A Spectral Clustering Approach to Speaker Diarization

Huazhong Ning, Ming Liu, Hao Tang, Thomas Huang,

Beckman Institute, U. of Illinois at Urbana-Champaign
{hning2,mingliu1,haotang2,huang}@ifp.uiuc.edu

Abstract

In this paper, we present a spectral clustering approach to explore the possibility of discovering structure from audio data. To apply the Ng-Jordan-Weiss (NJW) spectral clustering algorithm to speaker diarization, we propose some domain specific solutions to the open issues of this algorithm: choice of metric; selection of scaling parameter; estimation of the number of clusters. Then, a postprocessing step – "Cross EM refinement" – is conducted to further improve the performance of spectral learning. In experiments, this approach has performance very similar to the traditional hierarchical clustering on the audio data of Japanese Parliament Panel Discussions, but it runs much faster than the latter.

Index Terms: Speaker Diarization, Spectral Clustering, Cross EM refinement, BIC.

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 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 hierarchical clustering along with certain model selection metrics (e.g. BIC) to group the audio segments into an appropriate number of clusters [1, 2]. Tranter and Reynolds presented a hybrid system developed to allow the benefits of their CUED and MIT-LL systems to be exploited in a single system [3]. 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 [4]. Auguera, *et al.* introduced a "purification" module that tries to keep the clusters acoustically homogeneous throughout the hierarchical clustering process [5].

However, to the best of our knowledge, there are only a few methods (e.g. [6]) which use spectral clustering to explore the structure of audio data, especially in speaker diarization. Spectral clustering can handle very complex and unknown cluster shapes in which cases the commonly used methods such as K-means and learning a mixture model using EM may fail. It relies on analyzing the eigen-structure of an affinity matrix, rather than on estimating an explicit model of the data distribution [7, 8, 9]. In this paper, we apply the Ng-Jordan-Weiss (NJW) algorithm [7], a typical spectral clustering approach, to explore the possibility of discovering structure from audio data which is high dimensional and temporal. The affinity matrix is built on the KL distance which is approximated based on unscented transformation [10] to save on computational cost. The scaling parameter is selected by considering the statistics of the distances, and the number of clusters is determined by searching the drop in the magnitude of the eigenvalues or by rotating the normalized eigenvectors [8]. This approach generates results comparable to that of hierarchical clustering but achieves much higher speed than the latter. We also conduct a postprocessing step called "cross EM refinement" (detailed in our previous work [11]) that is based on the idea of cross validation and EM algorithm, which further improves the performance.

The rest of the paper is organized as follows. Section 2 provides the overview of our speaker diarization system. Section 3 describes our solutions to the open issues in spectral clustering so that it can be applied to audio data. Section 4 presents the experimental evaluations, and section 5 compares spectral clustering and hierarchical clustering and 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. Segmentation of each non-silent audio segment into homogenous segments based on the Baysian Information Criteria (BIC) [1, 2, 11]. This segmentation algorithm intends to yield the set of homogenous segments which maximizes the BIC metric.

4. 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.

5. Tied GMM construction [4] to train a background Gaussian mixture model on the entire set of speech segments, and then obtain the GMM coefficients for each speech segment using the EM algorithm while keep the Tied GMM Gaussian components.

6. Local clustering to merge similar adjacent speech segments when their KL distance is above a predefined threshold. This local clustering is based on the observation that it is highly probable that

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

adjacent segments belong to the same speaker.

7. Spectral clustering to group the speech segments into the number of clusters, after carefully building the affinity matrix and estimate the number of clusters.

8. "Cross EM refinement" [11] to refine the spectral clustering result. This is based on the idea of cross validation and EM algorithm.

In step 6 of the above operations, local clustering is introduced before spectral clustering to both improve the performance and reduce the computational cost of spectral clustering. Firstly, the success of spectral clustering depends heavily on the accuracy of the metric measurement, in this paper, the KL distance. Local clustering will increase the average length and produce better estimation of GMMs of the segments, and in turn improve the accuracy of the KL distance. The effect is salient especially when the lengthes of many segments are less than 3 seconds before local clustering. Secondly, local clustering decreases the number of the segments, which in turn reduces the computational cost of spectral clustering because the cost depends only on the number but not the length of the segments. Our experimental evaluations have shown that the number of audio segments will be reduced by nearly 66% after local clustering. Therefore, the average length of the segments increases to greater than 3 seconds, and the total time complexity of spectral clustering decreases to about $1/9$ of the original complexity, compared to the operation without local clustering.

In step 8 of the above operations, we perform "Cross EM refinement" [11] because the local and spectral clustering algorithms generate speaker diarization results which still have large room for improvement due to the following reasons. First, estimation of GMMs of the segments involves approximations which cause errors, especially when the audio segments are very short (less than 3 seconds). Second, the KL distance is approximated based on unscented transformation [10] which introduces errors. Third, the local clustering algorithm may induce some errors as well. The refinement effect is very salient especially when the spectral clustering generates relatively bad results.

3. Spectral Clustering

Spectral clustering can handle very complex and unknown cluster shapes, and in this case the commonly used methods such as K-means and learning a mixture model using EM may fail. It relies on analyzing the eigen-structure of a affinity matrix, rather than on estimating an explicit model of data distribution [7, 8, 9]. It is expected that it can also handle high dimensional audio data. We use a modification of the Ng-Jordan-Weiss (NJW) algorithm [7]. For completeness of the text we first briefly review their algorithm.

Given a set of points (speech segments) $S = \{s_1, \ldots, s_n\}$ that we want to cluster into k subsets:

1. Form the affinity matrix $A \in R^{n \times n}$ defined by $A_{ij} = exp(-d^2(s_i, s_j)/\sigma^2)$ if $i \neq j$, and $A_{ii} = 0$, where $d(s_i, s_j)$ is distance function and σ^2 is scaling parameter.
2. Define D to be the diagonal matrix whose (i, i)-element is the sum of A's i-th row, and construct the normalized affinity matrix $L = D^{-1/2}AD^{-1/2}$.
3. Manually or automatically select the number of clusters k.
4. Find $x_1, x_2, \ldots, x_k$, the k largest eigenvectors of L, and form the matrix $X = [x_1, x_2, \ldots, x_k] \in R^{n \times k}$.
5. Re-normalize the rows of X to have unit length yielding $Y \in R^{n \times k}$, such that $Y_{ij} = X_{ij}/(\sum_j X_{ij}^2)^{1/2}$.
6. Treat each row of Y as a point in R^k and cluster them into k clusters via k-means.
7. Assign the original point s_i to cluster j if and only if row i of the matrix Y was assigned to cluster j.

As to this algorithm, there are still three open issues to be solved: (1) choice of metric, i.e. definition of $d(s_i, s_j)$, and the fast algorithm to calculate it; (2) Selection of the appropriate scale σ; (3) Estimating automatically the number of clusters, i.e, k. We present our domain specific solutions as follows.

3.1. KL Distance

Spectral clustering, as well as most of other clustering methods, depends heavily on the choice of metric. Since the length of the audio segments vary, the commonly used Euclidean metric may fail in this case. And a natural distance measure between two audio segments s_i and s_j is KL distance [10]:

$$KL(f||g) = \int f(x) \log \frac{f(x)}{g(x)} dx \quad (1)$$

$$= \int f \log f dx - \int f \log g dx \quad (2)$$

where distributions f and g are GMMs of s_i and s_j. To make the distance symmetric, we take the KL distance as $d(s_i, s_j) = KL(f||g) + KL(g||f)$.

There are no closed form expression for the KL distance of two GMMs. One approach is Monte-Carlo simulations to approximate it. However, the Monte-Carlo techniques have such drawbacks as extensive computational cost, slow converges properties, different approximations by different computations, and requirement of original data. Therefore the more efficient approximation based on unscented transformation [10] is used in this paper. Unlike the Monte-Carlo approach that chooses sample points randomly, the unscented transformation considers only the "sigma" points.

$$x_{i,j} = \mu_i + (\sqrt{d\Sigma_i})_j \qquad j = 1, \ldots, d \quad (3)$$

$$x_{i,d+j} = \mu_i - (\sqrt{d\Sigma_i})_j \qquad j = 1, \ldots, d \quad (4)$$

where d is the dimension of x, μ_i and Σ_i are the mean and covariance of the i-th component of the GMM, and $(\sqrt{d\Sigma_i})_j$ is the j-th column of $\sqrt{d\Sigma_i}$. Then given $f(x) = \sum_{i=1}^{c} \alpha_i f_i(x)$ is a GMM,

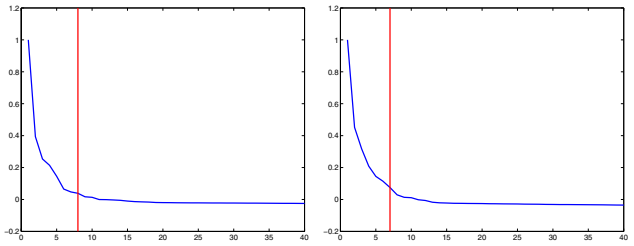
$$\int f \log g dx \approx \frac{1}{2d} \sum_{i=1}^{c} \alpha_i \sum_{j=1}^{2d} \log g(x_{i,k}) \quad (5)$$

Considering that all the GMMs of the segments have the same Gaussian components (components of the Tied GMM), the computational cost can be further reduced by calculating beforehand the probabilities of each Gaussian component at all the sigma points.

3.2. Scale Parameter

The scaling parameter σ^2 is some measure of when two points are considered similar and controls how rapidly the affinity A_{ij} falls off with the distance between s_i and s_j. Usually σ^2 is manually selected as a constant. Ng et al. [7] suggested selecting σ^2 automatically by searching over a range of values of σ^2, and pick the value that gives the tightest clusters of the rows of Y. This increases the computational cost and leaves the range of σ^2 to be specified manually.

Instead of selecting a single constant scaling parameter, we calculate a scaling parameter σ_{ij} for each pair of data points s_i



(a) (b)

Figure 1: Eigenvalues. The top 40 eigenvalues of L of two audio records. The vertical lines indicate the ground truth and locate closely to the drastic drops

and s_j by considering the statistics of the distances from s_i and s_j to all other data points. Thus the affinity between the pair of data points s_i and s_j can be written as:

$$A_{ij} = exp(-\frac{d^2(s_i, s_j)}{\sigma_{ij}^2}) \tag{6}$$

In this paper, we choose σ_{ij} as the multiplication of two variances of two groups of distances

$$\sigma_{ij}^2 = \beta\sqrt{var(d(s_i, \cdot))var(d(\cdot, s_j))} \tag{7}$$

where β is a predefined scalar and $var(\cdot)$ calculates the variance and $d(s_i, \cdot)$ are distances from s_i to all of other data points ($d(\cdot, s_j)$ is similar). The selection of scaling parameters σ_{ij}'s has two advantages: firstly the distances are normalized so that the multiple scale problem can be solved to some extent; secondly it is done fully automatically.

3.3. Estimating the Number of Clusters

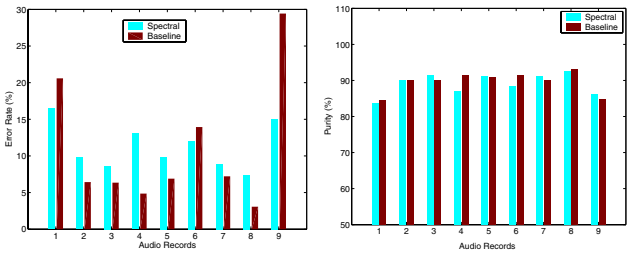
In speaker diarization, there is no prior knowledge available on the number of the speakers in the audio records. It means the number of clusters has to be estimated automatically. This can be done by analyzing the magnitude of the eigenvalues or the structure of the eigenvectors of the normalized affinity matrix L [8].

According to the theory of the spectral clustering [7], the number of clusters should be equal to the multiplicity of the eigenvalue 1 if the data-set is "clean" (the clusters are cleanly separated). However, the audio data in this paper captured in reality is far from "clean", so the top eigenvalues may deviate from 1. A tricky approach is to search for a drastic drop (where gradient is greater than a predefined threshold) in the magnitude of the eigenvalues. Though it lacks theoretical justification, it works well for the audio data in our experiments. Figure 1 gives two examples where the real number of clusters (vertical lines) locate very closely to the drastic drops.

We also try an alternative approach proposed by Zelnik-Manor et al. [8] which relies on analyzing the structure of the eigenvectors. It generates results similar to that of the eigenvalue approach but requires much more computational cost than the latter

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.



(a) (b)

Figure 2: Comparisons of the performance without "Cross EM refinement" of our system and the baseline system . (a) Comparisons of the error rate; (b) Comparisons of the purity.

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 [12] as our evaluation criterion:

$$derr = \frac{T_{falarm} + T_{miss} + T_{wrong}}{T_{ref}} \tag{8}$$

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 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:

$$purity = \frac{pure\ time}{total\ system\ speaker\ time} \tag{9}$$

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 shows the performance of our speaker diarization systems on the nine Japanese Parliament audio records. To reveal the effectiveness of this spectral clustering approach, we have also implemented the speaker diarization system based on traditional hierarchical clustering [11, 4], and it was tested using the same test data set. This implementation is equivalent to the current state-of-the-art speech diarization approaches [4], 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 11.25% and 89.14% respectively for our system with spectral approach. Compared with the average performances of 10.98% and 89.67% for the baseline system, it can be seen that the two systems generate very similar results. Figure 2 uses the bars to illustrate the performance comparisons of the two systems. On average our system is worse than the baseline system by only 0.3% in $derr$ and by only 0.5% in $purity$. However, the standard deviation of the error rate of our system that is 3.09% is much smaller than that of the baseline system which is 8.76%. It means that our system has more stable performance than the baseline system. Moreover, our system is much faster than the

Table 1: Speaker Diarization Error and Purity of both spectral and hierarchical clustering, with and without cross EM refinement.

File Information			Spectral Clustering				Hierarchical Clustering			
			Without EM Refinement		With EM Refinement		Without EM Refinement		With EM Refinement	
file	len(sec.)	#spkrs	error (%)	purity (%)	error (%)	purity (%)	error (%)	purity (%)	error (%)	purity (%)
1	2366	8	16.47	83.77	13.90	86.35	20.59	84.66	14.37	88.00
2	2201	7	9.79	90.21	9.55	90.55	6.43	90.21	6.18	90.51
3	1878	7	8.65	91.54	9.11	91.18	6.35	90.26	5.10	91.07
4	1475	8	13.03	86.97	10.59	89.52	4.86	91.42	5.73	90.61
5	2457	9	9.88	91.32	8.73	91.82	6.90	90.85	4.90	91.75
6	1876	9	12.07	88.57	11.30	89.68	13.94	91.48	6.86	91.03
7	1938	11	8.91	91.09	7.03	93.07	7.22	90.16	6.45	90.95
8	1260	6	7.41	92.59	6.91	93.31	3.07	93.30	3.19	93.20
9	2699	11	15.04	86.24	14.64	86.73	29.43	84.72	26.27	84.51
avg.	2017	8.4	11.25	89.14	10.20	90.25	10.98	89.67	8.78	90.18
std.	464.6	1.74	3.09	2.95	2.73	2.48	8.76	2.99	7.26	2.52

baseline system (see Section 5). Therefore these two methods are really comparable on the Japanese Parliament audio dataset. It is hard to predict which one remarkably outperforms the other.

It worth mentioning that our "Cross EM refinement" does help to improve the performance for both our system and the baseline system. On average it achieves a relative improvement of 10% for $derr$ and 1% for $purity$ in our system, and 21.6% for $derr$ and 1% for $purity$ in the baseline system. The average improvements are not very salient because the results of some audio records (e.g. audio 2, 3, and 8) are already quite good even without the EM refinement. However, for those audio records that the two systems cannot handle well (i.e. audio 1, 4, and 9 for our system, and audio 1, 6 and 9 for the baseline system), the relative improvement is as much as 56% for $derr$ and 2% for $purity$.

5. Discussions and Conclusions

Here we qualitatively compare the computational cost of the spectral clustering approach and of the hierarchical approach. As to the spectral approach, the computational cost of eigen-decomposition and k-mean clustering can be ignored, and most of the cost falls onto the calculation of affinity matrix which after unscented transformation approximation is drastically reduced. While hierarchical clustering needs $O(n^3)$ operations of calculating ΔBIC and estimating GMM [11], therefore it is much slower than our spectral clustering approach.

Spectral clustering is a global approach and optimal with respect to some criteria while hierarchical clustering is a greedy approach and achieves a suboptimal solution. From this point of view, the spectral approach should have higher performance than the hierarchical approach. It is expected to be true if the affinity matrix exactly characterizes the data. However, KL distances are far from accurate if the average length of the segments is too short (< 3 seconds), while hierarchical clustering avoids this problem by accumulating the segments in the iterative steps. In this case, hierarchical clustering may achieve much better performance. Fortunately, thanks to our "Local Clustering" (see Section 2), the average length of the segments is much greater than 3 seconds on the Japanese Parliament audio data, so that our spectral approach has performance very close to that of the hierarchical approach.

In conclusion, we present a spectral learning approach to speaker diarization. Some domain specific solutions are proposed to solve the open issues of spectral clustering, and then applied to the audio records of Japanese Parliament Panel Discussions. It generates similar results as the hierarchical approach but it is much faster than the latter. A postprocessing step ("Cross EM refinement") further improves the performance of our system.

6. Acknowledgment

This work was supported by National Science Foundation Grant CCF 04-26627.

7. References

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

[4] Q. Jin and T. Schultz, "Speaker segmentation and clustering in meetings," *ICSLP*, 2004.

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

[6] D.P.W. Ellis and J.C.Liu, "Speaker turn segmentation based on between-channel differences," *NIST Meetings Workshop*, 2004.

[7] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," *NIPS*, 2002.

[8] L. Zelnik-Manor and P. Perona, "Self-tuning spectral clustering," *NIPS*, 2004.

[9] F.R. Bach and M.I. Jordan, "Learning spectral clustering," *NIPS*, 2004.

[10] J. Goldberger and H. Aronowitz, "A distance measure between gmms based on the unscented transform and its application to speaker recognition," *Proc. of Interspeech*, 2005.

[11] H. Ning, W. Xu, Y. Gong, and T. Huang, "Improving speaker diarization by cross em refinement," *ICME*, 2006.

[12] NIST, "Rich transcription 2004 spring meeting recognition evaluation plan," *http://www.nist.gov/speech/*, 2004.