Abstract
Speaker diarization is often performed before automatic speech recognition (ASR) to label speaker segments. In this paper we present two simple schemes to improve the speaker diarization performance. The first is to iteratively refine GMM speaker models by frame level re-labeling and smoothing of the decision likelihood. The second is to use word level alignment information from the ASR process. We focus on the CHIL lecture meeting data. Our experiments on the NIST RT06 evaluation data show that these simple methods are quite effective in improving our baseline diarization system, with alignment information providing 1% absolute reduction in diarization error rate (DER) and the re-label smoothing providing an additional 3.51% absolute reduction in DER. The overall system generates a DER that is 6.8% relative better than the top performing system from the RT06 evaluation.

Index Terms: speaker diarization, CHIL, lecture meeting, DER

1. Introduction
Integrated Project CHIL (“Computers in the Human Interaction Loop”), aims to create people-friendly computing by monitoring how people interact, exchange information and collaborate or socialize in meetings. An important initial step towards this goal is to be able to automatically generate transcripts of the conversational speech in the meeting under “always-on” audio capturing from far-field (non-intrusive) microphone sensors.

Seminar meetings recorded by distant microphones in noisy reverberant settings, multiple speakers with often overlapping speech, as well as the non-native accents and technical content of CHIL lectures make this data very challenging to the ASR research community. NIST Rich Transcription (RT) meeting recognition evaluation therefore takes the CHIL lecture meeting data as one set of test data, contrast to the interactive conference meeting data. There are three tasks being evaluated this year at RT07: speaker diarization (SPKR), speech-to-text (STT), and speaker-attributed STT (SASTT). Although speech activity detection (SAD) is removed from the evaluation, it remains an important step prior to speaker diarization.

Speaker diarization is generally performed after SAD and before STT.

The goal of speaker diarization is to label each speech segment from the SAD output with speaker information. This task is sometimes referred to as the “who spoke when” task. Typically, the number of speakers in the recording or the characteristics of the recording are unknown a priori. This information needs to be determined automatically for the diarization task. As a result of the NIST speaker diarization evaluations from the past few years the research community has made significant progress on the diarization task [1-8].

There are two main approaches to the speaker diarization problem: the first is a bottom-up approach (i.e. hierarchical, agglomerative clustering) [14, 10, 15], and the second is a top-down approach with evolutive hidden Markov models (E-HMM) [7], starting with one speaker and detecting and adding speakers in succession. Agglomerative clustering methods generally involve several steps: Initially the speech segments as determined from the SAD output are investigated for possible speaker change points [16]. The output of change point detection is then fed into a speaker clustering procedure. The clustering stops when a predetermined criterion is satisfied (for example a drop in overall data likelihood from a merge). The limitation of this approach is that the errors in the first two steps would carry over to the final clustering step.

An improvement is to optimize the segmentation and clustering jointly with an iterative procedure as in [11] using Gaussian Mixture Models (GMMs) for each cluster. Recently ICSI proposed purification algorithms to the iterative segmentation scheme to improve performance. Impure segments are removed before the cluster merging step, and impure frames are removed from GMM train and clustering merging [14]. LIMSI proposed to use speaker identification combined with BIC to improve performance [10, 9]. The speaker ID idea is not quite suitable for the CHIL lecture meetings in which there is one main speaker. The remaining speakers in the lecture usually do not have enough data to generate a reliable MAP adapted speaker model (some speakers talk as little as 2 seconds). The same problem occurs with the E-HMM approach where speaker models are needed. This approach
usually detects the largest speaker well but misses speakers with sparse data. [7].

Although it is permitted to use the STT output to assist with speaker diarization in the NIST RT evaluation, most of the systems did not take advantage of the word output from STT. Only LIMSI had exploited spoken cues (“Back to you, Bob”) to add information to the diarization output for Broadcast News data [12], and remove short-duration silence segments [10].

In this paper, we implement simple yet effective schemes to improve speaker diarization for CHIL lecture meetings: the first is to build GMM speaker models from an available segmentation (for example from a SAD output) and refine the label of each frame with these GMM models followed by smoothing the labeling decision with its neighbors. A result of the reclassification and smoothing is the possibility that the original segments can be further segmented, in effect locating speaker change points within the initial segmentation. In a comparison to our RT06 clustering system with change-point detection, the refinement detects the speaker change-point with more reliable global speaker models as opposed to using local windows to make the change-point decision; the second scheme is to use word information from the speaker independent ASR process. There are two advantages of using word alignment information from a SI decoding: one is to filter out non-speech segments which constitute false alarm errors, and input more accurate speech segments to the speaker clustering step; the other is to use word alignment information to remove short silence, background noise and vocal noise that do not discriminate speakers and causes overlaps of cluster models. Only useful speech frames are used to train and compare cluster models. This technique of using word output from the recognizer is similar to ICSI’s segment-based purification and frame-based purification algorithms [14]. The difference being acoustic cues are used in ICSI’s purifying algorithms, in contrast our approach takes advantage of both acoustic and language models from the SI decoding.

The remainder of the paper is organized as follows: Section 2 describes our baseline speaker diarization system used for the NIST RT06 STT task last year, and Section 3 presents ways of improvement to the baseline system. Section 4 is devoted to the experimental study and discussions, and Section 5 concludes the paper.

2. Baseline Diarization Systems

Our baseline speaker diarization system was developed for the EARS transcription system. The framework is similar to the one described in [15]. The diarization runs through the following steps:

- speech/non-speech segmentation (SAD): Owing to the fact that our interest in this paper is to compare the performance of the speaker diarization process we remove the impact of the SAD and allow for fair comparison across the various sites submitting to the NIST 2006 evaluation. We use the SAD output from the site with best diarization performance as a starting SAD input to our system.

- speaker change detection: speech segments from SAD are segmented further into homogeneous segments using Ajmera and Wooters’ change-point detection procedure [16] with a modified BIC criteria. This step is optional, i.e. the SAD input can go directly to the speaker clustering procedure.

- speaker clustering: a single Gaussian density is used to model a cluster. Clusters are initialized into a pre-specified number using $K$-means and a Mahalanobis distance measure. The cluster merging stops when the number of desired clusters is reached, which is set by the development data.

All the above procedures use PLP features. We skip the re-segmentation step [15] because we haven’t obtained any gain from this step. This simple scheme of speaker segmentation would be OK for speech-to-text recognition. But the diarization error rate (DER) is rather high (above 60%). In the following section we present simple ways to improve upon the baseline speaker diarization system.

3. Improvements to the Speaker Diarization System

We made the following modifications upon the baseline speaker segmentation scheme:

1. As opposed to using 24 dimensional PLP features, we switch to 19-dim MFCC features with no energy term. This change was made since the 19 dim MFCC are widely used in speaker identification systems.

2. Instead of using a fixed number for speaker clustering, we first over-segment the data into, say 8 clusters, merge clusters according to Mahalanobis distance function, and stop the merging process when a threshold value is reached, which is determined by the development data.

3. At this point we change the distance function from a Mahalanobis distance measure to likelihood gain. Each cluster is modeled by one Gaussian with full-covariance matrix. At each step in the bottom up clustering process we combine the two nodes which result in the smallest likelihood loss if combined. We stop when no two nodes can be combined with a loss smaller than a pre-specified threshold.
The speaker diarization performance is measured in terms of Diarization Error Rate (DER): the error time includes speaker error time, missed speaker time and false alarm speaker times. The latter two are coming from the SAD errors which as we previously mentioned (section 2) is beyond our control in this setup since we are using the SAD segmentation from the top performing RT06 system. Therefore to improve SPKR performance we need to reduce speaker errors and possibly false alarms. We propose to use the word alignment information from a simple speaker-independent word decoding to achieve these goals:

1. Remove segments with only silence, background noise and vocal noise. These segments can exist because of the SAD step failure. This step is referred as “non-speech removal”. The correctly-removed non-speech segments would reduce false alarm errors.

2. In the baseline speaker clustering step, ignore frames that correspond to silence, background noise and vocal noise when computing single-Gaussian models for each cluster. This step is referred to as “speech-frame only”, performed after silence removal.

The first usage is segment-based purification, and the second usage is frame-based purification [14]. By identifying non-speech frames and removing them from training speaker models, one would certainly expect better speaker clustering results. With word alignment information, there is no need to estimate parameter \( P \) to remove \( P\% \) frames in each cluster from GMM training in [14].

Our baseline SPKR system and the above enhancement by using alignment information do not break the speech segments input by SAD, which has a detrimental effect on the subsequent speaker models as well as increasing the speaker error rate if speaker change points occur within a segment. To resolve this problem, we add an iterative refinement process after the merging stops: We build GMMs for each speaker cluster based on the previous decision. We then re-label each frame by averaging scores of the speaker models for 0.75 seconds on each side of the current frame and then assigning the frame to the speaker with the largest smoothed score. The advantage of frame level refinement is to break a segment if there are speaker change points within the segment. Therefore we have generated superior speaker models from all segments assigned to a speaker. We have found that two iterations are sufficient.

4. Experiments and Results

Our SAD and SPKR systems are developed for the RT07 evaluation. Since the official results from NIST are not available, we choose to report our experimental results on the NIST RT06a evaluation of CHIL lecture meeting data, and use RT05 CHIL evaluation data (from CHIL site UKA only) as development data (dev) to determine the threshold used in our clustering step. RT06a CHIL evaluation data (eval) contains a total of 190 minutes of lecture meetings recorded in the smart rooms of five CHIL sites: AIT, IBM, ITC, UKA, and UPC. NIST decided to score only 28 lectures from this set, with 4 lectures from AIT, 4 lectures from IBM, 2 lectures from ITC, 14 lectures from UKA and 4 lectures from UPC. In this study, experiments of the speaker diarization (SPKR) systems are conducted with a single distant microphone (SDM) specified by NIST.

As already mentioned the speaker diarization performance is measured in terms of Diarization Error Rate (DER) as defined by the NIST RT evaluations. DER is calculated by first finding the optimal one-to-one mapping between the reference speakers and the system hypothesized speakers, and then computing the percentage of time that is wrongly assigned according to the optimal mapping. The overall DER is computed as the fraction of speaker time that is not attributed correctly to a speaker. The error time includes speaker error time, missed speaker time and false alarm speaker times, thus taking the SAD errors into account [3]. Overlapping speech is also included in the DER calculation. The performance of SAD is also measured in the same DER as diarization except all speakers are labeled as speech. The contribution to the SAD error rate to the overall DER can be removed, which is the procedure chosen here (section 2). With the use of the SAD segmentation from the top performing speaker diarization system as input to the proposed diarization scheme we feel that a fair comparison of pure speaker diarization performance is achievable.

Owing to the fact that NIST changed this year from a manual segmentation reference to a forced alignment reference, the numbers presented here differ significantly from the officially published evaluation results last year on SAD and speaker error. But nonetheless the scoring is accomplished with the same references.

Table 1 illustrates the performance of the three systems on the development data set. The IBM Baseline system is as described in section (2), which uses the 24 dimensional PLP features. The IBM 1 system is based on the 19 dimensional MFCC, with over segmented clusters, merged by Mahalanobis distance, the IBM 2 system uses 19 dimensional MFCC and the likelihood merge, inclusion of word level alignment information is represented by “align”, and the iterative GMM refinement specified by “refine”, as discussed in section (3). The refinement is run for two iterations.

Table 2 illustrates the impact of the word level alignment information and refined labeling steps as discussed in section 3. Alignment provides approximately 1% absolute reduction in DER, with an additional 3.51% ab-
Table 1: Diarization error rate of various IBM systems on the development data (RT05 CHIL evaluation data).

<table>
<thead>
<tr>
<th>systems</th>
<th>DER</th>
<th>Speaker error time</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM Baseline</td>
<td>8.6</td>
<td>616.51</td>
</tr>
<tr>
<td>IBM 1</td>
<td>11.8</td>
<td>805.22</td>
</tr>
<tr>
<td>IBM 2</td>
<td>8.5</td>
<td>590.23</td>
</tr>
<tr>
<td>IBM 1 align</td>
<td>7.7</td>
<td>515.33</td>
</tr>
<tr>
<td>IBM 2 align</td>
<td>7.6</td>
<td>496.98</td>
</tr>
<tr>
<td>IBM 2 align+refine</td>
<td>8.3</td>
<td>512.68</td>
</tr>
</tbody>
</table>

Table 2: Diarization error rate of the best performing system from the RT06 evaluation and the IBM baseline system along with the systems incorporating the proposed enhancements (on the 28 segment 2006 evaluation data).

<table>
<thead>
<tr>
<th>systems</th>
<th>DER</th>
<th>Speaker error time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best RT06</td>
<td>13.6</td>
<td>232.53</td>
</tr>
<tr>
<td>IBM Baseline</td>
<td>20.1</td>
<td>527.34</td>
</tr>
<tr>
<td>IBM 1</td>
<td>17.2</td>
<td>395.79</td>
</tr>
<tr>
<td>IBM 2</td>
<td>17.1</td>
<td>390.73</td>
</tr>
<tr>
<td>IBM 1 align</td>
<td>15.8</td>
<td>342.79</td>
</tr>
<tr>
<td>IBM 2 align</td>
<td>16.3</td>
<td>375.71</td>
</tr>
<tr>
<td>IBM 2 align+refine</td>
<td>12.8</td>
<td>216.72</td>
</tr>
</tbody>
</table>

Overall we have a 6.8% relative improvement over the best performing system in the RT06s evaluation. Based on table 1, the results on the evaluation data are somewhat unexpected, although all of the development data is coming from a single site (UKA), where speaker interaction is minimal. After threshold tuning on the development test set, the results across all systems are quite close, with the outlier being the IBM 1 system with a DER of 11.8%. It appears that we have reached a limit of performance for all systems after threshold tuning for this development data set. Table 2 indicates the ability of the systems to generalize from these tuned thresholds moving from the baseline system to the align, refined system. The data from the evaluation set contains more speaker interaction.

5. Conclusion and Future Work

In this paper, we have presented two simple techniques that provide us with improved speaker diarization performance: use of word level alignment information and an iterative labeling, GMM building step.

We are currently investigating use of phoneme level information in the alignment step, with the expectation that removing those phonemes that don’t carry speaker discriminative information should enhance our speaker models.

6. Acknowledgment

The work was partially supported by the European Commission under integrated project CHIL.

7. References