Speaker Clustering in Multi-party Conversation

Masafumi Nishida, Yuki Ishikawa, Seiichi Yamamoto

Graduate School of Engineering, Doshisha University, Kyoto 610-0321, Japan

{mnishida,seyamamo}@mail.doshisha.ac.jp, dtl0720@mail4.doshisha.ac.jp

Abstract

Speech feature variations are mainly attributed to variations in phonetic and speaker information included in speech data. If these two types of information are separated from each other, more robust speaker clustering can be achieved. We propose a speaker clustering method using principal component analysis transformation by separating speaker information from phonetic information, under the assumption that a space with large within-speaker variance is a "phonetic subspace" and a space with small within-speaker variance is a "speaker subspace". We carried out comparative experiments of the proposed method with conventional methods based on Bayesian information criterion and Gaussian mixture model in an observation space. The experimented results showed that the proposed method can achieve higher clustering accuracy than conventional methods.

1 Introduction

In automatic interaction management, it is important to improve interactions by making interaction smooth and natural, and be able to elicit and to provide communicative signals that allow the user to take the turn. Recently there has been growing interest in the automatic analysis of conversational data so as to further our understanding of human-human communication and multimodal signaling of social interactions. Due to advance technology, it is possible to study communicative behavior and social signaling patterns using automatic analysis techniques. Besides speech and speaker recognition, also motion capture and gesture recognition technology can be used, while the development in eye-tracker technology allows us to study gaze behaviour in an objective manner.

Chen et al. (2009) investigated combining verbal with nonverbal cues (i.e., hand gesture and eye gaze) to detect floor control shifts in multi-party meetings. Jokinen et al. (2010) showed that eye-gaze is an important cue in deciding turn-taking: the use of eye-gaze information improves classification accuracy of turn-taking significantly, compared with the use of only speech features or dialogue acts. Battersby (2011) studied interactions with a motion tracker device, and points out that the speaker's gesturing behavior differs from that of the addressees, and that head and hand movements are also different between primary and secondary addressees.

In this paper, we focus on speaker clustering based on speaker recognition technique in multi-party conversations. Speaker clustering is a technique for clustering utterances from the same speaker, and is useful for retrieving the utterances of a specific speaker and for improving automatic speech recognition performance based on speaker adaptation of the acoustic model. Speaker clustering has been studied mainly for broadcast news audio, multi-party conversations, and telephone conversations (Tranter and Reynolds, 2006) (Reynolds and Torres-Carrasquillo, 2005).

In previous studies, Chen et al. (1998) presented a maximum likelihood approach for acoustic change detection; the detection of a turn is based on the Bayesian information criterion (BIC), a model selection criterion well-known in statistics. Furthermore, Cheng et al. (2010) proposed three divide-and-conquer approaches for BIC-based speaker segmentation. The three approaches are used to detect speaker changes by recursively partitioning a large analysis window into two sub-windows and recursively verifying the merging of two adjacent audio segments using ΔBIC, a widely adopted distance measure of two audio segments. Iso (2010) proposed a method for representing a speech segment with a vector of Vector quantization (VQ) code frequencies by using a cosine between two vectors as their similarity measure. The clustering is done using a spectral clustering algorithm with cluster number estimation based on an eigen structure of the similarity matrix. Nishida et al. (2005) proposed a flexible framework in which an optimal speaker model (GMM or VQ) is automatically selected based on the BIC and according to the amount of training data available. Reynolds et al. (1998) presented the cross likelihood ratio (CLR), and Le et al. (2007) presented the normalized cross likelihood ratio (NCLR) and the advantages of using it in a speaker diarization system.

For speaker identification and verification, Nishida et al. (2001) proposed a method based on a statistical speaker model (GMM) in the "speaker subspace" which is created using all speech data projected to the speaker subspace where the phonetic information is suppressed. The speech data include two types of information, phonetic and speaker. Phonetic information is attributed to the phonetic features in speech data, and speaker information is attributed to the speaker features in speech data. In particular, phonetic information varies depending on the speech data. Therefore, if these two types of information are separated from each other, robust speaker recognition can be achieved.

Conventional speaker-clustering methods do not distinguish between phonetic and speaker information. We propose a speaker clustering method based on a statistical speaker model (GMM) in the “speaker subspace”, which is created using all speech data projected to the speaker subspace where the phonetic information is already suppressed. In speaker clustering, we believe that our method is effective in separating speaker from phonetic information because the variance in duration of each segment enlarges variation of phonetic information in the segment more in comparison with speaker identification and verification. We carried out speaker clustering experiments with three methods. The first method was a hierarchical agglomerative clustering method based on the BIC in an observation space. The second method was a hierarchical clustering method based on CLR using GMM in an observation space. The third method is the proposed method based on GMM in the speaker subspace obtained from an observation space. Our proposed method clusters using the CLR.

The remainder of this paper is organized as follows: Section 2 explains speaker clustering based on GMM in speaker subspace, Section 3 describes our speaker clustering experiments and section 4 concludes the paper.

2 Speaker Clustering based on GMM in Speaker Subspace

2.1 Separation of phonetic and speaker subspaces

We describe a separation method of phonetic and speaker information. The speech feature variation is mainly caused by the variation in the phonetic information included in speech data. This insight enables the separation of the phonetic and speaker information based on this variance. Principal component analysis (PCA) is conducted to locate each speaker’s speech data of phonetic information in a subspace constructed using the principal component axes (lower order axes), and speaker information in a complementary subspace constructed using the higher order axes. We call the subspace with the large variation constructed using the lower axes “phonetic subspace”, and the subspace with the small variation constructed using the higher axes “speaker subspace”.

A sequence of speech data $\left\{x_t^{(s)}\right\}\left(t = 1,2,\dots,N^{(s)}\right)$ of a segment s is observed in an n-dimensional observation space. Its mean vector $\bar{x}^{(s)}$ and covariance matrix $R^{(s)}$ are then computed from the training data as follows:

$$\bar{x}^{(s)} = \frac{1}{N}\sum_{t=1}^{N} x_t^{(s)} \tag{1}$$

$$R^{(s)} = \frac{1}{N}\sum_{t=1}^{N}(x_t^{(s)} - \bar{x}^{(s)})(x_t^{(s)} - \bar{x}^{(s)})^T \tag{2}$$

The covariance matrix R can be composed of eigenvectors and a matrix of eigenvalues as follows:

$$R^{(s)} = \Phi^{(s)}\Lambda^{(s)}\Phi^{(s)T}, \tag{3}$$

where $\Lambda^{(s)}$ is a diagonal matrix whose diagonal components are eigenvalues $\lambda_i^{(s)}$ $(i = 1,\dots,k,\dots n)$ of $R^{(s)}$, and $\Phi^{(s)}$ is a matrix whose columns are eigenvectors $\varphi_i^{(s)}$ $(i = 1,\dots,k,\dots n)$ of $R^{(s)}$.

The eigenvalues $\lambda_i^{(s)}$, which are obtained by eigenvalue decomposition, represent a variance in the eigenvectors $\varphi_i^{(s)}$, which are orthonormal bases. In this study, a space constructed by eigenvectors corresponding to the largest eigenvalues up to k numbers is the phonetic subspace, which represents the phonetic information. A space constructed by $(n-k)$ eigenvectors corresponding to the remaining small $(n-k)$ eigenvalues is the speaker subspace, which is complementary to the phonetic subspace. The speaker subspace represents the speaker information. Consequently, the input speech can be separated into phonetic and speaker information by projecting both type of information to the speaker and phonetic subspaces, respectively.

2.2 Speaker clustering based on projection to speaker subspace

Clustering ideally produces one cluster for each speaker in a conversation and assigns all segments from each speaker to a single cluster. Gaussian mixture models are trained using the speech data projected to the speaker subspace for each segment.

The Mel-frequency cepstral coefficient (MFCC) is commonly used in speaker recognition and is obtained from the log filter-bank amplitudes using a discrete cosine transform (DCT). However DCT is not designed to transform a space by taking into account data distribution as well as correlation of feature parameters. In this study, we used PCA instead of DCT to diagonalize a data covariance matrix and decorrelate the feature parameters of the log filter-bank amplitudes. This PCA, which we used instead of DCT for signal processing, can also construct respective speaker subspace.

A sequence of speech data $\left\{x_t^{(s)}\right\}$ of a segment s observed in an n-dimensional observation space is projected to the speaker space by using Eq. (4) and the speaker model (GMM) is trained in the speaker subspace by using the projected speech data.

$$\hat{x}_t^{(s)} = P^{(s)T}(x_t^{(s)} - \bar{x}^{(s)}) \tag{4}$$

The orthogonal matrix $P^{(s)}$ has columns that are higher order eigenvectors $\varphi_i^{(s)}(i = k, \dots, n)$, which were obtained with PCA for the segment. Figure 1 shows an example of the projection to the speaker subspace. The speaker subspaces of segments *A* and *B*, shown with rectangles, are respectively denoted by P_A and P_B. The regions enclosed by ellipses indicate the speech data. The speaker subspace is a space constructed by axes whose variance is small. Therefore, after projecting the speech data of segments *A* and *B* to each speaker subspace, a within-speaker variance becomes smaller than that in an observation space, leaving a fixed between-speaker variance.

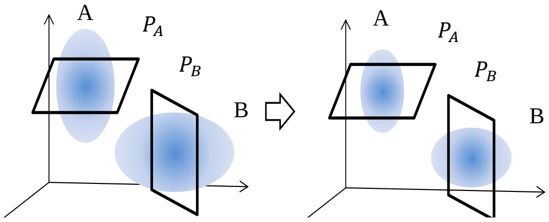

Figure 1: Projection to speaker space

Figure 2 shows a conceptual diagram of the projected phonetic subspace. The orthonormal basis vector $\varphi_1^{(s)}$ configures the phonetic subspace, and the orthonormal basis vectors $\varphi_2^{(s)}$ and $\varphi_3^{(s)}$ configure the speaker subspace. The input feature vector x_t can be divided into phonetic vector $x^{(s)}_{phoneme}$ and speaker vector $x^{(s)}_{\text{speaker}}$ by using Eqs. (5) and (6), respectively. $x^{(s)}_{phoneme}$ shows the phonetic vector projected to the phonetic subspace, and $x^{(s)}_{\text{speaker}}$ shows the speaker vector projected to the speaker subspace.

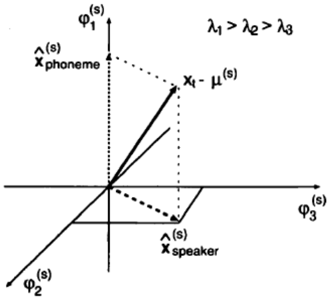
$\varphi_1^{(s)}$
$\lambda_1 > \lambda_2 > \lambda_3$
$\hat{x}^{(s)}_{\text{phoneme}}$
$x_t - \mu^{(s)}$
$\varphi_3^{(s)}$
$\hat{x}^{(s)}_{\text{speaker}}$
$\varphi_2^{(s)}$

Figure 2: Phonetic vector and speaker vector

$$x^{(s)}_{phoneme} = \sum_{i=1}^{k} (x_t - \bar{\mathrm{x}}^{(s)}, \quad \varphi_i^{(s)})\varphi_i^{(s)} \tag{5}$$

$$x^{(s)}_{\text{speaker}} = \sum_{i=k+1}^{n} (x_t - \bar{\mathrm{x}}^{(s)}, \quad \varphi_i^{(s)})\varphi_i^{(s)} \tag{6}$$

A common approach used in speaker clustering is hierarchical agglomerative clustering with a CLR consisting of the following steps:

1. Form one cluster from each segment.

2. Construct a speaker subspace in the segment by performing PCA.

3. Project speech data in the segment to the speaker subspace by using Eq. (4).

4. Construct a statistical speaker model (GMM) in the respective speaker subspaces.

5. Compute the CLR as pair-wise distances between each cluster (Reynolds et al. ,1998). The CLR $d_{ij}$for clusters i and j is given by Eq. (7).

$$d_{ij} = \log\frac{P(X_i|\lambda_i)}{P(X_i|\lambda_j)} + \log\frac{P(X_j|\lambda_j)}{P(X_j|\lambda_i)} \tag{7}$$

$$\log P(X_i|\lambda_j) = \frac{1}{n_i}\sum_{k=1}^{n_i} \log P(x_{ik}|\lambda_j)$$

where X_i is a segment of cluster i, x_{ik} is its kth frame feature of the segment, n_i is the number of frames of a segments, λ_i is the parameters of GMM for cluster i, and $\log P(X_i|\lambda_j)$ is the average log likelihood of the segment of cluster i given by model λ_i.

6. Merge the closest pairs of clusters, if the minimum distance between the clusters is smaller than the threshold θ.

7. Update distances of remaining clusters to form a new cluster by using the unweighted pair-group method using arithmetic averages (UPGMA) (Sneath and Sokal, 1973) by Eq. (8).

$$d(r,s) = \frac{1}{n_r n_s}\sum_{i=1}^{n_r}\sum_{j=1}^{n_s} dist(x_{ri}, x_{sj}), \tag{8}$$

   where r and s are the cluster number, n_r and n_s indicate the number of segments in each cluster, and $dist(x_{ri}, x_{sj})$ is obtained by Eq. (7).

8. Iterate steps 5-7. The clustering process finishes if all distances between clusters are not smaller than the threshold θ.

3 Experiments

3.1 Experimental Setup

We used corpus of spontaneous Japanese (CSJ) as evaluation data. The CSJ consists of 3302 talks (662 hours, 1417 speakers) collected from academic conference presentations and extemporaneous speeches (Maekawa, 2003). The talks are segmented into utterances at every pause of longer than 300 milliseconds. We chose utterances of multiple speakers randomly from the CSJ to make the test sets as close to actual multi-party conversations as possible. We used five test sets (1-5), each of which consisted of five speakers. The duration of an utterance ranged from 30 to 70 seconds. In addition, we also used another five test sets (6-10), each of which consisted of 10 speakers. The duration of an utterance ranged from 20 to 50 seconds. The duration of one speaker's total speech was about 100 seconds. There are not overlapping utterances in the test tests. Table 1 lists the detail of each test set.

The speech data was sampled at 16 kHz, analyzed with an analysis window size of 25 ms with 10-ms overlap, and parameterized into 24 cepstral coefficients obtained using a 24-channel Mel-frequency spaced filter-bank.

Table 1: Details of each test set

Test set No.	Number of speakers	Number of segments	Total segments time (min)
1	5	55	44.5
2	5	57	45.1
3	5	59	44.4
4	5	58	44.8
5	5	55	45.0
6	10	177	95.0
7	10	181	93.7
8	10	183	93.5
9	10	174	81.4
10	10	171	91.5

We carried out speaker clustering experiments with three methods: The first method was a hierarchical agglomerative clustering method based on BIC in an observation space with 24 dimensional MFCC parameters. The second method was a hierarchical clustering method based on the CLR using GMM in an observation space with 24 dimensional MFCC parameters. The third method was the proposed method based on GMM in the speaker subspace obtained from an observation space with 24 channel log filter-bank amplitudes. Our method clustered using CLR.

The clustering results were aligned with the ground truth speaker labels to measure their accuracy based on the diarization error rate (DER) (Iso, 2010):

$$\mathrm{DER} = \frac{T_{miss} + T_{wrong}}{T_{ref}}, \tag{9}$$

where T_{miss} is the total length of segments not aligned with the speaker labels, T_{wrong} is the total length of segments aligned with the wrong speaker labels, and T_{ref} is the total length of all segments in a test set. We also calculated the purity metric (Iso, 2010):

$$\mathrm{Purity} = \frac{T_{pure}}{T_{ref}}, \tag{10}$$

where T_{pure} is the total length of the speaker label, which is the longest utterances for each cluster.

3.2 Experimental results

Table 2 lists the clustering results for test sets 1-5, and Table 3 lists the clustering results for test sets 6-10. The parameter α for the BIC is the turning parameter, MN indicates the number of mixtures of the GMM, and SD for the proposed method indicates the dimensions of the speaker subspace. To investigate the phoneme -dependency of each eigenvector axis, we compared 20 combinations of dimensions with 1-20th, 1- 21st, 1-22nd, 1-23rd, 1-24th, 2-20th, 2-21st, …, and 4-24th eigenvectors.

Table 2: Clustering results for the test sets 1-5

	DER(%)	Purity(%)	Parameter
BIC	8.8	90.5	$\alpha = 1.5$
GMM	10.1	89.4	MN = 2

Proposed method	6.8	92.2	MN = 4
			SD = 2–21

Table 3: Clustering results for the test sets 6-10

	DER(%)	Purity(%)	Parameter
BIC	10.8	87.9	$\alpha = 1.2$
GMM	12.8	86.4	MN = 4

Proposed method	7.1	92.2	MN = 4
			SD = 2–21

Tables 2 and 3 show that the proposed method obtained a higher clustering accuracy than that obtained with the conventional methods based on the BIC and GMM, for both groups of test sets. Test sets 5-10 contained five speakers and test sets 6-10 contained 10 speakers. Therefore, the proposed method can obtain high clustering accuracy with a variation in the number of speakers.

Figures 3 and 4 show the relation between clustering accuracy and the number of mixtures for the conventional GMM and the proposed method for test sets 1-5 (Fig. 3) and 6-10 (Fig.4). The optimal number of mixtures of the GMM varies because GMM of two mixtures is best for test sets 1-5 and GMM of four mixtures is best for test sets 6-10. However, the optimal number of mixtures of the proposed method does not depend on the number of speakers.

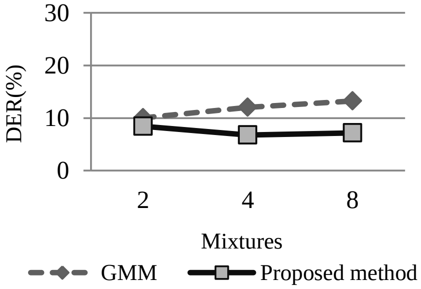

Figure 3: DER in each mixture for test sets 1-5

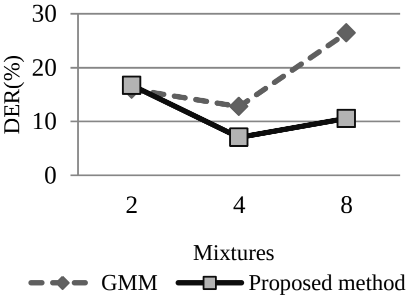

Figure 4: DER in each mixture for test sets 6-10

A preliminary experiment, showed that the first axis of PCA should not be used for configuring the low-dimensional axes of the speaker subspace in the proposed method. Therefore, Fig. 5 shows the DER when the higher-dimensional axes of the speaker subspace are reduced. The number of mixtures is four for all cases.

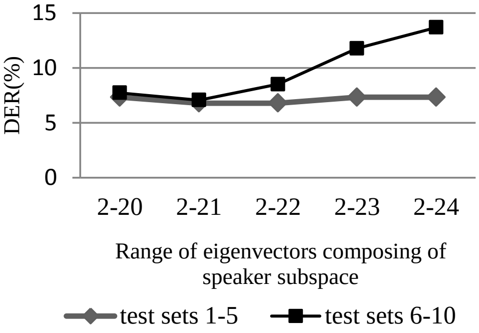

Figure 5: DER in various ranges of eigenvectors composing the speaker subspace

As clearly shown in the Fig. 5, the best DER was obtained when SD was 2-21 for both test sets. However, in each test set, the best DER varied by the dimensions of the speaker subspace because the variation in utterance lengths was large. Therefore, we will study how to select the optimal dimensions of the speaker subspace by considering the variability of phoneme in speech data.

The average number of clusters with the BIC was 5.0, the GMM was 5.8, and the proposed method was 5.6 for test sets 1-5. The standard deviation was 0.71 for the BIC, 1.30 for the GMM, and 0.89 for the proposed method. For test sets 6-10, the average number of clusters was 11.6, 13.8, and 14.0, for the BIC, GMM and proposed method, respectively. The standard deviation by the BIC was 0.55, the GMM was 2.56 and the proposed method was 1.22. The proposed method used a threshold for the CLR to stop the clustering process. For future work, we will use BIC as a stopping criterion of clustering for the proposed method to improve the estimation accuracy of the number of speakers.

4 Conclusions

We proposed a speaker-clustering method using a GMM trained in speaker subspace using speech data projected to the speaker subspace. The proposed method used PCA transform to construct the speaker subspace.

From the results of the speaker clustering experiments, the DER with the BIC was 8.8% for test sets 1-5 and 10.8% for test sets 6-10, that with the CLR using a GMM was 10.1% for test sets 1-5 and 12.8% for test sets 6-10, and that with the proposed method was 6.8% for test sets 1-5 and 7.1% for test sets 6-10. Therefore, the proposed method obtained a higher speaker clustering accuracy than that with the conventional methods. The experiments also demonstrated that separating the phonetic and speaker subspaces using PCA was effective.

For future work, we will evaluate the proposed method on the National Institute of Standards and Technology (NIST) databases to demonstrate its generality. It is also necessary to study how to select the optimal number of dimensions of the speaker subspace. Moreover, we will study on speaker clustering for test data included overlapping utterances.

References

Lei Chen and Mary P. Harper. 2009. *Multimodal Floor Control Shift Detection*, Proc. ICMI-MLMI.

Kristiina Jokinen, Kazuaki Harada, Masafumi Nishida, and Seiichi Yamamoto. 2010. *Turn-alignment Using Eye-gaze and Speech in Conversational Interaction*, Proc. Interspeech, pp.2018-2021.

Stuart Battersby. 2011. *Moving Together: the Organization of Non-verbal Cues During Multiparty Conversation*, PhD Thesis.

Sue E. Tranter and Douglas A. Reynolds. 2006. *An Overview of Automatic Speaker Diarization Systems*, IEEE Transactions on Audio, Speech, and Language Processing, Vol.14, No.5, pp.1557-1565.

Douglas A. Reynolds and Pedro A. Torres-Carrasquillo. 2005. *Approaches and Applications of Audio Diarization*, Proc. ICASSP, Vol.5. pp.953-956.

Scott Chen and Ponani Gopalakrishnan. 1998. *Speaker Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion*, Proc. DARPA Broadcast News Transcription and Understanding Workshop, pp.127-132.

Shih-Sian Cheng, Hsin-Min Wang, Hsin-Chia Fu. 2010. *BIC-based Speaker Segmentation Using Divide-and-conquer Strategies with Application to Speaker Diarization*, IEEE Transactions, Vol.18, pp.141-157.

Kenichi Iso. 2010. *Speaker Clustering Using Vector Quantization and Spectral Clustering*, Proc. ICASSP, pp. 4986 – 4989.

Masafumi Nishida and Tatsuya Kawahara. 2005. *Speaker Model Selection Based on the Bayesian Information Criterion Applied to Unsupervised Speaker Indexing*, IEEE Transactions on Speech and Audio Processing, Vol.13, No.4, pp. 583-592.

Douglas A. Reynolds, Elliot Singer, Beth A. Carlson, Gerald C. O'Leary, Jack J. McLaughlin, and Marc A. Zissman. 1998. *Blind Clustering of Speech Utterances based on Speaker and Language Characteristics*, Proc. ICSLP, pp.3193-3196.

Viet-Bac Le, Odile Mella, and Dominique Fohr. 2007. *Speaker Diarization using Normalized Cross Likelihood Ratio*, Proc.Interspeech, pp.1869-1872.

Masafumi Nishida and Yasuo Ariki. 2001. *Speaker Recognition by Separating Phonetic Space and Speaker Space*, Proc. EUROSPEECH, Vol. 2, pp. 1381-1384.

Peter Sneath and Robert R. Sokal. 1973. *Numerical Taxonomy*, W. H. Freeman and Company.

Kikuo Maekawa. 2003. *Corpus of Spontaneous Japanese: Its Design and Evaluation*, Proc. ISCA & IEEE Workshop on SSPR, pp.7-12.