

SPEAKER RECOGNITION BASED ON SUBSPACE METHODS

Y. Aiki and K. Doi

Ryukoku University, Seta Otsu-shi 520-21, Japan

ABSTRACT

This paper proposes a method to separate speaker information from phonetic information included in speech data. A speaker recognition method using the separated speaker information is also proposed and shown to be equivalent with a method based on speaker subspace. The validity of this proposed method was verified by carrying out simple speaker recognition experiments.

1 INTRODUCTION

Speech recognition sometimes requires who is speaking, because intention of an utterance becomes different depending on speakers. This requirement leads to an idea that speaker recognition should be integrated into speech recognition. In other words, they are ought to be carried out simultaneously.

From this view point, this paper investigates speaker information and phonetic information included in speech signal and tries to separate them each other. As a result, using the separately extracted speaker information, we carry out speaker recognition.

A speech production model, we assume, is that speech data presented in an observation space as a sequence of short time spectra is produced by projecting speech data presented in an individual speaker space to the observation space. We also assume that the speech data presented in an individual speaker space are neutral and include only phonetic information. The individual speaker space is defined by orthonormal bases computed from the speech data. This concept coincides with a subspace method proposed by Oya.[1]

According to the subspace method, we prepare a projection matrix to each speaker by computing the orthonormal bases of the speech data. Using the projection matrices, we recognize who is speaking by finding a speaker subspace to which the projection length of the input speech data is maximized.

2 SEPARATION MODEL

2.1 Singular value decomposition

Let us consider a situation where we are observing speech data X_A of speaker A and speech data X_B of speaker B in an observation space shown in Fig.1. The

speech data are a sequence of spectral feature vectors x_{At} and x_{Bt} obtained at time t by short time spectral analysis. We denote the speech data X_A as a matrix whose row is a spectral feature vector x_{At}^T , ($1 \leq t \leq M$). The column of the matrix corresponds to frequency i , ($1 \leq i \leq N$).

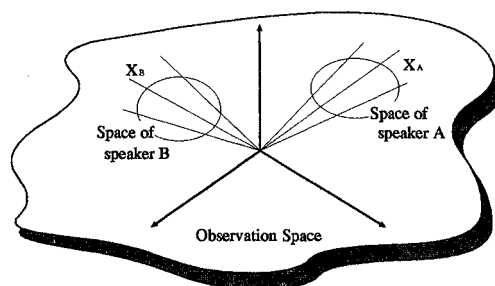


Figure 1: Observation space and speaker space

By singular value decomposition, the speech data matrix X_A is decomposed as

$$X_A = U_A \Sigma_A V_A^T \quad (1)$$

Here U_A and V_A are the matrices whose columns are eigenvectors of $X_A X_A^T$ and $X_A^T X_A$ respectively, and Σ_A is the singular value matrix of X_A .

If r numbers of the larger singular values are selected from the matrix Σ_A , the matrix U_A becomes $M \times r$ dimension and the row still corresponds to time. The matrix V_A^T becomes $r \times N$ dimension and the column corresponds to frequency. A new interpretation of equation (1) is that the speech data matrix $U_A \Sigma_A$ is produced by projecting the speech data matrix X_A in the observation space to the individual speaker space through the orthogonal transformation matrix V_A . The reverse projection is also interpreted as that the individual speaker space is projected to the observation space through the orthogonal transformation matrix V_A^T .

Since the speech data matrix $U_A \Sigma_A$ is represented in own speaker space, we can say that speaker information is less included in $U_A \Sigma_A$ than the speech data matrix X_A presented in the observation space. This interpretation means that $U_A \Sigma_A$ has mainly phonetic information and V_A has mainly speaker information.

2.2 Eigenvalue decomposition

Here we investigate more the speaker information matrix V_A . V_A is the matrix whose columns are eigenvectors of a correlation matrix $X_A^T X_A$. It is expressed by eigenvalue decomposition as follows:

$$X_A^T X_A = V_A \Sigma V_A^T \quad (2)$$

The eigenvectors of the correlation matrix $X_A^T X_A$ are shown to be orthonormal bases, of the speech data X_A , computed based on a criterion that the total distance between an observed speech vector x_{At} and the orthonormal bases is minimized (See APPENDIX).

2.3 Subspace method

Since the speaker information matrix V_A is composed of orthonormal bases $\{v_{A1}, \dots, v_{Ar}\}$ of the speech data X_A , the matrix V_A constructs the speaker subspace. Here we consider a distance from an arbitrary speech feature vector x in the observation space to the speaker subspace V_A . It is presented as follows using a projection matrix P_A from the observation space to the speaker subspace.

$$Dist(V_A, x) = \| P_A x - x \| \quad (3)$$

where the projection matrix P_A is defined as:

$$P_A = \sum_{k=1}^r v_{Ak} v_{Ak}^T \quad (4)$$

because the projected vector $P_A x$ is presented as $\sum_{k=1}^r v_{Ak} (v_{Ak}^T x)$.

Equation (4) is written as follows:

$$P_A = V_A V_A^T \quad (5)$$

Equation(5) means that the projection matrix from the observation space to the speaker subspace is obtained using the orthonormal bases of the speech data X_A . Once the projection matrix P_A is obtained, the distance from the speech feature vector x in the observation space to the speaker subspace (the projected vector $P_A x$) can be computed. The speaker is identified as one with the subspace nearest to the input speech feature vector x .

In practice, the length of the projected vector $\| P_A x \|^2$ is computed, instead of the distance, and the speaker is identified as one with the maximum length of the projected vector to his subspace. It can be expressed as follows:

$$x^T P_A x > x^T P_B x \quad \text{then the speaker is A} \quad (6)$$

The decision boundary of the subspace method is defined as non-linear.

$$\begin{aligned} \text{Boundary}_{AB} &= \{x | x^T P_A x = x^T P_B x\} \\ &= \{x | x^T (P_A - P_B) x = 0\} \end{aligned} \quad (7)$$

3 SPEAKER RECOGNITION

Fig.2 shows a system of speaker training and speaker recognition based on the subspace method.

3.1 Speaker training

The training step of individual speaker information proceeds as follows:

- (1) Speech signals sampled at 12KHz are converted into frequency domain by 256-point FFT analysis for each speaker.
- (2) The frequency domain below 6,000Hz is converted into log scale and divided into 16 bands.
- (3) A sequence of 16 order spectral feature vectors x_{At} is obtained by averaging and log-scaling the energy at each band.
- (4) A correlation matrix $X_A^T X_A$ is computed.
- (5) Orthonormal bases v_{Ak} of the speech data X_A is obtained by eigenvalue decomposition of the correlation matrix $X_A^T X_A$. The matrix V_A is composed of the bases v_{Ak} .
- (6) The projection matrix P_A is computed by equation (5). Since V_A and P_A are the matrices with 16x16, speaker information of each speaker is presented by 256 parameters.

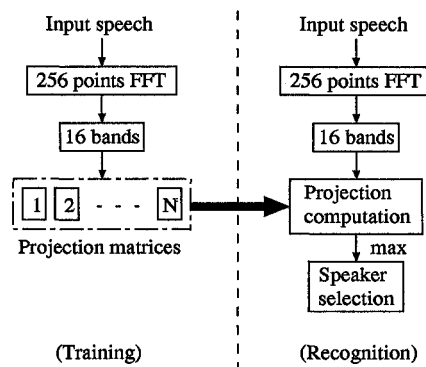


Figure 2: Speaker training and recognition

3.2 Speaker recognition

Speaker recognition is performed by applying the projection matrices of each speaker to input speech. The step proceeds as follows.

- (1) A sequence of 16 order spectral feature vectors x_t , ($1 \leq t \leq K$) is obtained from speech signal in the same way as the training step.

- (2) Projected length of the speech vector x_i to the subspace of speaker A is computed using the projection matrix P_A by equation (6). The averaged length y^A of the input speech projected to the subspace of speaker A is computed as follows:

$$y^A = \frac{1}{K} \sum_{i=1}^K x_i^T P_A x_i \quad (8)$$

- (3) Speaker is identified by selecting the maximum averaged length of the input speech projected to the subspace.

4 RECOGNITION EXPERIMENT

4.1 Database and experimental condition

Speech data are sentences spoken by 15 speakers (10 males and 5 females) at three time sessions during 10 months.[2] These sessions are denoted as 90-8, 91-3 and 91-6 in this paper.

The 15 speech sentences, including each 5 utterances with normal speed, fast speed and slow speed respectively were selected from each speaker for training. The averaged duration of the sentences is about 4 seconds. They were 256-point FFT analyzed and a time sequence of 16 order spectral feature vectors was produced. Table 1 shows the condition of speech analysis.

Table 1: Condition of speech analysis

Sampling frequency	12KHz
High-pass filter	$1 - 0.97z^{-1}$
Feature analysis	FFT spectrum(256)
Feature parameter	16 band energy
Frame length	20ms
Frame shift	5ms
Window type	Hamming window

4.2 Experiment 1

To investigate the dimension of the speaker subspace, we computed a cumulated contribution rate CCR which is defined by the following expression:

$$CCR(s) = \frac{\sum_{i=1}^s \alpha_i}{\sum_{i=1}^{16} \alpha_i} \quad (9)$$

Here, α_i indicates the i th eigenvalue obtained by eigenvalue decomposition of the correlation matrix $X_A^T X_A$. Table 2 shows the cumulated contribution rate as a function of a number of dimensions. From the table, it can be seen that the subspace dimension is enough up to 5.

Table 2: Cumulated contribution rate(%)

1dim	2dim	3dim	4dim	5dim	6dim
97.6	99.2	99.8	99.9	100.0	100.0

4.3 Experiment 2

The projection matrices P_k ($1 \leq k \leq 15$) were produced using the speech data at one time session 90-8 for each speaker, and registered as speaker templates.

Speaker recognition was carried out using 15 different sentences (text independent) spoken at normal, fast and slow speed for each 5 sentences. The time session used for the recognition was 90-8, 91-3 and 91-6. The speaker was identified by selecting the longest projection vector on the 15 speaker subspaces.

The result was shown in Fig.3. In the figure, the horizontal axes is the subspace dimension and the results at the three time sessions were shown by different lines. The highest recognition rate was 100 % at the same time session as training, and the lowest was 91 % at the different time session. From the figure, it can be seen that the subspace dimension is enough at 5 which corresponds well with the result by [Experiment 1].

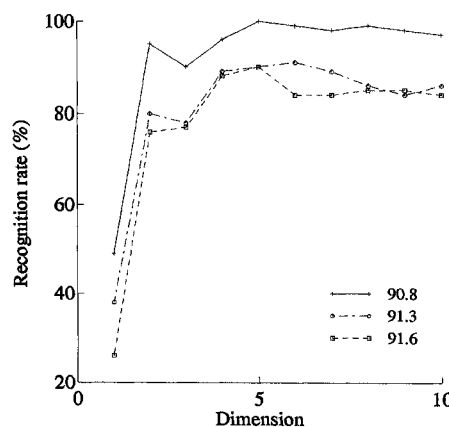


Figure 3: Result of speaker recognition (Trained at 90-8)

4.4 Experiment 3

To solve a problem that the recognition system is suffered from difference of the time session, speech data over two time sessions were used for training. Namely, the projection matrices P_k ($1 \leq k \leq 15$) were produced using the speech data over two time sessions for each speaker.

Speaker recognition was carried out using 15 different sentences spoken at normal, fast and slow speed for 5 sentences respectively. The time session for the recognition was remaining one among 90-8, 91-3 and 91-6.

The result was shown in Fig.4. In the figure, the horizontal axes is the subspace dimension and the results at the three time sessions were shown by different lines. The line part1 shows the result recognized at time session 91-6, trained over 90-8 and 91-3. The line part2 and part3 are the result at time session 91-3 and 90-8 respectively, trained over remaining two sessions. The highest recognition rate was 94.0 %. From the figure, it can be seen that training over two sessions improves the recognition a little.

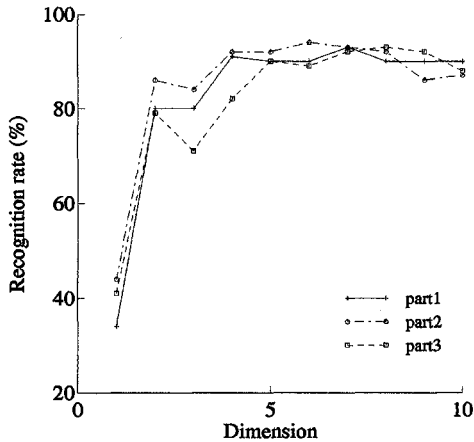


Figure 4: Result of speaker recognition (Trained over two time sessions)

5 CONCLUSION

A model to separate speaker information from phonetic information was described on the basis of singular value decomposition. The speaker recognition using the separated speaker information was shown to be equivalent with the speaker subspace method. The experimental results show the effectiveness of the proposed method. Further work will be required to solve the problem that speaker recognition is degraded when speaking time is different.

APPENDIX

We prove that the eigenvectors of $X_A^T X_A$ are the orthonormal bases, of the speech data X_A , computed based on a criterion that the total distance between the observed speech vector x_{At} and the orthonormal bases is minimized.

(Proof)

Let l denote one of the orthonormal bases of the speech data $X_A = \{x_{A1}, \dots, x_{An}\}^T$. We compute l by minimizing the distance between the speech vector x_{At} and the base. Here we assume $\|l\|^2 = l^T l = 1$. The projected length z_t of the speech vector x_{At} to the orthonormal basis l is expressed as:

$$z_t = l^T x_{At} \quad (10)$$

The mean square $H(l)$ of the distance between the speech vector x_{At} and the projected vector is computed by subtracting square of the projected length $(l^T x_{At})^2 = l^T x_{At} x_{At}^T l$ from square of the speech vector length $\|x_{At}\|^2 = x_{At}^T x_{At}$ as follows:

$$\begin{aligned} H(l) &= E\{x_{At}^T x_{At} - l^T x_{At} x_{At}^T l\} \\ &= E\{x_{At}^T x_{At}\} - l^T E\{x_{At} x_{At}^T\} l \\ &= L - l^T S l \end{aligned} \quad (11)$$

Here, L is the mean square of the speech vector length, and S is the correlation matrix $E\{x_{At} x_{At}^T\} = X_A^T X_A$ of the speech data X_A .

Lagrange's method is used to minimize equation(11). The augmented function is:

$$F(l) = L - l^T S l + \lambda(l^T l - 1) \quad (12)$$

The following equation is obtained by differentiating partially equation (12) and equalizing it to zero.

$$\frac{\partial F(l)}{\partial l} = -2S l + 2\lambda l = 0 \quad (13)$$

Then

$$S l = \lambda l \quad (14)$$

It can be seen that the basis l is an eigenvector of the correlation matrix $S = X_A^T X_A$. In addition, the following equation is obtained.

$$H(l) = L - l^T S l = L - l^T \lambda l = L - \lambda \quad (15)$$

This means that the mean square $H(l)$ of the distance between the speech vector and the projected vector is computed by subtracting the square of the projected length λ from the square of the speech vector length L . Since minimizing the $H(l)$ equals to maximizing the square of the projected length λ , the first basis l_1 is obtained as an eigenvector for the maximum eigenvalue.

After computing the first basis l_1 , the second and further bases l_2, \dots, l_n are computed as the eigenvectors whose eigenvalues are the second large and so on.

By this step, a set of orthonormal bases $V_A = (l_1, \dots, l_n)$ is obtained. Equation (14) is rewritten as the following equation.

$$S(l_1, \dots, l_n) = (l_1, \dots, l_n) \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix} \quad (16)$$

Then finally the following equation is obtained.

$$S V_A = V_A \Sigma \quad \text{therefore} \quad X_A^T X_A = V_A \Sigma V_A^T \quad (17)$$

This indicates that the orthonormal bases V_A of the speech data X_A is obtained by eigenvalue decomposition of the correlation matrix $X_A^T X_A$.

ACKNOWLEDGEMENT

We wish to express our deep appreciation to Dr., Prof. S.Furui and Miss T.Matsui at NTT human interface laboratories for allowing us to use their speaker recognition database.

REFERENCES

- [1] E.Oja, "Subspace Methods of Pattern Recognition", Research Studies Press, England, 1983.
- [2] T.Matsui and S.Furui, "Concatenated Phoneme Models for Text-variable Speaker recognition", ICASSP93, pp.11-391-11-394, 1993.