A COMPARATIVE STUDY OF SPECTRAL CLUSTERING FOR I-VECTOR-BASED SPEAKER CLUSTERING UNDER NOISY CONDITIONS

Naohiro Tawara, Tetsuji Ogawa and Tetsunori Kobayashi

Department of Computer Science, Waseda University, Tokyo, Japan

ABSTRACT

The present paper dealt with speaker clustering for speech corrupted by noise. In general, the performance of speaker clustering significantly depends on how well the similarities between speech utterances can be measured. The recently proposed i-vector-based cosine similarity has yielded the state-of-the-art performance in speaker clustering systems. However, this similarity often fails to capture the speaker similarity under noisy conditions. Therefore, we attempted to examine the efficiency of spectral clustering on i-vector-based similarity for speech corrupted by noise because spectral clustering can yield robustness against noise by non-linear projection. Experimental comparisons demonstrated that spectral clustering yielded significant improvement from conventional methods, such as agglomerative clustering and k-means clustering, under non-stationary noise conditions.

Index Terms— spectral clustering, i-vector, noise-robust speaker clustering.

1. INTRODUCTION

Speaker clustering is a technology to estimate which utterances are from the same speaker, and it has been an important role in speaker diarization. Two main approaches have been taken in speaker clustering; the bottom-up and top-down approaches. In the bottom-up approach, utterances are clustered by iteratively merging the most similar pair of clusters until a stopping criterion is met (e.x. Bayesian information criteria [1]). In the top-down approach, the utterances are clustered by identifying the most appropriate cut that separates the dissimilar clusters.

In both approaches, the clustering performance significantly depends on how accurately the speaker similarity can be measured. Accurate estimation of the similarity between speech utterances, however, is difficult when the speech data are corrupted by background noise. This difficulty arises because the similarity is significantly affected by the degree of similarity between background noises as well the degree of between speakers, which can make the speaker similarity less confident. A noise-robust similarity measurement, therefore, is required to achieve noise-robust speaker clustering.

Cosine similarity between i-vectors has yielded reasonable performance in speaker recognition systems because the i-vectors are separably distributed on the unit hypersphere after being projected onto the discriminative space obtained by linear discriminative analysis (LDA) and within-class covariance normalization (WCCN) [2]. The more sophisticated probabilistic LDA (PLDA) is also employed for scoring [3]. In addition, the i-vector-based approach is shown to be applicable to noisy data after using noisy data for training the discriminative space [4].

The i-vector approach has also been applied to speaker clustering problems [5, 6]. In speaker clustering, an utterance is represented by an i-vector, and k-means clustering based on cosine similarity is applied for clustering those vectors. Spectral clustering was evaluated as an alternative to k-means clustering [7, 8, 6]. Generally, spectral clustering works better than k-means clustering because of its capability of classifying data with a type of manifold embedding. However, the authors of [6] concluded that the simple k-means clustering algorithm provides a sufficiently high accuracy because i-vectors are linearly separable on the unit hypersphere, and the cosine distance is, hence, a valid measurement. The assumption that i-vectors are separably distributed on the unit hypersphere is not always true under noise conditions, because the i-vector contains not only speaker information but also noise information, which cannot be perfectly eliminated by front-end processing methods such as LDA and WCCN. In the present paper, therefore, the effectiveness of spectral clustering is investigated for various types of noisy speech utterances. The results obtained in the present study can be useful for developing noise-robust speaker clustering and diarization systems.

The rest of the present paper is organized as follows. Section 2 briefly reviews i-vector extraction. In section 3, some examples of the failure of the conventional i-vector-based approach and the purpose of employing spectral clustering are presented. Section 4 provides a brief explanation of spectral clustering and indicates the effectiveness of this approach for speech data corrupted by noise. In Section 5, experimental comparisons are conducted to demonstrate the effectiveness of spectral clustering for noisy speech. Section 6 concludes this paper and discusses some directions for future work.

2. I-VECTOR AND COSINE SIMILARITY SCORING

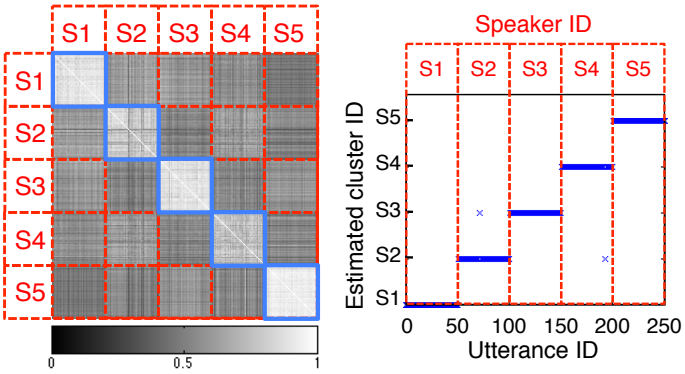
For the i-th utterance, the speaker- and channel-dependent GMM supervector $\boldsymbol{m}_i$ is written by the factor analysis model as

$$\boldsymbol{m}_i = \boldsymbol{m}_0 + \boldsymbol{T}\boldsymbol{x}_i, \tag{1}$$

where $\boldsymbol{m}_0$ denotes the speaker- and channel-independent GMM (universal background model; UBM) supervector; $\boldsymbol{T}$, a rectangular low-rank matrix, represents the total variability; and $\boldsymbol{x}_i$ is a random vector having a standard normal distribution. The vector $\boldsymbol{x}_i$ is called the "i-vector" and is calculated using an EM algorithm. i-vectors contain speaker information as well as intra-speaker variability derived from difference in channels, phoneme contexts, and environmental noises. Such nuisance information, however, was successfully eliminated by using LDA and WCCN [2]. The cosine distance is applied to measure the similarity between a pair of i-vectors, which is extracted as follows:

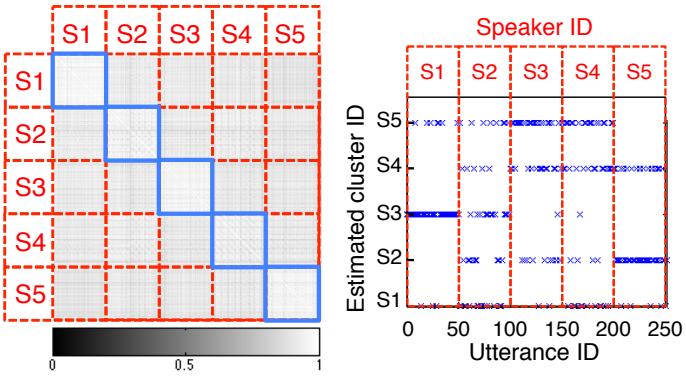
$$\cos(\boldsymbol{x}_i, \boldsymbol{x}_j) = \frac{\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_j}{||\boldsymbol{x}_i|| \cdot ||\boldsymbol{x}_j||}, \tag{2}$$

where $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ denote the i-vectors extracted from the i-th and j-th utterances, respectively.



(a) Clean

(b) Clean



(c) Noisy

(d) Noisy

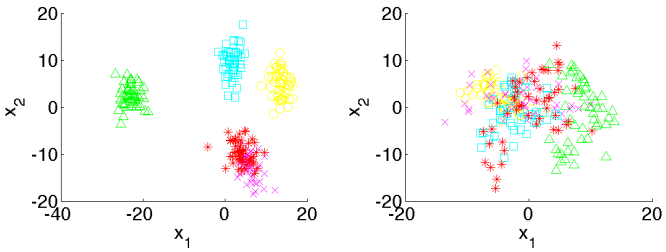
Fig. 1. Similarity matrix obtained from (a) clean and (c) noisy utterances. Clustering result obtained by applying k-means clustering on i-vectors from (b) clean and (d) noisy utterances.

3. SPEAKER CLUSTERING UNDER NOISY CONDITIONS

We compare the similarity matrices calculated from clean and noisy speech utterances to investigate the effect of noise on the cosine similarity score extracted from i-vectors.

Fig. 1 (a) depicts the similarity matrix of 500 utterances from five speakers in a clean environment. Fig. 1 (c), on the other hand, depicts the similarity matrix of the same sentences with non-stationary noise overlapped. In these figures, five rectangles drawn with dashed lines, S1 to S5, correspond to each speaker's utterances, and regions with blue dashed lines indicate the similarity between utterances from the same speaker. These figures indicate that, for clean utterances, i-vector-based similarity between the same speaker's utterances is larger than that between utterances from different speakers; these similarities are almost equivalent for noisy utterances. Figs. 1 (b) and (d) visualize the speaker clustering results obtained using k-means clustering for clean and noisy utterances, respectively. These results indicate that i-vector-based similarity can fail to capture the speaker similarity for noisy utterances even when it can for the corresponding clean utterances. Therefore, i-vector-based similarity is insufficient to measure the speaker similarity for noisy utterances.

Further analysis is performed by investigating the distribution of i-vectors. Figs. 2 (a) and (b), respectively, depict distributions of i-vectors obtained from clean and noisy utterances from five speakers. Each i-vector is projected onto two-dimensional space by LDA/WCCN projection followed by principal component analysis. In each figure, different colors correspond to different speakers.



(a) Clean (b) Noisy

Fig. 2. The i-vectors of five speakers after LDA/WCCN projection onto two-dimensional space. Each color corresponds to speaker.

Algorithm 1 Algorithm of Ng-Jordan-Weiss spectral clustering [9].

1: Calculate cosine-based distance between all pair of i-vectors $D(\boldsymbol{x}_i, \boldsymbol{x}_j) = 1 - \cos(\boldsymbol{x}_i, \boldsymbol{x}_j), \forall i, j$.
2: Calculate adjacency matrix $\boldsymbol{W} \in \mathbb{R}_+^{n \times n}$, where $(\boldsymbol{W})_{ij} = \exp\{-D(\boldsymbol{x}_i, \boldsymbol{x}_j)\}$ for $i \neq j$ and zero otherwise.
3: Calculate the diagonal matrix $\boldsymbol{D}$ whose (i,i)-th components is sum of i-th row of $\boldsymbol{W}$ (i.e. $(\boldsymbol{D})_{ii} = \sum_{j=1}^{n} w_{ij}$), and construct the graph Laplacian $\mathbf{L} = \boldsymbol{I} - \boldsymbol{D}^{-\frac{1}{2}} \boldsymbol{W} \boldsymbol{D}^{-\frac{1}{2}}$.
4: Select $\boldsymbol{t}_1, \cdots, \boldsymbol{t}_K$, K smallest eigenvectors of $\mathbf{L}$ and form $\boldsymbol{T} = [\boldsymbol{t}_1, \cdots, \boldsymbol{t}_K] \in \mathbb{R}^{n \times K}$.
5: Normalize each row of $\boldsymbol{T}$ to have unit length (i.e. $\{\hat{\boldsymbol{T}}\}_{ij} = \{\boldsymbol{T}\}_{ij} / (\sum_k t_{ik}^2)^{1/2}$)
6: Cluster row vectors of $\hat{\boldsymbol{T}}$ via cosine similarity-based k-means clustering.

These figures clearly show that the distribution of each speaker's utterances spread over a large and complex region in the case of noisy utterances, whereas it spreads over a relatively small region in the case of clean utterances. This result also explains the insufficiency of k-means clustering: it assumes that each speaker's utterances are sampled from a Gaussian distribution.

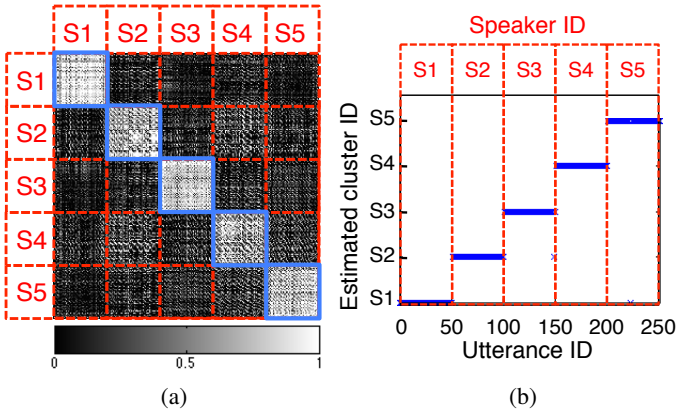
4. SPECTRAL CLUSTERING

Spectral clustering on i-vector-based similarity is carried out to handle noise corruption in speaker clustering. Here, we describe a brief explanation of spectral clustering and how it works for speech corrupted by noise.

Spectral clustering is a top-down approach to determine the optimal assignment of utterances to speakers; it assumes any pair of samples in the same cluster has high similarity, while those from a different cluster should have low similarity. Assume that an indicator vector $\boldsymbol{t}_i = [t_{i1}, \cdots, t_{ij}, \cdots, t_{in}]^{\mathrm{T}} \in \mathbb{R}^n$ represents the assignment of the j-th sample to the i-th cluster. A component t_{ij} is equal to $1/\sqrt{(\boldsymbol{D})_{ii}}$ if the j-th sample is assigned to the i-th cluster and equal to zero otherwise. For any indicator vector $\boldsymbol{t}_i$, we have

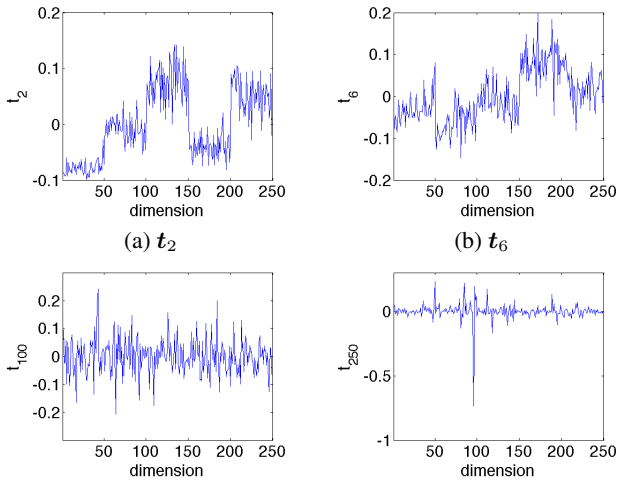
$$\boldsymbol{t}_i^{\mathrm{T}} \boldsymbol{L} \boldsymbol{t}_i = \frac{1}{2} \sum_{j=1}^{n} \sum_{j'=1}^{n} w_{jj'} (t_{ij} - t_{ij'})^2, \tag{3}$$

where $w_{jj'}$ denotes the similarity between the j-th and j'-th utterances and n denotes the number of utterances. Eq. 3 indicates that $\boldsymbol{t}_i^{\mathrm{T}} \boldsymbol{L} \boldsymbol{t}_i$ will be small if two samples with large similarity (i.e., $w_{jj'}$ is large) have similar coordinates (i.e., t_{ij} and $t_{ij'}$ are close). This implied that the indicator vector $\boldsymbol{t}_i$ obtained by minimizing Eq. 3 under



(a) (b)

Fig. 3. (a) Similarity matrix calculated from normalized eigenvectors of the Laplacian matrix of noisy utterances of five speakers. (b) Clustering result obtained by k-means clustering using the eigenvectors-based features.



(a) $\boldsymbol{t}_2$ (b) $\boldsymbol{t}_6$

(c) $\boldsymbol{t}_{100}$ (d) $\boldsymbol{t}_{250}$

Fig. 4. The (a) second, (b) sixth, (c) 100th and (d) 250th smallest eigenvectors of the Laplacian matrix calculated from noisy utterances.

the constraint of all indicator vectors being orthogonal can indicate a valid clustering result in which pairs of utterances from the same speaker have a large similarity and those from the different speakers have small similarity. The solution of this minimization problem is obtained as the K smallest eigenvectors of $\boldsymbol{L}$. A more detailed analysis of this algorithm is presented in [9].

Fig. 3 (a) depicts the similarity matrix of noise-corrupted utterances calculated using the smallest 20 normalized eigenvectors of the Laplacian matrix $\boldsymbol{L}$. This figure shows that these eigenvectors are reasonable features to measure the similarity in speakers because the similarity thus obtained between the same speaker's utterances is higher than that from different speakers. Fig. 3 (b) depicts the clustering result obtained by k-means clustering on the eigenvectors-based features. This figure shows that spectral clustering can perfectly cluster the utterances even when they are corrupted by noise.

Next, we present a more detailed analysis of the effectiveness of eigenvector-based features for measuring the speaker similarity under the noise conditions. The spectral clustering utilizes the K smallest eigenvectors of the Laplacian matrix as features of each utterance. Therefore, we have

$$\boldsymbol{L} = \boldsymbol{T}\boldsymbol{\Lambda}\boldsymbol{T}^{\mathrm{T}} = \sum_{i=1}^{n} \lambda_i \boldsymbol{t}_i \boldsymbol{t}_i^{\mathrm{T}}, \tag{4}$$

where $\boldsymbol{\Lambda} = [\lambda_1, \cdots, \lambda_N]$ and $\boldsymbol{T} = [\boldsymbol{t}_1 \cdots, \boldsymbol{t}_N]$ denote the eigenvalues and eigenvectors of the Laplacian matrix, respectively. Roughly speaking, the Laplacian matrix of the noisy utterances can be factorized into two matrices: the similarity matrix primarily obtained from speech signals and that from noise signals. Here, there exists a clear pattern that intra-speaker similarity is relatively higher than inter-speaker similarity. The similarly in terms of noise signals, on the other hand, is consistently low because there are no correlations between noise signals. Considering this fact and that $0 = \lambda_1 < \lambda_2, \cdots < \lambda_n$, smaller eigenvectors should correspond to a pattern representing speaker similarity, whereas larger eigenvectors should correspond to a pattern representing similarity in noise.

Fig. 4 depicts the second, sixth, 100th, and 250th smallest eigenvectors of the Laplacian matrix of noisy utterances. This figure shows that the smaller eigenvectors restore speakers' similarity patterns while the larger ones restore the similarity patterns of noise.

5. SPEAKER CLUSTERING EXPERIMENTS

Experimental comparisons were performed to demonstrate robustness of spectral clustering against noise. For that purpose, the following three methods were evaluated under various types of noise.

- **GMM-HAC**: Agglomerative clustering in which each cluster is modeled as a Gaussian mixture model (GMM). Similarity between clusters is defined as a cross likelihood ratio between these GMMs [10].
- **IV-KMEANS**: k-means clustering using cosine distance between i-vectors [5, 6].
- **IV-SC**: Spectral clustering using cosine distance between i-vectors.

The present experiments were conducted using the corpus of spontaneous Japanese (CSJ) [11].

5.1. Experimental setups

The clean and noisy speech utterances from CSJ were used for evaluation. Note that speech data from CSJ are basically uncorrupted by noise. The clean evaluation sets were constructed as follows. All of the lecture speech in CSJ were divided by the utterance on the basis of silence. Then, 10 speakers were randomly selected. Finally, their 50 utterances were randomly selected. Each utterance is from both the same and different lectures. Four combinations of different speakers yielded four evaluation sets and the resulting performance was the average over those four sets.

In addition, noisy speech data were developed by overlapping each utterance with seven types of noise at the signal-to-noise ratio (SNR) of about 0 dB. The noise includes four types of environmental noise (Crowd, Party, Street, and Station) sampled from JEIDA noise database [12] and three types of background music sampled from RWC music databases [13]. Note that crowd and party noises are stationary while street and station noises are non-stationary. The other experimental setups for evaluating noisy speech were the same as for clean speech.

The evaluation criteria was the K value, which is the geometric mean of the average speaker purity and average cluster purity [14].

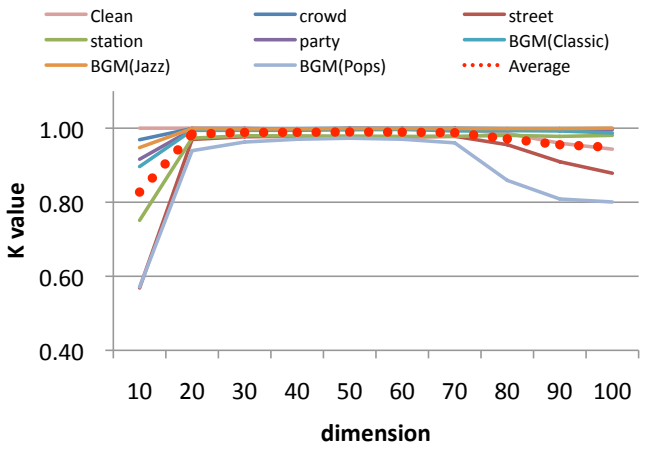


Fig. 5. Clustering accuracy as a function of number of eigenvectors.

5.2. Front-end processing

Acoustic feature parameters consisted of 12-dimensional mel-frequency cepstral coefficients (MFCCs) plus log-energy and their delta parameters, yielding a 26 dimensional vector for every 10 ms. A gender-independent UBM of 128 Gaussians with diagonal covariance matrices was trained on the speech data taken from the Japanese newspaper article sentence (JNAS) [15] and Continuous Speech Corpus for Research (ASJ-JIPDEC) databases [16]. Those speech data were overlapped with four types of noise (air conditioner, car, factory, and plant) from JEIDA noise database and the total variability, LDA, and WCCN matrices were trained on those noisy data. The 150-dimensional i-vectors were extracted and finally projected onto 100-dimensional vectors.

5.3. Experimental results

5.3.1. Number of Eigenvectors

In ideal spectral clustering, samples from different clusters are infinitely far apart, yielding always zero-similarity between those samples. In this case, the eigenvectors of the Laplacian matrix are consistent to the optimal indicator vectors and the optimal number of eigenvectors coincides the number of clusters. However, this assumption is not always true in the noise conditions because the similarity between distinct speakers' utterances can be large. Figure 5 depicts the clustering accuracy as a function of the number of eigenvectors. Eight solid lines describe the clustering accuracy obtained from spectral clustering in clean and seven noise conditions. This figure shows that the highest clustering accuracy was achieved when the number of eigenvectors was larger than that of speakers. This was noticeable particularly in the noise conditions because further eigenvectors are required to recover the Laplacian matrix from the noisy utterances. The experiments below used 50 eigenvectors providing the highest performance for all conditions.

5.3.2. Clustering Accuracy

Table 1 lists the K values obtained using three clustering methods for clean and noisy speech data. This result demonstrates that the clean utterances were almost perfectly clustered, irrespective of methods. However, the agglomerative clustering (GMM-HAC) and k-means clustering on i-vectors (IV-KMEANS) consistently deteriorated clustering accuracy in noisy conditions and yielded significant degradation particularly in the non-stationary and BGM (pops) noise conditions. Note that IV-KMEANS outperformed GMM-HAC in the stationary noise conditions but not in the non-stationary noise conditions. The i-vector-based similarity is therefore sufficient to handle the stationary noise corruptions but insufficient for the non-stationary noise. In contrast, spectral clustering on i-vectors (IV-SC) worked in both stationary and non-stationary noise conditions with small or almost no degradation in clustering performance compared with the clean conditions. We also evaluated these methods on relatively short utterances. Here, the average duration is about 10 seconds. The result is shown in Table 2, and demonstrates that IV-SC also outperformed GMM-HAC and IV-KMEANS for these short utterances.

Table 1. K values obtained from Speaker clustering experiment. Average duration of each utterance is about 20 seconds.

Environment		GMM-HAC	IV-KMEANS	IV-SC
Clean		0.955	**1.000**	**1.000**
Stationary noise	Crowd	0.906	0.997	**1.000**
	Party	0.907	0.958	**0.999**
Non-Stationary noise	Street	0.425	0.540	**0.976**
	Station	0.591	0.591	**0.979**
BGM	Classic	0.769	0.930	**0.996**
	Jazz	0.821	0.989	**0.999**
	Pops	0.301	0.383	**0.973**

Table 2. K values obtained from Speaker clustering experiment. Average duration of each utterance is about 10 seconds.

		GMM-HAC	IV-KMEANS	IV-SC
Clean		0.900	**1.000**	**1.000**
Stationary noise	Crowd	0.672	0.809	**0.981**
	Party	0.727	0.752	**0.964**
Non-Stationary noise	Street	0.225	0.331	**0.876**
	Station	0.398	0.470	**0.820**
BGM	Classic	0.355	0.604	**0.964**
	Jazz	0.467	0.789	**0.983**
	Pops	0.193	0.263	**0.665**

6. CONCLUSION

i-vector-based spectral clustering was applied to speaker clustering in various types of noise and yielded significant gains from conventional agglomerative and i-vector-based k-means clustering.

This work assumes the correct number of clusters is known and does not focus on estimating it. Several attempts have been made to discover the number of clusters equal to the optimal number of eigenvectors on the basis of the largest gradient of eigenvalues for clean speech data [8, 6]. However, the present work showed that to achieve high accuracy of speaker clustering in the noise conditions, more eigenvectors than the number of actual clusters should be used. This implies that eigenvalue-based method is no more applicable to estimate the number of clusters for noisy speech data and the alternative approach is required. An attempt will be made as a future work to estimate the optimal number of clusters using more sophisticated manner, e.g., self-tuning spectral clustering [17]. In addition, PLDA scoring [3] will be utilized to construct more robust similarity matrix.

7. REFERENCES

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

[2] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Speech Audio Process.*, vol. 19, no. 4, pp. 788–798, 2011.

[3] P. Kenny, "Bayesian speaker verification with heavy- tailed priors," in *Odyssey: The Speaker and Language Recognition Workshop*, June 2010.

[4] L. Ferrer, M. McLaren, N. Scheffer, Y. Lei, M. Graciarena, and V. Mitra, "A noise-robust system for nist 2012 speaker recognition evaluation," in *Interspeech*, Aug. 2013, pp. 1981–1985.

[5] S. Shum, N. Deha, E. Chuangsuwanich, D. Reynolds, and J. Glass, "Exploiting intra-conversation variability for speaker diarization," in *interspeech*, Aug. 2011, pp. 945–948.

[6] S. Shum, N. Dehak, E. Chuangsuwanich, D. Reynolds, and J. Glass, "On the use of spectral and iterative methods for speaker diarization," in *Interspeech*, Sept. 2012.

[7] H. Tang T. Huang H. Ning, M. Liu, "A spectral clustering approach to speaker diarization," in *ICSLP*, May 2006.

[8] K. Iso, "Speaker clustering using vector quantization and spectral clustering," in *ICASSP*, Mar. 2010, pp. 4986–4989.

[9] A. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *NIPS*, Dec. 2001.

[10] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," in *Digital Signal Processing*, Jan. 2000, vol. 10, pp. 19–41.

[11] T. Kawahara, H. Nanjo, and S. Furui, "Automatic transcription of spontaneous lecture speech," in *Proc. IEEE workshop on Automatic Speech Recognition and Understanding*, 2001.

[12] S. Itahashi, "A noise database and japanese common speech data corpus," *Journal of the Acoustical Society of Japan*, vol. 47, no. 12, pp. 951–953, 1991, in Japanese.

[13] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Popular, classical, and jazz music database," in *3rd International Conference on Music Information Retrieval (ISMIR)*, Oct. 2002, pp. 287–288.

[14] A. Solomonoff, "Clustering speakers by their voices," in *ICASSP*, May 1998, pp. 757–760.

[15] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," *Acoust Soc Jpn E*, vol. 20, no. 3, pp. 199–206, 1999.

[16] "ASJ continuous speech corpus for research (ASJ-JIPDEC)," *National Institute of Information*, `http://research.nii.ac.jp/src/en/ASJ-JIPDEC.html`.

[17] A. Y. Ng, M. I. Jordan, and Y. Weiss, "Self-tuning spectral clustering," in *NIPS*, Dec. 2004, pp. 1601–1608.