SPEAKER RECOGNITION BY SEPARATING PHONETIC SPACE AND SPEAKER SPACE

M.Nishida and Y.Ariki

Department of Electronics and Informatics
Ryukoku University, Seta, Otsu-shi, Shiga, 520-2194, JAPAN
nishida@arikilab.elec.ryukoku.ac.jp

Abstract

In speaker recognition, it is a problem that speech feature varies depending on sentences and time difference. This variation is mainly attributed to the variation of phonetic information and speaker information included in speech data. If these two kinds of information are separated each other, robust speaker recognition will be realized. In this study, we propose a speaker identification and speaker verification method by separating the phonetic information and speaker information by a subspace method, under the assumption that a space with large within-speaker variance is a "phonetic space" and a space with small within-speaker variance is a "speaker space". We carried out comparative experiments of the proposed method with a conventional method based on GMM in an observation space as well as in a space transformed by LDA. As a result, we could construct a robust speaker model with a few model parameters using a few training data by the proposed method.

1. Introduction

In speaker recognition, it is a problem that speech feature varies depending on sentences and time difference. This variation is mainly attributed to the variation of phonetic information and speaker information included in speech data. If these two kinds of information are separated each other, robust speaker recognition will be realized by using only the speaker information. However, it is difficult to separate the phonetic information and speaker information included in speech data at present.

In speaker recognition, GMM(Gaussian Mixture Model) has been conventionally used and is statistically constructed using features of speech data[1]-[3]. The conventional GMM is based on a statistical method but not based on the separation of the phonetic information and speaker information included in speech data.

We have already proposed speaker recognition based on a subspace method in order to extract the speaker information included in speech data [4]-[6]. In this proposal, we noticed that speech feature variation is mainly caused by the variation of the phonetic information, not the speaker information included in speech data. This insight leads us to the separation of the phonetic information and speaker information based on the variance. Namely by performing PCA(Principal Component Analysis) to each speaker's speech data, phonetic information locates in a subspace constructed by the principal component axes(lower order axes), and speaker information locates in a complementary subspace constructed by the higher order axes. According to this insight, we call the subspace with the large variation constructed by the lower axes "phonetic space" and the subspace with the small variation constructed by the higher axes "speaker space".

In this study, we propose a speaker identification and speaker verification method based on a statistical speaker model(GMM) in the "speaker space" using the speech data projected to the speaker space where the phonetic information is already suppressed.

In this paper, we carried out comparative experiments to show an effectiveness of our proposed method with the conventional two methods: a method based on GMM in an observation space and a method based on GMM in a space transformed by LDA(Linear Discriminant Analysis)[7].

2. Conventional Space Transformation by LDA

2.1. LDA for Multiple Classes

We discuss the LDA defined by a linear transformation of an n-dimensional feature space for multiple classes($c>2$). A sequence of training data $\{x_t^{(i)}\}$ $(t=1,2,\cdots,N^{(i)})$ of a speaker i is observed in an n-dimensional observation space. A mean vector of this training data and a covariance matrix are denoted by $\mu^{(i)}$ and $\Sigma^{(i)}$. A priori probability of speaker i and a mean vector of total speakers are denoted by $P(\omega^{(i)})$ and μ respectively. A within-class covariance matrix $\Sigma_W = \Sigma_{i=1}^{c} P(\omega^{(i)})/N^{(i)} \Sigma_{t=1}^{N^{(i)}} (x_t^{(i)} - \mu^{(i)})(x_t^{(i)} - \mu^{(i)})^T$ and a between-class covariance matrix $\Sigma_B = \Sigma_{i=1}^{c} P(\omega^{(i)})(\mu^{(i)} - \mu)(\mu^{(i)} - \mu)^T$ are defined as follows: where $N^{(i)}$ denotes the number of training data of speaker i and N denotes the number of training data of total speakers. Here, a priori probability $P(\omega^{(i)})$ of each speaker can be represented by $P(\omega^{(i)}) = N^{(i)}/N$.

The within-class covariance matrix $\tilde{\Sigma}_W = A^T \Sigma_W A$ and the between-class covariance matrix $\tilde{\Sigma}_B = A^T \Sigma_B A$ after linear transformation are represented using a transformation matrix A. A Fisher's criterion $J_\Sigma(A) = |A^T \Sigma_B A| / |A^T \Sigma_W A|$ which can represent a class separation is defined as a ratio of a between-class variance to a within-class variance. To obtain the optimal transformation matrix A, the above described Fisher's criterion $J_\Sigma(A)$ is maximized. The solution is obtained by maximizing $|A^T \Sigma_B A|$ under a condition of $A^T \Sigma_W A = I$. This can be solved as an eigenvalue problem $\Sigma_B A = \Sigma_W A \Lambda$. Λ denotes a diagonal matrix whose diagonal components

are eigenvalues $\lambda_i (i = 1, \cdots, k, \cdots, n)$. If Σ_W is a non-singular matrix, the transformation matrix A is obtained by an eigenvalue decomposition of ${\Sigma_W}^{-1}\Sigma_B$. Therefore, eigenvectors for the large eigenvalues up to k numbers construct a k-dimensional space.

2.2. Speaker Recognition in Space Transformed by LDA

A sequence of training data $\{x_t^{(s)}\}$ of a speaker s is observed in an n-dimensional observation space. LDA is applied to all the speech data of all the speakers and eigenvectors $a_i (i = 1, \cdots, k, \cdots, c-1)$ are obtained by the eigenvalue decomposition. The transformation matrix A is a matrix whose columns are the eigenvectors and a subspace constructed by $\{a_i\}$ is a space with the largest separation for all the speakers. Then the training data is projected to the subspace by Eq.(1).

$$\hat{x}_t^{(s)} = A^T x_t^{(s)} \tag{1}$$

After projecting a training data of each speaker to the subspace, the speaker model of GMM is trained in the subspace.
Fig.1 shows an example of the space transformation by LDA. As shown in Fig.1, the separation ratio of all the

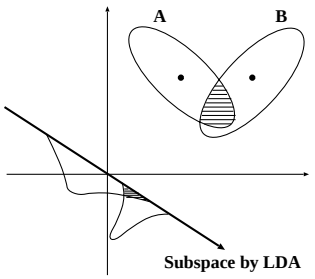

Figure 1: Space transformation by LDA

speakers becomes largest after the projection to the subspace obtained by LDA.
In speaker identification, a sequence of input feature vector $\{x_t\}$ is projected to the subspace by Eq.(1) and a log likelihood $\log P(\hat{x}_t|\lambda^{(c)})$ of each customer c is computed by GMM. An identification result is obtained as customer s with maximum log likelihood shown in Eq.(2).

$$s = \arg\max_c \log P(\hat{x}_t|\lambda^{(c)}) \tag{2}$$

In speaker verification, a sequence of input feature vector $\{x_t\}$ is projected to the subspace by Eq.(1) and a log likelihood $\log P(\hat{x}_t|\lambda^{(s)})$ of the claimed speaker s is computed by GMM. Then the log likelihood is normalized based on likelihood ratio shown in Eq.(3).

$$\log \bar{P}(\hat{x}_t|\lambda^{(s)}) = \log P(\hat{x}_t|\lambda^{(s)}) - \max_{i \neq s} \log P(\hat{x}_t|\lambda^{(i)}) \tag{3}$$

In this normalization method, the log likelihood $\log P(\hat{x}_t|\lambda^{(s)})$ of the claimed speaker is subtracted by a maximum likelihood of the other speaker except the claimed speaker. If the normalized log likelihood is larger than a threshold, the speaker is accepted as the true speaker.

3. Projection to Speaker Space

3.1. Subspace Separation

Speech data includes both of phonetic information and speaker information. If they are separated each other, robust speaker recognition will be realized by using only the speaker information. Here we propose a separation method of these information.
The speech feature variation is mainly caused by the variation of the phonetic information included in speech data. This insight leads us to the separation of the phonetic information and speaker information based on the variance. Namely by performing PCA to each speaker's speech data, phonetic information locates in a subspace constructed by the principal component axes(lower order axes), and speaker information locates in a complementary subspace constructed by the higher order axes. According to this insight, we call the subspace with the large variation constructed by the lower axes "phonetic space", and the subspace with the small variation constructed by the higher axes "speaker space".
In this study, we propose a speaker identification and speaker verification method based on a statistical speaker model(GMM) in the proposed "speaker space" using the speech data projected to the speaker space where the phonetic information is already suppressed.
A sequence of training data $\{x_t^{(s)}\}$ $(t = 1, 2, \cdots, N^{(s)})$ of a speaker s is observed in an n-dimensional observation space. Then a mean vector $\mu^{(s)}$ and a covariance matrix $R^{(s)}$ are computed from the training data as follows:

$$\mu^{(s)} = \frac{1}{N^{(s)}} \sum_{t=1}^{N^{(s)}} x_t^{(s)} \tag{4}$$

$$R^{(s)} = \frac{1}{N^{(s)}} \sum_{t=1}^{N^{(s)}} (x_t^{(s)} - \mu^{(s)})(x_t^{(s)} - \mu^{(s)})^T \tag{5}$$

By eigenvalue decomposition, the covariance matrix $R^{(s)}$ is decomposed as:

$$R^{(s)} = \Phi^{(s)} \Sigma^{(s)} {\Phi^{(s)}}^T \tag{6}$$

Here, $\Sigma^{(s)}$ is a diagonal matrix whose diagonal components are eigenvalues $\lambda_i^{(s)} (i = 1, \cdots, k, \cdots, n)$ of $R^{(s)}$. $\Phi^{(s)}$ is a matrix whose columns are eigenvectors $\varphi_i^{(s)} (i = 1, \cdots, k, \cdots, n)$ of $R^{(s)}$.
The eigenvalues $\lambda_i^{(s)}$ which are obtained by the eigenvalue decomposition represent a variance on the eigenvectors $\varphi_i^{(s)}$ which are orthonormal bases. In this study, a space constructed by the lower order eigenvectors corresponding to the large eigenvalues up to k numbers is called "phonetic space". A space constructed by the higher order eigenvectors corresponding to the remaining small $n-k$ eigenvalues is a complementary subspace of the phonetic space and is called "speaker space". Therefore, the phonetic space constructed by axes with large variance is a space where the phonetic information is powerfully represented. On the other hand, the speaker space, which is complementary to the phonetic space, is a space where the speaker information is well represented by suppressing the phonetic information. Consequently, the input speech can be separated into phonetic information and

speaker information by projecting to the speaker space and to the phonetic space respectively.
In this study, a speaker identification and speaker verification are carried out in the following way:

1. Construct a speaker space for each speaker by performing PCA for his/her speech data and selecting the higher order axes.
2. Project a training data to his/her speaker space.
3. Construct a statistical speaker model GMM in the respective speaker space.

We call the proposed method "SSGMM(Speaker Space GMM)".

3.2. Speaker Recognition by Projection to Speaker Space

MFCC(Mel-Frequency Cepstral Coefficient) is commonly used in speaker recognition. MFCC is obtained from the log filter-bank amplitudes using DCT(Discrete Cosine Transform). However DCT is not designed to transform a space by considering a data distribution as well as correlation of feature parameters. In this study, we employ PCA instead of DCT to diagonalize a data covariance matrix and decorrelate the feature parameters of the log filter-bank amplitudes. This PCA used instead of DCT for signal processing can also construct respective speaker space. Therefore, this PCA plays two roles.
A sequence of training data $\{x_t^{(s)}\}$ of a speaker s is observed in an n-dimensional observation space. A subspace constructed by the higher order eigenvectors ${\varphi_i}^{(s)}(i = k, \cdots, n)$ which were obtained by PCA for the training data is a speaker space. The orthogonal matrix $P^{(s)}$ is a matrix whose columns are the eigenvectors ${\varphi_i}^{(s)}$. Training data is projected to the speaker space by Eq.(7) and the speaker model GMM is trained in the speaker space by using the projected training data.

$$\hat{x}_t^{(s)} = {P^{(s)}}^T (x_t^{(s)} - \mu^{(s)}) \tag{7}$$

Fig.2 shows an example of the projection to the speaker space. In Fig.2, P_A and P_B, shown by rectangles, denote

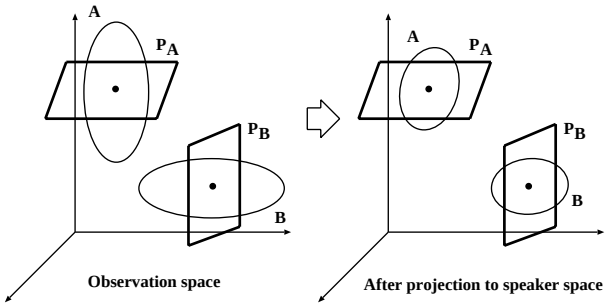

Figure 2: Projection to speaker space

a speaker space of speaker A and speaker B respectively. The regions enclosed by ellipses indicate the speech data. As shown in Fig.2, the speaker space is a space constructed by axes whose variance is small. Therefore, after projecting the training data of each speaker A and B to each speaker space, a within-speaker variance becomes small compared with that in an observation space, leaving a between-speaker variance fixed.

In speaker identification, a sequence of input feature vector $\{x_t\}$ is projected to the speaker space of each customer c by Eq.(7) and a log likelihood $\log P(\hat{x}_t|\lambda^{(c)})$ is computed by GMM of each customer c. An identification result is obtained as customer s with maximum log likelihood computed by Eq.(2).
In speaker verification, a sequence of input feature vector $\{x_t\}$ is projected to the speaker space of a claimed speaker s by Eq.(7) and a log likelihood $\log P(\hat{x}_t|\lambda^{(s)})$ is computed by GMM of the claimed speaker s. The log likelihood is normalized based on likelihood ratio shown in Eq.(3). If the normalized log likelihood is larger than a threshold, the speaker is accepted as the true speaker.

4. Speaker Recognition Experiments

4.1. Experimental Condition

In the experiment, sentences were uttered by 30 speakers(20 males and 10 females) at two time sessions over ten months. The duration of the speech data is 4 sec at average. The speech data was sampled at 16kHz, the analysis window size was 20ms with 5ms overlap and was parameterized into 21 cepstral coefficients obtained by 24-channel mel-frequency spaced filter-bank analysis. In this study, we carried out speaker recognition experiments by three methods: The first method is a conventional method based on GMM in an observation space with 21 dimensional MFCC parameters. The second method is a conventional method based on GMM in a subspace obtained by LDA from the observation space with 21 dimensional MFCC parameters. The third method is the proposed method(SSGMM) based on GMM in the speaker space obtained from an observation space with 24 channel log filter-bank amplitudes.
In speaker identification experiment, 30 speakers are customers. Five sentences uttered at first time session were used for training. Therefore the duration of the training data is 20sec at average. Each 15 sentences uttered at second time session were used for evaluation. In speaker verification experiment, the 30 speakers are divided into two: 15 speakers are true speakers (customers) and the other 15 speakers are impostors for each customer. Five sentences uttered at first time session were used for training. The duration of the training data is 20sec at average. Each 15 sentences uttered at second time session were used for evaluation.

4.2. Experimental Results

We carried out comparative experiments among the proposed method(SSGMM), a conventional method based on GMM in an observation space and in a subspace obtained by LDA to show an effectiveness of our proposed method.
Experimental results of speaker identification are shown in Table 1. The results were evaluated by IER(Identification Error Rate). Experimental results of speaker verification are shown in Table 2. The results were evaluated by EER(Equal Error Rate).
In Table 1, "GMM" denotes the IER by a conventional method based on normal GMM in an observation space(21 dimensional MFCC parameters). "LDA" denotes the IER by a conventional method based on GMM in a subspace obtained by LDA(21 dimensional space) from the observation space(21 dimensional MFCC parameters).

"SS" denotes the IER by the proposed method(SSGMM) based on GMM in the speaker space(4-24 dimensional space) obtained from an observation space(24 channel log filter-bank amplitudes).

Table 1: Speaker identification result(%)

mixture number	GMM	LDA (21dim)	SS (4-24dim)
2	18.44	16.67	6.44
4	12.00	11.33	5.11
8	9.56	8.22	6.44
64	8.89	6.00	7.78

As a result from Table 1, the IER was 5.11% by SSGMM(4 mixtures). The IER was reduced by 57% by SSGMM(4 mixtures) compared with the conventional method based on GMM(4 mixtures). The IER was reduced by 55% by SSGMM(4 mixtures) compared with the conventional method based on GMM(4 mixtures) in a subspace obtained by LDA. Furthermore, the IER was reduced by 43% by SSGMM(4 mixtures) compared with the conventional method based on GMM(64 mixtures). The IER was reduced by 15% by SSGMM(4 mixtures) compared with the conventional method based on GMM(64 mixtures) in a subspace obtained by LDA. Therefore, we can reduce the mixture components by 1/16 by the proposed method, still reducing the IER by more than 15%.

In Table 2, "GMM" denotes the EER by a conventional method based on normal GMM in an observation space(21 dimensional MFCC parameters). "LDA" denotes the EER by a conventional method based on GMM in a subspace obtained by LDA(21 dimensional space) from the observation space(21 dimensional MFCC parameters). "SS" denotes the EER by the proposed method(SSGMM) based on GMM in the speaker space(4-24 dimensional space) obtained from an observation space(24 channel log filter-bank amplitudes).

Table 2: Speaker verification result(%)

mixture number	GMM	LDA (14dim)	SS (4-24dim)
2	3.96	4.52	2.44
4	2.98	3.91	1.79
8	2.28	3.07	1.88
64	2.00	2.68	1.70

As a result from Table 2, the EER was 1.79% by SSGMM(4 mixtures). The EER was reduced by 40% by SSGMM(4 mixtures) compared with the conventional method based on GMM(4 mixtures). The EER was not reduced by GMM in a subspace obtained by LDA compared with the speaker identification. Furthermore, the EER was reduced by 11% by SSGMM(4 mixtures) compared with the conventional method based on GMM(64 mixtures). Therefore, we can reduce the mixture components by 1/16 by the proposed method.

As a result from Table 1 and Table 2, high identification and verification performance were obtained by the proposed method(SSGMM), so that the proposed speaker space was proven to be a space where the speaker information is well presented by suppressing the phonetic information, under the assumption that a space with large within-speaker variance is the phonetic space and a space with small within-speaker variance is the speaker space. Further, it was shown that a robust speaker model can be constructed by a few training data with 20sec at average and a few model parameters(4 mixture GMM). As a result, the training, identification and verification can be performed in a real time.

5. Conclusion

In this study, we proposed the speaker identification and speaker verification method using the speaker model trained by GMM in each speaker space, after projecting the training data to the speaker space and suppressing the phonetic information, under the assumption that a space with large within-speaker variance is the phonetic space and a space with small within-speaker variance is the speaker space.

As a result of the speaker identification and speaker verification experiments, the IER was reduced by 43% by SSGMM(4 mixtures) compared with the conventional method based on GMM(64 mixtures). The EER was reduced by 11% by SSGMM(4 mixtures) compared with the conventional method based on GMM(64 mixtures). Therefore, a robust speaker model can be constructed by a few training data and a few model parameters, and real time training, identification and verification can be performed by the proposed method.

6. Acknowledgement

We wish to express our deep appreciation for Dr.Matsui belonged to ATR and NTT Cyber Space Laboratory offered a speaker recognition database which we used in this study.

7. References

[1] D.A.Reynolds and R.C.Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models," IEEE Trans.SAP, vol.3, no.1, pp.72-83, 1995.

[2] S.V.Vuuren, "Comparison of Text-independent Speaker Recognition Methods on Telephone Speech with Acoustic Mismatch," Proc.ICSLP, vol.3, pp.1788-1791, 1996.

[3] D.Tran and M.Wagner, "A Proposed Likelihood Transformation for Speaker Verification," Proc.ICASSP, vol.2, pp.1069-1072, 2000.

[4] Y.Ariki and K.Doi, "Speaker Recognition based on Subspace Method," Proc.ICSLP, vol.4, pp.1859-1862, 1994.

[5] Y.Ariki, S.Tagashira and M.Nishijima, "Speaker Recognition and Speaker Normalization by Projection to Speaker Subspace," Proc.ICASSP, vol.1, pp.319-322, 1996.

[6] M.Nishida and Y.Ariki, "Speaker Verification by Integrating Dynamic and Static Features using Subspace Method", Proc.ICSLP, vol.3, pp.1013-1016, 2000.

[7] E.Batlle, C.Nadeu and Jose.A.R.Fonollosa, "Feature Decorrelation Methods in Speech Recognition, A Comparative Study," Proc.ICSLP, vol.3, pp.951-954, 1998.