IMPROVING SPEAKER DIARIZATION BY CROSS EM REFINEMENT

Huazhong Ning

Beckman Institute
U. of Illinois at Urbana-Champaign
Urbana, IL 61801
hning2@ifp.uiuc.edu

Wei Xu, Yihong Gong

NEC Laboratories America, Inc.
10080 N. Wolfe Road
Cupertina, CA 95070
{xw,ygong}@sv.nec-labs.com

Thomas Huang

Beckman Institute
U. of Illinois at Urbana-Champaign
Urbana, IL 61801
huang@ifp.uiuc.edu

ABSTRACT

In this paper, we present a new speaker diarization system that improves the accuracy of traditional hierarchical clustering-based methods with little increase in computational cost. Our contributions are mainly two fold. First, we include a preprocessing called "local clustering" before the hierarchical clustering algorithm to merge very similar adjacent speech segments. This local clustering aims to reduce the number of segments to be clustered by the hierarchical clustering, so as to dramatically increase the processing speed. Second, we perform a postprocessing called "cross EM refinement" to purify the clusters generated by the hierarchical clustering. This algorithm is based on the idea of cross validation and EM algorithm. Our experimental evaluations show that the proposed cross EM refinement approach reduces the speaker diarization error by up to 56%, with an average reduction of 22% compared to the traditional hierarchical clustering method.

Keywords: *Cross EM Refinement, Hierarchical Clustering, BIC, Speaker Diarization*

1. INTRODUCTION

Speaker diarization (also called speaker segmentation) is the task of segmenting a multi-speaker audio document into homogeneous parts and then clustering the resulting parts into groups which each contains only the voice of a single speaker. With the explosive growth of audio documents both on the Internet and in corporate information archives, speaker diarization techniques have been receiving more and more attentions because they are valuable enabling tools for developing various advanced audio access and playback functionalities. To promote research in this area, NIST has initiated the speaker diarization contest [1] since 2002, and the number of participants for the contest has been increasing steadily each year.

Given an unknown audio document, generally there is no prior knowledge available on the number nor the profiles of the speakers within the document. Therefore, we must employ unsupervised clustering techniques to detect the number of speakers, and to segment/cluster different speakers appropriately.

There is a large volume of literature on speaker diarization research. Most methods use a mixture of Gaussians to model audio segments, and use the hierarchical clustering method along with certain model selection metrics (e.g. BIC) to group the audio segments into an appropriate number of clusters [1, 2]. In [3] Moraru, *et al.* investigated two possibilities for combining their previous LIA and CLIPS systems, first using an hybridization strategy and then merging the proposed segmentations. Tranter and Reynolds also presented a hybrid system developed to allow the benefits of their CUED and MIT-LL systems to be exploited in a single system [4]. Jin and Schultz used a tied GMM for both segmentation and clustering, which is also adopted as part of our speaker diarization system due to its accuracy and speed [5]. Auguera, *et al.* introduced a "purification" module that tries to keep the clusters acoustically homogeneous throughout the hierarchical clustering process [6]. This approach shares the same goal as our cross EM refinement, but takes a different approach from ours.

In this paper, we present a new speaker diarization system that improves the accuracy of traditional hierarchical clustering-based methods with little increase in computational cost. Our contributions are mainly two fold. First, to reduce the number of segments to be clustered by the hierarchical clustering algorithm, so as to dramatically increase the processing speed, we conduct a preprocessing called "local clustering" to merge very similar adjacent segments. This local clustering is based on the observation that it is highly probable that adjacent segments belong to the same speaker, or the same type of audio sound. Second, which is more important, to further purify the hierarchical clustering result, we perform a postprocessing called "cross EM refinement" that is based on the idea of cross validation and EM algorithm. Our experimental evaluations show that the proposed cross EM refinement approach reduces the speaker diarization error by up to 56%, with an average reduction of 22% compared to the traditional hierarchical clustering method.

The rest of the paper is organized as follows. Section 2 provides the overview of our speaker diarization system.

[1] http://www.nist.gov/speech/tests/rt/rt2006/spring/

Section 3 describes the proposed cross EM refinement algorithm in detail. Section 4 presents the experimental evaluations, and section 5 concludes the paper.

2. SPEAKER DIARIZATION SYSTEM OVERVIEW

Our speaker diarization system consists of the following major steps:

1. Silence detection to detect and remove the silent segments whose time length is above the predefined threshold.

2. Feature extraction to compute the 20 mel-frequency cepstral coefficients (MFCC), whereby to form the feature vector for each non-silent audio segment.

3. GMM construction to train a Gaussian mixture model on the entire set of non-silent audio segments, and then obtain the GMM coefficients for each non-silent audio segment using the EM algorithm.

4. Segmentation of each non-silent audio segment into homogenous segments based on the Baysian Information Criteria (BIC). This segmentation algorithm intends to yield the set of homogenous segments which maximizes the BIC metric.

5. Speech segment detection to detect the audio segments that contain human speech only. This is achieved by a binary classifier able to classify each audio segment into either the speech or non-speech class. The resulting speech segments are used as input to the subsequent clustering operations.

6. Local clustering to merge similar adjacent speech segments based on the BIC metric. This local clustering is based on the observation that it is highly probable that adjacent segments belong to the same speaker, or the same type of audio sound.

7. Hierarchical clustering to group the speech segments into the number of clusters that maximize the BIC metric. This operation is similar to the approach described in [5].

8. Cross EM refinement to refine the hierarchical clustering result. This is based on the idea of cross validation and EM algorithm (See Section 3 for detailed descriptions).

In the entire speaker diarization process, the GMM and BIC have been utilized in every segmentation and clustering step (i.e., the step 4, 6, and 7) to determine whether audio segments should be merged or split, and when the segmentation/clustering processes should be terminated. In fact the decision making problem on whether segments should be merged or split can be modelled as the model selection problem, and BIC has been widely used as a model selection criterion. Let $\mathcal{D} = \{X_i : i = 1, ..., N\}$ be the data set to be clustered. The BIC defines the quality of model M to represent data set $\mathcal{D}$ as follows:

$$BIC(M) = \log L(\mathcal{D}, M) - \frac{\lambda}{2}\rho(M)\log(N) \qquad (1)$$

where $L(\mathcal{D}, M)$ is the maximum likelihood of the data set $\mathcal{D}$ with the model M, $\rho(M)$ is the number of free parameters in the model M, and λ is the penalty of the model complexity. Theoretically λ is 1 but in practice it should be tuned due to the imperfectness of the model.

Let M_0 be the model that considers the data set $\mathcal{D}$ as one cluster, and M_1 be the alternative model that considers $\mathcal{D}$ as being formed by two clusters. Using the BIC, the choice between M_0 and M_1 becomes as simple as computing $BIC(M_0)$ and $BIC(M_1)$, and selecting the one with a larger BIC value.

In step 6 of the above operations, the local clustering is introduced before the hierarchical clustering to reduce the computational cost. The time complexity of the hierarchical clustering algorithm is proportional to $O(n^2 \log n)$, where n is the number of total audio segments to be clustered. This operation is very computationally expensive when n is large. In contrast, as the local clustering considers only adjacent audio segments for possible merges, its time complexity is $O(n)$, which is a huge saving compared to $O(n^2 \log n)$. By performing the local clustering operation on the audio segments, we aim to remarkably reduce the number of segments to be processed by the hierarchical clustering. Our experimental evaluations have shown that the number of audio segments will be reduced by nearly 66% after the local clustering, and therefore, the total time complexity decreases to $O(n)+O((n/3)^2 \log(n/3))$, about $1/9$ of the original complexity, compared to the operation using the hierarchical clustering only.

As described above, both the local and hierarchical clustering use the BIC metric to determine whether to merge two segments or not. In our implementation, a very small λ in Eq.(1) is used for local clustering (1.5 compared to 5.0 for hierarchical clustering). Because small λ favors splitting, this is to ensure that there is no or very little over-clustering at this step.

3. CROSS EM REFINEMENT

The local and hierarchical clustering algorithms generate speaker diarization results which still have large rooms for improvement due to the following reasons. First, the hierarchical clustering uses a greedy search for grouping audio segments, which generates only suboptimal solutions [1]. Second, BIC itself involves approximations which cause errors, especially when the audio segments are very short (less than *3* seconds). Third, the local clustering algorithm may induce some errors as well.

We introduce the "cross EM refinement" algorithm to refine the hierarchical clustering results. We start with the clusters generated by the hierarchical clustering algorithm, and train a GMM Θ_k for each cluster k using the EM algorithm. The obtained GMMs Θ_k, $k = 1, 2, \ldots, K$ are

applied to the entire data set $\mathcal{D}$ to generate a new clustering result (E-step). The new clustering result is then used to update each GMM Θ_k (M-step). This EM process is repeated until the clustering result converges. Note that our cross EM refinement algorithm involves two EM processes: one is the EM process that iteratively purifies the clustering result (we call it outer EM), the other is the EM process used by the M-step of the outer EM to update the GMM Θ_k for each cluster k (we call it inner EM).

The outer EM process can be more accurately described as follows. At the E-step, the expected value of the conditional probability $\hat{P}(k|X_i)$, the probability that the audio segment X_i belongs to the cluster k, is computed for each k and X_i, and the cluster label $\Upsilon(X_i)$ of each segment $X_i \in \mathcal{D}$ is computed using the equation $\Upsilon(X_i) = \arg\max_k \hat{P}(k|X_i)$. This E-step results in a new clustering result $\mathcal{D}_k^{new} = \{X_i|\Upsilon(X_i) = k\}$. At the M-step, GMM Θ_k for each cluster k is updated using the clustering result $\mathcal{D}_k^{new}$, $k = 1, 2, \ldots, K$, from the E-step. Again, the EM algorithm is used to update the GMM parameters including the mean, covariance matrix of each Gaussian component, as well as the Gaussian mixture coefficients. These two steps are iterated until convergence.

To prevent the GMM from overfitting during the EM iterations, we randomly and equally divide each cluster $\mathcal{D}_k$ into two groups $\mathcal{D}_k^{(1)}$ and $\mathcal{D}_k^{(2)}$. In the M-step, one group is chosen for training the GMM, and in the E-step, only the speech segments in the other group are updated on their cluster labels. In the next E- and M-step, the two groups are switched for the training and re-labelling purposes. Because the division of each cluster into two groups, and switching between the two groups for training and relabelling are similar to the idea of cross validation in some sense, we call this algorithm "cross EM refinement".

In summary, our cross EM refinement algorithm consists of the following main steps:

1. Take the clusters $\mathcal{D}_k$, $k = 1, 2, \ldots, K$, generated by the hierarchical clustering, randomly and equally divide each cluster $\mathcal{D}_k$ into two groups $\mathcal{D}_k^{(1)}$, $\mathcal{D}_k^{(2)}$, and set $\mathcal{D}^{(2)} = \mathcal{D}_1^{(2)} \cup \cdots \cup \mathcal{D}_K^{(2)}$.

2. M-step: For each cluster $\mathcal{D}_k$, use $\mathcal{D}_k^{(1)}$ to train the GMM Θ_k using the EM algorithm (inner EM).

3. E-step: For each segment $X_i \in \mathcal{D}^{(2)}$, compute

$$\hat{P}(k|X_i) = P(X_i|\Theta_k), \quad k = 1, ..., K \quad (2)$$
$$\Upsilon(X_i) = \arg\max_k \hat{P}(k|X_i) \quad (3)$$

   For each cluster $\mathcal{D}_k$, update the group $\mathcal{D}_k^{(2)}$ as follows:

$$\mathcal{D}_k^{(2)} = \{X_i|X_i \in \mathcal{D}^{(2)}, \Upsilon(X_i) = k\}. \quad (4)$$

4. For each cluster $\mathcal{D}_k$, switch between the two groups $\mathcal{D}_k^{(1)}$ and $\mathcal{D}_k^{(2)}$.

5. If the clustering result converges, terminate the process; otherwise, go to Step 2.

In the above cross EM refinement process, training GMM in the M-step is very time consuming, especially for high-dimensional and large data sets such as audio segments in our case. Fortunately, in each iteration of the GMM training, the GMM parameters can be initialized by the result either from the hierarchical clustering or from the previous iteration. Started from these initial values, the GMM training process converges very quickly (less than 5 iterations in our experiments). Furthermore, the entire EM refinement process also converges very fast ($\leq$ 3 iterations) when the hierarchical clustering generates a reasonable result. Therefore, our cross EM refinement algorithm improves the speaker diarization accuracy with little increase in computational cost.

4. EXPERIMENTS

The test data we used in our experiments are audio records of Japanese Parliament Panel Discussions. There are nine such audio records with the lengths ranging from 20 to 45 minutes (See Table 1, columns $1 \sim 3$). All the nine audio files were labelled by human annotators to form the ground truth for performance evaluations. Each audio segment can take one of the following three labels: *silence*, *non-speech*, and *speech* with a unique speaker ID. Only one audio file was used for tuning the parameters of our speaker diarization system.

We use the following "diarization error" defined by the NIST Rich Transcription Evaluation [7] as our evaluation criterion:

$$derr = \frac{T_{falarm} + T_{miss} + T_{wrong}}{T_{ref}} \quad (5)$$

where T_{falarm} is the total length of the non-speech segments that were classified as speech, T_{miss} is the total length of the speech segments that were classified as either non-speech or silence, T_{wrong} is the total length of the speech segments that were correctly classified as speech, but clustered into wrong speaker groups, and T_{ref} is the total length of all speech segments in the ground truth. In addition to $derr$, we also introduce the following $purity$ metric:

$$purity = \frac{\textit{pure time}}{\textit{total system speaker time}} \quad (6)$$

For each speaker identified by the system, we find a reference speaker from the ground truth that shares the longest time with the system speaker. The *pure time* is the sum of all these shared times. The *purity* metric is useful for the applications which care less about over-segmentation (i.e., one speaker may be separated into multiple clusters) but more about the "cleanliness" of each cluster.

Table 1 (columns $4 \sim 7$) shows the performances of our speaker diarization system on the nine Japanese Parliament

Table 1. Speaker Diarization Error and Purity with and without the cross EM refinement.

File Information			Without EM Refinement		With EM Refinement	
file	length (sec.)	#spkrs	error (%)	purity (%)	error (%)	purity (%)
1	2366	8	20.59	84.66	14.37	88.00
2	2201	7	6.43	90.21	6.18	90.51
3	1878	7	6.35	90.26	5.10	91.07
4	1475	8	4.86	91.42	5.73	90.61
5	2457	9	6.90	90.85	4.90	91.75
6	1876	9	13.94	91.48	6.86	91.03
7	1938	11	7.22	90.16	6.45	90.95
8	1260	6	3.07	93.30	3.19	93.20
9	2699	11	29.43	84.72	26.27	84.51
avg.	2017	8.4	10.98	89.67	8.78	90.18

audio records. To reveal the effectiveness of our EM refinement algorithm, we have also implemented the speaker diarization system without the EM refinement process, and tested it using the same test data set. This implementation is equivalent to the current state-of-the-art speech diarization approaches [5], and serves as the baseline for performance comparisons. The performance scores of the two systems are displayed shoulder by shoulder in the table.

The average $derr$ and $purity$ are 8.78% and 90.18% respectively for our speech diarization system with the cross EM refinement. Compared with the average performances of 10.98% and 89.67% for the baseline system, our system achieves a relative improvement of 21.6% for $derr$ and 1% for $purity$. The average improvements are not very salient because the results of some audio records (e.g. audio 2, 3, 4 and 8) are already quite good even without the EM refinement. However, for those audio records that the baseline system cannot handle well (i.e. audio 1, 6 and 9), the relative improvement is as much as 56% for $derr$ and 2% for $purity$.

It is worthwhile to note that the performances are relatively low on audio records 1, 6 and 9 compared to others. The reasons include: (1) the clusters of these audio records are very unbalanced, with the largest cluster being as long as 547.53 seconds while the smallest one as short as 18.83 seconds; and (2) there are much more arguments in these audio records, which usually contain stronger background noises and more mixture of speeches. It is obvious from the evaluations that our speaker diarization system handles this kind of complex audio records better than traditional hierarchical clustering-based approaches, with the reduction on $derr$ by up to 56%.

5. CONCLUSIONS

This paper presents a new approach that improves the performance of the traditional hierarchical clustering in speaker diarization while it requires little extra computational cost. The new approach includes a preprocessing before hierarchical clustering and a postprocessing after that. The preprocessing called "Local Clustering" is performed to merge the very similar temporally adjacent speech segments, so that the number of segments fed to hierarchical clustering, and in turn the computational cost, is highly reduced. The postprocessing called "cross EM refinement" is used to purify the clusters generated by the hierarchical clustering. Intensive experiments show the effectiveness of our approach.

6. REFERENCES

[1] Scott Shaobing Chen and P.S. Gopalakrishnan, "Clustering via the bayesian information criterion with applicationsin speech recognition," *ICASSP*, vol. 2, pp. 645–648, 1998.

[2] A. Tritschler and R.A. Gopinath, "Improved speaker segmentation and segments clustering using the bayesian information criterion," *EUROSPEECH*, 1999.

[3] Daniel Moraru, Sylvain Meignier, Corinne Fredouille, Laurent Besacier, and Jean-Francois Bonastre, "The elisa consortium approaches in broadcast news speaker segmentation during the nist 2003 rich transcription evaluation," *ICASSP*, 2004.

[4] S. E. Tranter and D. A. Reynolds, "Speaker diarisation for broadcast news," *Proc. Odyssey: The Speaker and Language Recognition Workshop in Toledo, Spain, ISCA*, pp. 337–344, 2004.

[5] Qin Jin and Tanja Schultz, "Speaker segmentation and clustering in meetings," *ICSLP*, 2004.

[6] B. Peskin X. Anguera, C. Wooters and M. Aguilo, "Robust speaker segmentation for meetings: The icsi-sri spring 2005 diarization system," *Proceedings of NIST MLMI Meeting Recognition Workshop, Edinburgh*, 2005.

[7] NIST, "Rich transcription 2004 spring meeting recognition evaluation plan," *http://nist.gov/speech/tests/rt/rt2004/spring/documents/rt04s-meeting-eval-plan-v1.pdf*, 2004.