CLUSTER CRITERION FUNCTIONS IN SPECTRAL SUBSPACE AND THEIR APPLICATION IN SPEAKER CLUSTERING

Trung Hieu Nguyen, Haizhou Li

Institute for Infocomm Research,
Department of Human Language Technology,
1 Fusionopolis Way, #21-01 Connexis,
South Tower, Singapore 138632

Eng Siong Chng

Nanyang Technological University,
School of Computer Engineering,
Block N4, Nanyang Avenue,
Singapore 639798

ABSTRACT

In this paper, we propose two cluster criterion functions which aim to maximize the separation between intra-cluster distances and inter-cluster distances. These criteria can automatically deduce the desired number of clusters based on their extremized values. We then propose an algorithm to apply our criterion functions in conjunction with spectral clustering. By exploiting the characteristic of spectral subspace,we show that the speakers are more separable in this subspace which will further enhance the effectiveness of our proposed criteria. The algorithm is used in our agglomerative hierarchical speaker diarization system to test on Rich Transcription 2007 conference data set and obtains very good results.

Index Terms— speaker diarization, criterion function, spectral clustering

1. INTRODUCTION

Clustering is the procedure to group data points into clusters such that the data points in the same cluster possess strong internal similarities. Generally, there are two major issues in clustering: determining number of clusters (cluster validity) and finding optimal partitioning (cluster criteria). Thus far, these two issues are handled separately with different criteria e.g. the sum-of-squared-error criterion cannot be used for cluster validity because it is monotonic decreasing with increasing number of clusters. In this paper, we propose two cluster criterion functions when extremized could concurrently solve both issues. These functions have a simple interpretation that they aim to maximize the separation between intra-cluster distances and inter-cluster distances.

Recently, spectral clustering methods get much attention because of their ability to handle many difficult clustering problems. However, not much has been investigated for speaker clustering within this framework. In this paper, we introduce an algorithm using our proposed criterion functions in spectral subspace and provide a mathematical analysis to this algorithm in the ideal case. Furthermore, we also show in the experiment that the speakers are more separable in the spectral subspace which is a desirable property for clustering. We then demonstrate the use of this algorithm in our agglomerative hierarchical speaker diarization system to estimate number of speakers. This approach has advantage compared to those using thresholds derived from development set to determine number of speakers [1, 2, 3] because it does not suffer from mismatch issues between development data and test data. Ajmera [4] proposed a system using a modified version of BIC. This system performs well in terms of having low diarization error rate (DER) and not requiring development data, however it usually generates many small clusters (which does not have much impact on DER) thus provides wrong number of speakers.

The paper is organized as follow: first we introduce two criterion functions in section 2, and then in section 3, we apply these functions in spectral subspace and provide detail analysis. We finally report some experimental results on speaker clustering using the proposed algorithm in section 4.

2. CLUSTERING CRITERION FUNCTIONS

Given a set of point $S = \{s_1, s_2, \ldots, s_n\}$ of n samples that we want to partition into c disjoint subsets $S_1, \ldots, S_c$. Let $d(s_i, s_j)$ be the similarity function between two points s_i and s_j. Define:

$$\begin{aligned} D_{intra} &= \{d(s_i, s_j) \,|\forall i, j \; \exists k : s_i \in S_k, s_j \in S_k\} \\ D_{inter} &= \{d(s_i, s_j) \,|\forall i, j \; \exists k \neq l : s_i \in S_k, s_j \in S_l\} \end{aligned}$$

We propose two criterion functions to measure the quality of partitioning.

2.1. T_s criterion

Let $m_1, \sigma_1, n_1, m_2, \sigma_2, n_2$ be respectively the mean, standard deviation, size of D_{intra} and D_{inter}.

$$T_s\left(D_{intra}, D_{inter}\right) = \frac{|m_2 - m_1|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \quad (1)$$

The interpretation of this criterion function: T_s measures the difference between the mean distance among points in the same clusters and the mean distance among points across different clusters taking into account the variance of these distances. The optimal partitioning is defined as one that maximizes T_s.

2.2. ρ criterion

This criterion function is based on Mann-Whitney U test [5]. First, we sort the value of $D = \{D_{intra} \cup D_{inter}\}$ in ascending order and assign a ranking order for each element of D.

$$\begin{aligned} R_1 &= \sum_{x_i \in D_{intra}} rank(x_i) \\ U_1 &= R_1 - \frac{||D_{intra}||(||D_{inter}|| + 1)}{2} \\ \rho &= \left| \frac{U_1}{||D_{intra}||.||D_{inter}||} - 0.5 \right| \times 2 \end{aligned} \quad (2)$$

where $rank(x_i)$ is the order of x_i in the sorted sequence of D, $||.||$ is the cardinal of the set. ρ can take values between 0 and 1.A ρ of 0 represents complete overlap while a value of 1 represent complete separation. This criterion function has a simple interpretation: ρ measures the overlap between the set of distances among points in the same clusters and the set of distances among points across different clusters based on the ranking order of these distances, the actual values of distances are not important. The optimal partitioning is defined as one that maximizes ρ.

3. SPECTRAL SUBSPACE AND CLUSTERING CRITERION FUNCTIONS

In this section, we propose an algorithm to measure clustering quality using the above mentioned functions in spectral subspace.

3.1. Algorithm

Given a set of point $S = \{s_1, s_2, \ldots, s_n\}$ of n samples that we want to partition into c disjoint subsets $S_1, \ldots, S_c$. Let $d\left(s_i, s_j\right)$ be the similarity function between two points s_i and s_j.

1. Form the affinity matrix $A \in R^{n \times n}$ defined by

$$A_{ij} = exp\left(\frac{-d^2\left(s_i, s_j\right)}{\sigma_i \sigma_j}\right) \quad (3)$$

   where $\sigma_i = d\left(s_i, s_K\right)$ is the distance from point s_i to its K'th neighbor. In all our experiments, a single value of $K = 7$ is used as suggested in [6].

2. Define D to be a diagonal matrix with $D\left(i,i\right) = \sum_{j=1}^{N} A_{ij}$ and construct the normalized affinity matrix $L = D^{-1/2} A D^{-1/2}$.

3. Find $x_1, x_2, \ldots, x_c$, the c largest eigenvectors of L (largest eigenvectors are those corresponding to largest eigenvalues), and form the matrix $X = [x_1 x_2 \ldots x_c] \in R^{n \times c}$.

4. Renormalize X to obtain matrix Y such that $Y_{ij} = \frac{X_{ij}}{\sqrt{\sum_{j=1}^{n} X_{ij}^2}}$.

5. Form the matrix $Z \in R^{n \times n}$ in which each element of Z is the cosine distance between rows i, j of Y: $Z\left(i,j\right) = \sum_{k=1}^{n} Y_{ik} Y_{jk}$.

6. Define D_{intra} and D_{inter} respectively as the set of intra-cluster distances and inter-cluster distances

$$D_{intra} = \{Z\left(i,j\right) | \forall i, j \; \exists k : s_i \in S_k, s_j \in S_k\}$$

$$D_{inter} = \{Z\left(i,j\right) | \forall i, j \; \exists k \neq l : s_i \in S_k, s_j \in S_l\}$$

   Now, we can compute the partitioning quality with either formula (1) or formula (2).

3.2. Analysis

This section will provide an analytical analysis of our algorithm in the ideal case where $A_{ij} = 0$ if s_i and s_j are in different clusters and $A_{ij} > 0$ otherwise. In other word, the points in different clusters are assumed to be infinitely far apart. Without loss of generality, we suppose that there are $c = 3$ actual clusters for ease of discussion. The matrix A and matrix L are block-diagonal matrices:

$$L = \begin{pmatrix} L^{(1)} & 0 & 0 \\ 0 & L^{(2)} & 0 \\ 0 & 0 & L^{(3)} \end{pmatrix}$$

where $L^{(k)}$ is the sub-matrix corresponding to cluster S_k.

3.2.1. The actual number of clusters is known

In this case

$$X = \begin{pmatrix} x^{(1)} & 0 & 0 \\ 0 & x^{(2)} & 0 \\ 0 & 0 & x^{(3)} \end{pmatrix}$$

where $x^{(k)}$ is an eigenvector of the sub-matrix $L^{(k)}$ (please refer to [6, 7] for detail derivations). When we renormalize each of X's rows to have unit length, we obtain:

$$Y = \begin{pmatrix} \vec{1} & 0 & 0 \\ 0 & \vec{1} & 0 \\ 0 & 0 & \vec{1} \end{pmatrix}$$

When each point s_i belongs to its true cluster, we have: D_{intra} contains all 1s and D_{inter} contains all 0s. Thus, $T_s = +\infty$ and $\rho = 1$. In all other cases where there is at least one point in a wrong cluster, D_{intra} and D_{inter} both contain some 0s and some 1s which will lead to $T_s < +\infty$ and $\rho < 1$.

3.2.2. The estimated number of clusters is less than the actual number of clusters

Supposed $c' = 2$ largest eigenvectors are selected.

$$Y' = \begin{pmatrix} \vec{1} & 0 \\ 0 & \vec{1} \\ 0 & 0 \end{pmatrix}$$

In this case, D_{intra} contains some 0s and some 1s no matter what how we cluster the data into 2 clusters. Hence, $T_s < +\infty$ and $\rho < 1$.

3.2.3. The estimated number of clusters is more than the actual number of clusters

Supposed $c'' = 4$ largest eigenvectors are selected.

$$Y'' = \begin{pmatrix} \vec{a} & 0 & 0 & \vec{b} \\ 0 & \vec{1} & 0 & 0 \\ 0 & 0 & \vec{1} & 0 \end{pmatrix}$$

where $\vec{a}$, $\vec{b}$ are two eigenvectors from the same sub-matrix, $\vec{a} \cdot \vec{b} = 0$ and at least one of them is strictly positive [7]. No matter what how we cluster the data into 4 clusters, D_{intra} has some 1s and some less than 1s, D_{inter} has some 0s and some greater than 0s. Thus, $T_s < +\infty$ but we cannot conclude about the value of ρ in this case although in practice it works well as shown later in the experiments.

4. EXPERIMENTS IN SPEAKER CLUSTERING

4.1. Similarity measure between two segments

In all our experiments, we used T_d as the similarity measure between two segments (clusters). Define:

$$\begin{aligned} f_1(x) &= logL(x|\lambda_1) - logL(x|\lambda_{UBM}) \\ f_2(x) &= logL(x|\lambda_2) - logL(x|\lambda_{UBM}) \\ S_1 &= \{f_1(x)|\forall x \in X_1\} \cup \{f_2(x)|\forall x \in X_2\} \\ S_2 &= \{f_1(x)|\forall x \in X_2\} \cup \{f_2(x)|\forall x \in X_1\} \end{aligned}$$

where X_1, X_2 are the sets of feature vectors from two segments;λ_1, λ_2 are the models estimated from X_1, X_2; λ_{UBM} is the universal background model. Let m_1, m_2, σ_1, σ_2, n_1, n_2 are respectively the mean, standard deviation, size of S_1 and S_2:

$$T_d = \frac{|m_1 - m_2|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \quad (4)$$

Note that although T_s and T_d have the same formula but they are contextually different. T_s is used as the stopping criterion; it is applied on the sets of intra-cluster and inter-cluster distances. T_d is used as the metric to determine which two closest clusters should be merged in hierrachical clustering; it is applied on the sets of frame log-likelihood scores.

4.2. Why spectral subspace?

For many problems in other domains [6, 7], the data points when mapped to spectral subspace will form tight groups and the groups are well separated which is a desirable characteristic for clustering. This experiment was carried out to answer the question whether the mapping can be applied for speaker clustering?

We extract 48 speakers from NIST Speaker Recognition 2004 in which 34 of them are females. Each speaker consists of several segments of 20 seconds audio data. We randomly select $4 - 10$ speakers from this group and compute the distances between every pair of speaker segments to form D_{intra}, D_{inter} in the original space as described in section 2 or in the spectral subspace as in section 3.1. The separation between D_{intra} and D_{inter} is then computed using T_s, ρ or equal error rate (EER) [8]. These procedures are repeated 100 times and the results are averaged and recorded in Table 1.

Table 1. *Separation between D_{intra} and D_{inter}.*

	T_s	ρ	$EER(\%)$
Original space	60.66	0.9851	2.74
Spectral subspace	607.22	0.9968	0.49

The obtained results have clear indications that the speakers are more separable in the spectral subspace, .

4.3. Agglomerative hierarchical speaker diarization

We demonstrate the use of the proposed algorithm in our agglomerative hierarchical speaker diarization system. Suppose from the system, we obtain $C^{(p)}, C^{(p-1)}, \ldots, C^{(2)}$ where $C^{(k)} = \{S_1, \ldots, S_k\}$ is a partitioning of data into k disjoint clusters. To determine the desired number of clusters, we measure the partitioning quality of $C^{(k)}$ and select the value of k which maximizes the quality. The procedure to compute the quality of $C^{(k)}$ is summarized as follow:

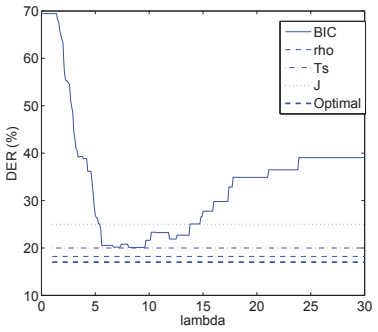
1. For each cluster S_i from $C^{(k)}$, S_i is divided uniformly into n_i segments of 10 seconds each: $S_i = \{s_{i1}, \ldots, s_{in_i}\}$. In total, there are $n = n_1 + \ldots + n_k$ segments.
2. Construct an affinity matrix $A \in R^{n \times n}$ as (3) using T_d as the distance metric between each pair of segments.
3. Follow the steps described in section (3.1) to compute the partitioning quality.

The system was tested on Rich Transcription (RT) 2007 conference data released by NIST for RT07 benchmark on the single distance microphone condition. The experimental results are reported in Table 2 with three evaluation criteria: diarization error rate (DER) [9], miss speakers and false alarm (FA) speakers. We also evaluate the alignment cost function J [6] to estimate the number of clusters.

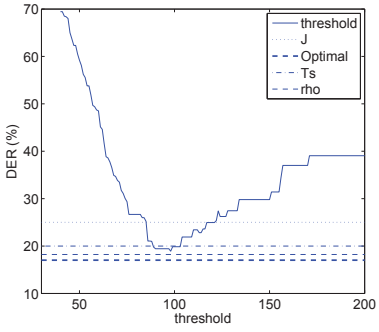
Table 2. *Results of speaker diarization system.*

Criterion Function	DER (%)	Miss speaker	FA speaker
ρ	18.22	4	1
T_s	19.99	6	1
J	25.00	9	1

In Figure 1, the proposed criterion functions are compared with Bayesian Information Criterion (BIC) and merging threshold. The experimental result shows that for this database: T_s is comparable with the best threshold and best possible value of λ while ρ is better than all other methods and it is approaching the optimal stopping criterion of our speaker diarizaton system. The optimal stopping criterion is the one that produces lowest diarization error rate (no other stopping criterion could be better than this limit).



(a) BIC



(b) Threshold

Fig. 1. DER with BIC and threshold as the stopping criterion

5. CONCLUSION

We have proposed two clustering criterion functions to measure partitioning quality as well as determine number of speakers. These functions are applied in spectral clustering framework and incorporated into our agglomerative speaker diarization system. The system performs very well on RT 2007 dataset when compared with state-of-the-art systems [10]. We have also shown that the speakers are more separable in spectral subspace.

6. REFERENCES

[1] H. Jin, F. Kubala, and R. Schwartz, "Automatic speaker clustering," *Proceedings of the DARPA Speech Recognition Workshop*, pp. 108–111, 1997.

[2] M. Siegler, U. Jain, B. Raj, and R.M. Stern, "Automatic segmentation, classification and clustering of broadcast news audio," *Proc. DARPA Speech Recognition Workshop*, pp. 97–99, 1997.

[3] D. van Leeuwen, "The TNO speaker diarization system system for NIST RT05s for meeting data," *NIST 2005 Spring Rich Transcrition Evaluation Workshop, Edinburgh, UK*, 2005.

[4] J. Ajmera and C. Wooters, "A robust speaker clustering algorithm," *Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on*, pp. 411–416.

[5] W. J. Conover, *Practical Nonparametric Statistics*, John Wiley & Sons, Inc., 1999.

[6] L. Zelnik-Manor and P. Perona, "Self-tuning spectral clustering," *Advances in Neural Information Processing Systems*, vol. 17, no. 1601-1608, pp. 16, 2004.

[7] A. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in Neural Information Processing Systems 14: Proceedings of the 2002 [sic] Conference*. MIT Press, 2002.

[8] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET Curve in Assessment of Detection Task Performance," *Fifth European Conference on Speech Communication and Technology*, 1997.

[9] NIST, "Spring 2007 (RT-07) Rich Transcription Meeting Recognition Evaluation Plan, http://www.nist.gov/speech/tests/rt/2007/docs/rt07-meeting-eval-plan-v2.pdf008," .

[10] NIST, "The Rich Transcription 2007 Speaker Diarization Results, http://www.nist.gov/speech/tests/rt/2007/workshop/RT07-SPKR-v7.pdf," .