Abstract

We address the speaker partitioning problem on datasets composed of two-speaker conversations. In such a situation, it is desirable to obtain a good overall diarization performance but even in that case, the performance of the partitioning problem can be severely degraded if some of the recordings are incorrectly segmented. We show that the performance of a bottom-up speaker clustering approach for the partitioning of two-speaker conversation datasets is sensitive to errors in the diarization, up to a point that the Diarization Error Rate for every recording should be as low as 1% to avoid degradation in performance due to the diarization process. Finally we propose a set of confidence measures along with a logistic regression approach to detect those conversations whose segmentation hypothesis is reliable enough to perform speaker clustering, showing that it enables an improvement in clustering performance at the expense of missing a small portion of the speakers in the dataset.

Index Terms: Speaker Partitioning, Speaker Clustering/Linking, Diarization, Confidence Measures

1. Introduction

It is usual in telephone environments to find large unlabeled datasets composed of two-speaker conversations recorded on a single channel. In some situations, these datasets are intended for speaker modeling, however, they are not directly useful, unless two problems are solved.

The first problem is the fact that two speakers are present on each recording. It has been shown that the presence of two speakers in a recording degrades severely the performance of a speaker characterization system [1]. To avoid this effect, a speaker diarization system is needed.

The second problem is the fact that a single speaker is usually present in several recordings within the dataset. Thus, it is mandatory to cluster those segments extracted from different recordings that belong to the same speaker, in order to avoid obtaining different models for a single speaker, and to provide robustness to the speaker models using more data for training.

Traditional solutions for the speaker diarization problem includes Agglomerative Hierarchical Clustering (AHC) strategies using mostly Bayesian Information Criterion (BIC) for both merging segments and determining the number of speakers [2]. On the other hand, recently, new approaches based on Joint Factor Analysis (JFA) have shown to outperform the traditional ones when tested on two-speaker telephone conversations [3].

Speaker clustering is a key aspect in the field of speaker diarization, specially when the number of speakers is priorly unknown. However, the speaker clustering task studied in this work differs significantly from that performed in speaker diarization tasks. The main difference is that in this case, we want to cluster speaker segments extracted from different sessions so inter-session variability must be taken into account. The problem of speaker clustering across different datasets has been studied previously as speaker linking [4] or from a more general point of view as the speaker partitioning problem [5].

In this work, we address the speaker partitioning problem on datasets composed of two-speaker telephone conversations. We propose a system to perform diarization and speaker clustering on the dataset, and we study the impact of the diarization error on the performance of the clustering process. Finally we propose a method for detecting correctly segmented recordings, based on a set of confidence measures, in order to improve the performance of the partitioning task.

2. System Description

The proposed approach for speaker partitioning of two-speaker conversation datasets follows four steps. Firstly, every conversation is processed by a speaker diarization system, in order to segregate the two speakers present. Then, an optional detection of correctly segmented conversations can be performed, in order to avoid feeding impure segments into the clustering process. Then, speaker clustering is performed over the remaining segments, and finally, a stopping criterion estimates the number of speakers in the dataset. Figure 1 shows the block diagram of this approach. Every block is detailed in the next subsections.

2.1. Diarization

To perform speaker diarization, we use the two-speaker segmentation system described in [1]. This system, by means of eigenvoice modeling, is able to extract a compact representation of the speaker present in a small window, so a sequence of these representations, known as speaker factors, can be extracted for a given recording. The speaker factors have shown to enable high separability among different speakers [3], [1].

In order to extract the speaker factors, we use a 1024 Gaussian Universal Background Model (UBM), and as feature vectors we use 19 Mel Frequency Cepstral Coefficients (MFCC) including C0, computed every 10 ms over a 25 ms window, plus delta features. For every frame, over a 1 s window, a speaker factor vector of dimension 100 is extracted. Dimensionality reduction from 100 to 50 and intra-session variability are achieved by means of LDA and Within Class Covariance Normalization (WCCN), as described in [6]. Once we obtain the sequence of
speaker factors for a recording, we perform the following steps:

- Initialization: the system applies PCA to the speaker factors, to obtain the best direction to separate the two speakers. Two clusters are obtained using uniquely this direction, and then refined using K-means.
- Core segmentation: a 2-component GMM is trained from the initial clusters, and each one of these Gaussians is assigned to a single speaker. The frames are reassigned using Viterbi segmentation in the speaker factor space. We repeat this process iteratively until convergence.

For training the UBM, the eigenvoice matrix and the LDA and WCCN transformations, we use data from the NIST Speaker Recognition Evaluations (SRE) 2004, 2005 and 2006.

2.2. Detection of correctly segmented conversations
After the diarization step, every recording can be analyzed in order to determine whether its segmentation hypothesis is acceptable for the clustering process or not. We expect incorrectly segmented recordings to degrade the performance of the clustering process since the segments they provide are not pure.

To avoid this, we need to detect those segmentation hypotheses with a diarization error rate below a certain level, in order to discard the rest of the dataset. In this work, diarization performance is measured in terms of Diarization Error Rate (DER), as described later in section 3. This way, we set a threshold for the DER, \( t_{\text{DER}} \), so that every segmentation hypothesis obtaining a DER below \( t_{\text{DER}} \) is considered to be correct.

We make use of a set of four confidence measures in order to determine whether a conversation is correctly segmented:

- BIC: for every recording, we compute the BIC for the hypothesis of having two speakers and for the hypothesis of having a single speaker. The confidence measure is the difference between both BIC values, as in [1].
- Speaker Verification using i-vectors and Probabilistic LDA (PLDA) [7]: the same speaker verification system used in the speaker clustering step (see section 2.3) can help to determine whether the two hypothetical segments obtained by the speaker diarization system actually belong to different speakers.
- Speaker Verification using speaker factors and cosine distance: a vector of speaker factors is computed for each segment obtained from a given recording, and after performing intra-session variability compensation, the cosine distance is computed as confidence measure.
- Eigenvalue Spread Ratio of the speaker factors: The diarization system considered assumes that the speaker factors belonging to a single speaker follow a normal distribution whose covariance matrix is the identity [1], [6]. The eigenvalue spread of the speaker factors for every hypothetical segment indicates whether this assumption is satisfied. Moreover, the eigenvalue spread of the speaker factors for the whole conversation should be higher since we expect to find a direction of high separability between two speakers. Therefore, the ratio between this last eigenvalue spread and the product of the eigenvalue spreads obtained for every hypothetical segment can be used as confidence measure.

In order to detect those recordings obtaining a DER below \( t_{\text{DER}} \), we train a Linear Logistic Regression model with the confidence measures described. The logistic regression model is trained on the 2213 conversations of the NIST SRE08 summed dataset, processed by the speaker diarization system described in section 2.1.

2.3. Speaker Clustering
The speaker clustering process is performed in the same fashion as in [4]. A speaker verification system is used to build a score matrix \( S \), whose entry \( s_{i,j} \) is the score obtained for the hypothesis of segments \( i \) and \( j \) to belong to the same speaker. Then, an iterative agglomerative clustering process is performed until one single cluster is obtained, keeping track of the sequence in which the different segments have been agglomerated. In every step of the iterative clustering, those segments obtaining higher score are merged. As in [4], every time two segments are merged in a single cluster we do not recompute the scores considering the new cluster. Instead of that, we use the initial score matrix, and every time two segments are to be merged, we merge the collections of segments they belong to.

To build \( S \), we use a gender independent i-vector PLDA speaker verification system [7]. For every hypothetical speaker, an i-vector of dimension 400 is extracted, centered and normalized to unit length, and the score obtained for two given i-vectors \( v_i, v_j \) extracted from the segments \( i, j \) is the log-likelihood ratio given by the PLDA model:

\[
S_{i,j} = \log \frac{P(v_i, v_j)}{P(v_i)P(v_j)}
\] (1)

The PLDA model considered uses a low rank eigenvoice matrix of dimension 100 and a full rank eigenchannel matrix. For training the total variability matrices and the PLDA model, the same data from NIST SRE 04, 05 and 06 as in the segmentation system was used. The features considered are 18 MFCC discarding c0 plus delta features, which are normalized applying feature warping [8].

2.4. Stopping Criterion
A stopping criterion must be considered in order to determine the hypothetical number of speakers for the dataset. Since we are using a speaker verification system for speaker clustering, the most straightforward stopping criterion is to set a threshold for the verification score. The threshold should be obtained through a calibration process that is out of the scope of this work, so we will not deal with it. However, it should be kept in mind that the stopping criterion is one of the most critical steps in the design of a speaker partitioning system.

3. Experimental Setup
3.1. Datasets
The datasets considered for testing are extracted from the NIST SRE 2008 and 2010. The first dataset is the NIST SRE08 summed channel, which contains 2213 five minute telephone conversations. Only 2894 out of the 4426 sides (2213 × 2) of the conversations that can be extracted belong to a speaker whose identity is provided by NIST. The partitioning problem is evaluated only on the segments corresponding to these 2894 sides of the conversation. This dataset contains 1040 different speakers.

To validate our approach for detecting correctly segmented recordings, the NIST SRE10 summed channel dataset is consid-
tered. This dataset contains 2794 recordings, and only 3198 out of the 5588 sides of the conversations that can be extracted belong to a known speaker. The partitioning problem is evaluated on these 3198 sides of the conversation. This dataset contains 461 different speakers.

3.2. Performance Evaluation

The performance of the speaker diarization system is evaluated in terms of DER. The speech/non-speech labels are extracted from the ASR transcripts provided by NIST, so perfect speech/non-speech detection is assumed. Since overlapped speech is in general low for telephone conversations, and its detection is out of the scope of this work, overlapped speech is not considered for DER computation.

The performance of the speaker clustering system is measured in terms of cluster impurity and speaker impurity as defined in [4]. The cluster impurity increases as segments containing different speakers are clustered together, while the speaker impurity increases as the set of segments belonging to a single speaker are assigned to different clusters. In every experiment, the point of equal impurity (EI) is obtained as measure of performance of the clustering system. To analyze the impact of the detection of correctly segmented recordings, we will study the EI as well as the fraction of speakers that have been kept in the dataset after discarding incorrectly segmented recordings.

4. Impact of the diarization error on the partitioning problem

To analyze the impact of the diarization error on the partitioning problem, we evaluate the proposed system without detection of incorrectly segmented recordings. We first consider several diarization systems obtaining different performances. The systems considered are listed below.

- A BIC AHC system with soft-clustering [3]
- A simplified speaker factor system that extracts 20 speaker factors without intra-session variability compensation, and does not use Viterbi resegmentation.
- The complete diarization system proposed in section 2.1
- The ground truth diarization labels.

Table 1: Diarization performance (DER), speaker clustering performance (EI) and number of clusters (C) obtained for different diarization systems, on the NIST SRE 08 summed dataset.

<table>
<thead>
<tr>
<th>system</th>
<th>DER</th>
<th>EI</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIC AHC</td>
<td>4.09%</td>
<td>18.87%</td>
<td>1314</td>
</tr>
<tr>
<td>simplified</td>
<td>2.89%</td>
<td>18.04%</td>
<td>1304</td>
</tr>
<tr>
<td>complete</td>
<td>1.31%</td>
<td>16.97%</td>
<td>1289</td>
</tr>
<tr>
<td>Ground truth</td>
<td>0.00%</td>
<td>15.38%</td>
<td>1122</td>
</tr>
</tbody>
</table>

Table 1 shows the diarization and speaker clustering performance for every speaker diarization approach in the NIST SRE 08 summed dataset, in terms of DER and EI, as well as the number of clusters (C) obtained for the EI point. It can be seen that the performance of the dataset partitioning degrades significantly as the performance of the diarization system degrades, even though the approaches for speaker diarization we are presenting here obtain very good performance in terms of DER. Even the complete speaker diarization system, that obtains a DER as low as 1.31%, degrades the EI from 15.38% to 16.97%. In this case, the most accurate speaker diarization system available is needed. So from now on, only the complete system described in 2.1 is considered.

This behavior can be explained by the fact that, although the overall DER is low, there are a few number of recordings with high DER. These recordings produce segments containing two speakers, that will mislead the iterative clustering process.

In order to find a threshold in the DER so that the recordings with DER below that threshold will not degrade the clustering performance, we rank all the recordings in the NIST SRE 08 summed dataset according to their DER. We split the ranked dataset into ten subsets of the same size, since the performance of the partitioning task depends on the size of the problem [4]. Thus, we obtain ten subsets of the same size containing recordings with different DER values.

Figure 2 shows the performance of the diarization system and the partitioning task for every subset. The solid curve represents the overall DER for every subset when they are sorted by descending DER, as well as the range of DER values for the recordings in the corresponding subset. We can see that only the first subset obtains high DER values, and that all recordings belonging to the last seven subsets obtain a DER below 1%.

The dashed curve represents the EI value when performing the partitioning task on every subset. The values are below those presented in table 1 since the subsets are smaller (one tenth of the dataset) and contain less speakers than the whole dataset used in that table (around 240 per subset, note that a single speaker can be in more than one subset). As we could expect, the EI decreases as the DER decreases, oscillating around 4% for the last seven subsets. It is interesting to note that even for low DER values, the EI can be very high, as it can be seen in for the second subset. This subset contains recordings with DER below 2.84%, but the EI is as high as 8.83%.

From this results we can conclude that the performance of the partitioning task is very sensitive to the DER obtained for the recordings in the dataset. According to figure 2, we see that the threshold in the DER value to obtain a performance not affected by the segmentation is very low. In this study, we consider $\text{th}_{\text{DER}} = 1\%$. This $\text{th}_{\text{DER}}$ value is quite different from that suggested in [3] for the task of speaker verification, which was as high as 10%. We believe that this difference is due to the fact that the partitioning problem is solved in this work using an iterative bottom-up clustering strategy, so slight degradations in the performance of speaker verification due to diarization errors will propagate through the speaker clustering process, degrading the final performance.
5. Speaker partitioning with detection of correctly segmented conversations

In order to evaluate the performance of the partitioning task when only those recordings detected as correctly segmented are fed into the clustering process, we run the partitioning system including the detection of correctly segmented recordings described in section 2.2, considering $\theta_{DER} = 1\%$.

Table 2: Performance of the partitioning task using detection of correctly segmented recordings, on the NIST SRE08.

<table>
<thead>
<tr>
<th>Segments</th>
<th>DER</th>
<th>EI</th>
<th>C</th>
<th>S</th>
<th>Segm</th>
</tr>
</thead>
<tbody>
<tr>
<td>All(2894)</td>
<td>1.31%</td>
<td>16.97%</td>
<td>1289</td>
<td>1040</td>
<td>0%</td>
</tr>
<tr>
<td>Correct(2146)</td>
<td>0.26%</td>
<td>12.18%</td>
<td>980</td>
<td>926</td>
<td>10.96%</td>
</tr>
<tr>
<td>Incorrect(748)</td>
<td>4.56%</td>
<td>15.29%</td>
<td>486</td>
<td>476</td>
<td>54.23%</td>
</tr>
<tr>
<td>Accept(1996)</td>
<td>0.39%</td>
<td>11.85%</td>
<td>934</td>
<td>896</td>
<td>13.85%</td>
</tr>
<tr>
<td>Reject(898)</td>
<td>3.50%</td>
<td>14.92%</td>
<td>565</td>
<td>517</td>
<td>50.29%</td>
</tr>
</tbody>
</table>

Table 2 shows the performance of the partitioning task when detecting correctly segmented recordings ($DER < 1\%$), for the NIST SRE08 summer dataset. The number of clusters obtained $C$, the number of speakers present in the detected recordings $S$ and the ratio of speakers missed during the discarding process ($\frac{S}{N-S}$) are shown for comparison. The accepted segments ("accept") are those that belong to a conversation detected by our system as correctly segmented, while the correct segments ("correct") are those that belong to a conversation that is actually correctly segmented ($DER < 1\%$).

We can see that our approach is able to detect 1996 segments with an overall DER as low as 0.39%, obtaining a reduction in the EI from 16.97% to 11.85%, at the expense of missing 13.85% of the speakers in the dataset. If we could select those correctly segmented recordings, the EI would increase slightly to 12.18%, but fewer speakers would be missed (10.96%).

Looking at the rejected subset of recordings ("reject"), we can see that the detection approach is working as desired, discarding a set of conversations whose overall DER is as high as 3.50% and would obtain an EI of 14.92%, which is very high given that the EI decreases as the subset is smaller.

Table 3: Performance of the partitioning task using detection of correctly segmented recordings, on the NIST SRE10.

<table>
<thead>
<tr>
<th>Segments</th>
<th>DER</th>
<th>EI</th>
<th>C</th>
<th>S</th>
<th>Segm</th>
</tr>
</thead>
<tbody>
<tr>
<td>All(3198)</td>
<td>2.45%</td>
<td>13.21%</td>
<td>771</td>
<td>461</td>
<td>0%</td>
</tr>
<tr>
<td>Correct(756)</td>
<td>0.32%</td>
<td>6.55%</td>
<td>332</td>
<td>309</td>
<td>32.97%</td>
</tr>
<tr>
<td>Incorrect(2242)</td>
<td>3.19%</td>
<td>13.55%</td>
<td>674</td>
<td>447</td>
<td>3.04%</td>
</tr>
<tr>
<td>Accept(2382)</td>
<td>1.58%</td>
<td>12.00%</td>
<td>654</td>
<td>440</td>
<td>4.56%</td>
</tr>
<tr>
<td>Reject(816)</td>
<td>5.14%</td>
<td>14.52%</td>
<td>371</td>
<td>325</td>
<td>29.50%</td>
</tr>
</tbody>
</table>

Table 3 shows the performance of the partitioning task when detecting correctly segmented recordings for the NIST SRE10 dataset. In this case, our approach detects many more conversations as correctly segmented that those that actually are: a total of 2382 out of 3198 segments are accepted as correctly segmented, while only 756 out of 2382 are actually correctly segmented. Because of this, the overall DER obtained for the accepted recordings is over 1%, and the EI is far over the EI obtained for the correct segments, 12.00% against 6.55%. However, the detection of correctly segmented recordings is still helping in this case: the DER of the accepted recordings is below the DER obtained for all recordings, and the EI is reduced from 13.21% to 12.00%, missing only 4.56% of the speakers.

Note that the difference between the performance obtained for the whole dataset and for the correct subset is in part due to the significant difference in the size of both datasets, and that 32.97% of the speakers are missed in the correct subset. The detection of correctly segmented recordings is introducing a high number of false alarms, but it is still discarding very wrong segmented recordings (DER of 5.14%), so it produces a little improvement in the EI, missing a small number of speakers.

6. Conclusions

We have presented a system for speaker partitioning of datasets containing two-speaker conversations, that makes use of state-of-the-art speaker diarization and speaker verification systems in order to segment the recordings and to perform speaker clustering on the obtained segments.

We have shown that the speaker clustering task is very sensitive to errors in the speaker segmentation hypotheses when performed in a bottom-up fashion, up to a point that considering recordings that obtain a DER over 1% in the clustering process will degrade the clustering performance. To avoid this effect, we have proposed an approach that makes use of four confidence measures to detect correctly segmented recordings, in order to discard harmful recordings. This approach enables to improve the performance of the clustering task, at the expense of missing some speakers present in the dataset.

For NIST SRE08, which is the development dataset, we are able to obtain a significant improvement in the performance of the clustering process in terms of EI reducing it from 16.97% to 11.85%, missing only 13.85% of the speakers present in the dataset. For NIST SRE10, the detection system accepts many incorrectly segmented recordings, but the DER for the accepted files is still reduced compared to the DER for all recordings. In this case, the reduction in terms of EI is not that significant, from 13.21% to 12.00%, but we are able to keep most of the speakers in the dataset, missing only 4.56% of them.

Finally, we have detected that there exist a trade-off between the performance of the clustering algorithm and the number of speakers kept in the dataset when detecting correctly segmented recordings. Thus, this approach should be trained depending on the application needs.

7. References


