EFFECTIVENESS OF SPEAKER NORMALIZED HMM BY PROJECTION TO SPEAKER SUBSPACE

Yasuo Ariki

Department of Electronics and Informatics
Ryukoku University, Seta Otsu-shi 520-21, Japan
ariki@rins.ryukoku.ac.jp http://arikilab.elec.ryukoku.ac.jp/

ABSTRACT

Conventional speaker-independent HMMs ignore the speaker differences and collect speech data in an observation space. This causes a problem that probability distribution of the HMMs becomes flat, and then causes recognition errors. To solve this problem, we construct the speaker subspace for an individual speaker and project his speech data to his own subspace. By this method we can extract speaker-independent phonetic information included in the speech data. Speaker-independent HMMs can be constructed using this phonetic information. In this paper, we describe the result of phoneme recognition experiments using the speaker-independent HMMs constructed by the speech data projected to the speaker subspaces.

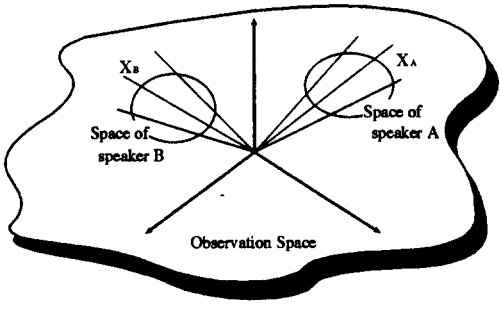

Figure 1: Observation space and speaker subspace

1. INTRODUCTION

Speaker-independent HMMs are usually constructed using various kinds of speech spoken by many speakers. This causes a problem that the probability distribution of the HMMs becomes flat and then causes recognition errors.

This flatness is explained in Fig.1. An individual speaker has his own subspace in which his phoneme characteristics are well represented. The subspaces of different speakers locate in different position in an observation space. However, conventional speaker-independent HMMs ignore the speaker subspaces and collect speech data in the observation space.

To solve this problem, the individual speaker subspace should be constructed using his own speech data and consequently speaker normalized phoneme data be produced by projecting the speech data to his own subspace. Speaker-independent HMMs can be trained by collecting the speaker normalized phoneme data. In this paper, we propose a method to construct the speaker normalized HMMs and show the effectiveness experimentally in phoneme recognition by comparing with the conventional speaker-independent HMMs.

2. SPEAKER SUBSPACE

As shown in Fig.1, we observe speech data X_A of speaker A and speech data X_B of speaker B in an observation space. 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, $(1 \leq i \leq N)$.

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

$$X_A = U_A \Sigma_A V_A^T \tag{1}$$

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

The eigenvectors of the correlation matrix $X_A^T X_A$ are orthonormal bases of the speech data X_A, computed based on a criterion that the total distance is minimized between the observed speech vectors x_{At} $(1 \leq t \leq M)$ and the orthonormal bases[1]. Then V_A is considered as the orthonormal bases of the speaker subspace. If r numbers of the larger singular values are selected from the matrix Σ_A, the number of dimensions of the matrix U_A becomes M x r and the row still corresponds to time. The number of dimensions of the matrix V_A^T becomes r x N and the matrix V_A^T is considered as the speaker subspace. This method of constructing the speaker subspace V_A is called the CLAFIC method[2].

Since the speech data matrix $U_A \Sigma_A$ is produced by projecting the speech data to the speaker subspace V_A, and represented in his own speaker subspace, 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 indicates that $U_A\Sigma_A$ is the speaker normalized data and has mainly phonetic information. Based on this idea, speaker normalization methods are described in more details.

3. SPEAKER NORMALIZATION

3.1. Canonical Correlation Analysis

A well known method of speaker normalization and adaptation is canonical correlation analysis[3]. The step of the canonical correlation analysis is summarized as follows (see APPENDIX (A));

STEP(1) Feature vectors in spoken sentences are matched by DP between speaker A and B, and the matched speech data X_A and X_B are obtained.

STEP(2) X_A and X_B are decomposed as $X_A = QR$ and $X_B = PS$ respectively by QR-decomposition.

STEP(3) $\Omega = Q^T P$ is computed and eigenvectors v'_{Ai} with the large eigenvalues are obtained by eigenvalue decomposition of the $\Omega\Omega^T$. In the same way, eigenvectors v'_{Bi} are obtained by eigenvalue decomposition of the $\Omega^T\Omega$. The axis $v_{Ai} = R^{-1}v'_{Ai}$ of speaker A and $v_{Bi} = S^{-1}v'_{Bi}$ of speaker B are computed.

3.2. CLAFIC Canonical Correlation Analysis

The canonical correlation analysis has a problem that the subspace produced by the canonical correlation analysis does not present the speech data in a compact and powerful way. It also causes the problem that the HMMs must be re-trained when a pair of speakers are changed, because the subspaces of a pair of speakers are simultaneously produced by the canonical correlation analysis.

To solve these two problems, we have already proposed CLAFIC canonical correlation analysis in which the subspace of speaker A is produced by singular value decomposition shown by Eq.(1) at first, and then the subspace of speaker B is produced as to maximize the correlation of the subspace axes between speaker A and B [4]. The step of the CLAFIC canonical correlation analysis is summarized as follows (see APPENDIX (B));

STEP(1) Feature vectors in spoken sentences are matched by DP between speaker A and B and the matched speech data X_A and X_B are obtained.

STEP(2) Orthonormal bases V_A of the speaker A is computed using the speech data X_A by singular value decomposition.

STEP(3) The axis v_B of speaker B is computed as follows in the way of maximizing the correlation between the axes v_A and v_B using speech data X_B.

$$v_B = \frac{\sqrt{C}\Sigma_{22}^{-1}\Sigma_{21}v_A}{\sqrt{v_A^T\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}v_A}} \quad (2)$$

where C is the variance on the axis v_A.

In this paper, we show experimentally the effectiveness of this CLAFIC canonical correlation analysis for multiple speaker normalization and also for constructing speaker normalized HMMs, compared with the conventional speaker-independent HMMs.

4. SPEAKER NORMALIZATION RESULT

4.1. Database and Experimental Condition

We carried out phoneme recognition experiments for multiple speakers using CLAFIC canonical correlation analysis. The number of phonemes is 46 kinds. The speech data used is ATR phoneme-balanced sentence set which includes 7 speakers and 150 spoken sentences for each speaker. The experimental condition is shown in Table1.

Table 1: Experimental condition

A	Sampling frequency	12kHz
A	High-pass filter	$1 - 0.97z^{-1}$
A	Feature parameter	LPC cepstrum(16th)
A	Frame length	20ms
A	Frame shift	5ms
A	Window type	Hamming window
HMM	Number of states	5 states 3 loops
HMM	Covariance matrix	Diagonal
HMM	Type	Mixture densities HMM
HMM	Number of Mixture	4

4.2. Effectiveness for Multiple Speakers

We selected the speaker MTK (male) as a base speaker. The subspace of the base speaker MTK was designed by the CLAFIC method using his 150 spoken sentences. His phoneme HMMs were constructed using speaker normalized speech data obtained by projecting 150 spoken sentences to his subspace. For remaining six speakers (3 males and 3 females), speaker subspaces were constructed by CLAFIC canonical correlation analysis using even numbered 75 sentences among 150. Odd numbered 75 sentences among 150 were projected to the speaker subspace and resulted in speaker normalized speech data. This speaker normalized data was recognized by the phoneme HMMs constructed by the speaker normalized data from MTK. The evaluation was based on the phoneme HMM evaluation algorithm without phoneme hand-labels[5].

The recognition result is shown in Table2. In the table, DEP indicates the speaker dependent phoneme recognition result and INDEP indicates the speaker independent-phoneme recognition result where 500 spoken sentences from 50 speakers were used for HMM construction. The averaged recognition rates of DEP and INDEP were 75.9% and 54.3% respectively. FIX indicates the phoneme recognition result without speaker normalization. Namely, speech data in the observation space was recognized by the HMMs of the speaker MTK, constructed in the observation space without speaker normalization. The averaged recognition rate of this method was 44.9%. CCCA indicates the result by the CLAFIC Canonical Correlation Analysis. The averaged recognition rate was 62.0%.

4.3. Speaker-Independent HMMs

In this experiment, 50 speaker subspaces were constructed using 50 spoken sentences respectively (selected from ASJ

database) by the CLAFIC Canonical Correlation Analysis. Then 500 spoken sentences (10 from each speaker) projected to the speaker subspaces were used for constructing speaker-independent HMMs. The recognition result is shown in Table2. In the table, CCCA-INDEP indicates the phoneme recognition result by speaker-independent HMMs constructed using the speaker normalized data. The averaged recognition rate of this method was 63.7%.

From the table, it can be said that phoneme recognition accuracy of the CLAFIC Canonical Correlation Analysis is 17% higher than that of FIX and 8% higher than that of the speaker-independent method (INDEP). The speaker-independent HMMs after speaker normalization (CCCA-INDEP) shows the accuracy by 1.7% higher than CCCA. This means that speaker normalization is well performed by CCCA.

Table 2: Phoneme recognition result for multiple speakers using speaker normalized data(%)

		FIX	CCCA	CCCA-INDEP	DEP	INDEP
MSH	(m)	59.5	61.5	63.6	74.9	55.0
MHO	(m)	43.2	56.8	59.2	70.1	47.2
MMY	(m)	52.2	61.7	59.8	77.0	48.3
FYM	(f)	39.0	62.5	65.7	77.5	58.5
FYM	(f)	38.0	67.0	70.9	80.4	62.7
FKS	(f)	37.6	62.2	63.0	75.6	54.3
Average		44.9	62.0	63.7	75.9	54.3

5. CONCLUSION

The effectiveness of speaker normalization by the CLAFIC canonical correlation analysis was shown through multiple speaker experiments. The phoneme recognition result showed that the accuracy of the CLAFIC canonical correlation analysis was 8% higher than that of the conventional speaker-independent HMMs which were trained by the speech data in the observation space. Further work is planned for investigating the speaker normalization ability by discriminant analysis instead of the CLAFIC method.

6. REFERENCES

[1] Y.Ariki and K.Doi, "Speaker Recognition based on Subspace Method," ICSLP'94, pp.1859-1862, 1994.

[2] E.Oja, "Subspace Methods of Pattern Recognition," Research Studies Press, England, 1983.

[3] K.Choukri G.Chollet Y.Grenier, "Spectral transformations through Canonical Correlation Analysis for speaker adaptation in ASR," ICASSP86, pp.2659-2662, 1986.

[4] Y.Ariki, S.Tagashira and M.Nishijima, "Speaker Recognition and Speaker Normalization by Projection to Speaker Subspace," ICASSP96, pp.319-322, 1996.

[5] Y.Minami, T.Matsuoka, K.Shikano, "Phoneme HMM Evaluation Algorithm without Phoneme Labeling," ICSLP'92, pp.1535-1538, 1992.

7. APPENDIX

(A) Canonical Correlation Analysis

Canonical correlation analysis finds two axes v_A and v_B whose correlation is maximized, after projecting the speech data X_A of speaker A and the speech data X_B of speaker B to the axes v_A and v_B respectively. This analysis is shown to be exactly same as finding two axes v_A and v_B on a criterion that the distance between the speech data of speaker A and the corresponding speech data of speaker B is minimized, after projecting them to the axes v_A and v_B respectively.

(Proof)
Let $X_A v_A$ denote a one-dimensional vector whose elements are speech data X_A of speaker A projected to the axis v_A. In the same way, let $X_B v_B$ denote a one-dimensional vector whose elements are speech data X_B of speaker B projected to the axis v_B. The total square distance d^2 between these two projected speech data $X_A v_A$ and $X_B v_B$ is computed as follows;

$$\begin{aligned} d^2 &= (X_A v_A - X_B v_B)^T (X_A v_A - X_B v_B) \\ &= v_A^T X_A^T X_A v_A + v_B^T X_B^T X_B v_B \\ &\quad - v_B^T X_B^T X_A v_A - v_A^T X_A^T X_B v_B \\ &= v_A^T \Sigma_{11} v_A + v_B^T \Sigma_{22} v_B - v_B^T \Sigma_{21} v_A - v_A^T \Sigma_{12} v_B \end{aligned} \quad (3)$$

Here following constraints are employed to find the solution.

$$v_A^T \Sigma_{11} v_A = 1 \quad (4)$$

$$v_B^T \Sigma_{22} v_B = 1 \quad (5)$$

Then d^2 is simplified as follows;

$$\begin{aligned} d^2 &= 1 + 1 - v_B^T \Sigma_{21} v_A - v_A^T \Sigma_{12} v_B \\ &= 2(1 - v_A^T \Sigma_{12} v_B) \end{aligned} \quad (6)$$

It can be seen that the analysis to minimize the total square distance d^2 between the two speech data projected to the axes v_A and v_B is same as the analysis to maximize the cross-correlation $v_A^T \Sigma_{12} v_B$ between the two axes v_A and v_B. Further it can be said that the two subspaces obtained by the canonical correlation analysis for two speakers are different in the observation space but relatively equivalent because the axis correlation is maximized. It should be noticed that the axes of the subpace obtained by the canonical correlation analysis are not orthonormal.

Now, a method to find the axes v_A and v_B is shown. The cross-correlation $v_A^T \Sigma_{12} v_B$ shown in Eq.(6) is maximized under the two constraints shown in Eq.(4) and Eq.(5). Using Lagrange's method of indeterminate multiplier, the augmented objective function is defined as follows;

$$\varphi(v_A, v_B) = v_A^T \Sigma_{12} v_B - \frac{\mu_1}{2} v_A^T \Sigma_{11} v_A - \frac{\mu_2}{2} v_B^T \Sigma_{22} v_B \quad (7)$$

Partial differentiations in terms of v_A and v_B are;

$$\frac{\partial \varphi}{\partial v_A} = \Sigma_{12} v_B - \mu_1 \Sigma_{11} v_A = 0 \quad (8)$$

$$\frac{\partial \varphi}{\partial v_B} = (v_A^T \Sigma_{12})^T - \mu_2 \Sigma_{22} v_B = 0 \quad (9)$$

Then the following equation is obtained;

$$\Sigma_{12} v_B = \mu_1 \Sigma_{11} v_A \quad (10)$$

$$\Sigma_{21} v_A = \mu_2 \Sigma_{22} v_B \quad (11)$$

Multiplying Eq.(10) by v_A^T from the left and multiplying Eq.(11) by v_B^T from the left lead to the following equations;

$$v_A^T \Sigma_{12} v_B = \mu_1 v_A^T \Sigma_{11} v_A = \mu_1 \quad (12)$$

$$v_B^T \Sigma_{21} v_A = \mu_2 v_B^T \Sigma_{22} v_B = \mu_2 \quad (13)$$

Then

$$\mu_1 = \mu_2 = \lambda \quad (14)$$

Using the above equation, Eq.(10) and Eq.(11) are expressed as follows;

$$\Sigma_{12} v_B = \lambda \Sigma_{11} v_A \quad (15)$$

$$\Sigma_{21} v_A = \lambda \Sigma_{22} v_B \quad (16)$$

Eq.(15) and Eq.(16) are expressed as follows;

$$\begin{pmatrix} 0 & \Sigma_{12} \\ \Sigma_{21} & 0 \end{pmatrix} \begin{pmatrix} v_A \\ v_B \end{pmatrix} = \lambda \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{pmatrix} \begin{pmatrix} v_A \\ v_B \end{pmatrix} \quad (17)$$

v_A and v_B are found by solving the above equation as follows. At first, the speech data X_A and X_B are decomposed by QR-decomposition as follow;

$$X_A = QR \quad (18)$$

$$X_B = PS \quad (19)$$

here Q and P are orthogonal matrices. From Eq.(18) and Eq.(19),

$$\Sigma_{11} = X_A^T X_A = R^T Q^T Q R = R^T R \quad (20)$$

$$\Sigma_{22} = X_B^T X_B = S^T P^T P S = S^T S \quad (21)$$

From Eq.(20) and Eq.(21),

$$\begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{pmatrix} = \begin{pmatrix} R^T & 0 \\ 0 & S^T \end{pmatrix} \begin{pmatrix} R & 0 \\ 0 & S \end{pmatrix} \quad (22)$$

By substituting Eq.(22) into Eq.(17)

$$\begin{pmatrix} 0 & \Sigma_{12} \\ \Sigma_{21} & 0 \end{pmatrix} \begin{pmatrix} v_A \\ v_B \end{pmatrix} = \lambda \begin{pmatrix} R^T & 0 \\ 0 & S^T \end{pmatrix} \begin{pmatrix} R & 0 \\ 0 & S \end{pmatrix} \begin{pmatrix} v_A \\ v_B \end{pmatrix} \quad (23)$$

By simplifying the above equation,

$$\begin{pmatrix} R^T & 0 \\ 0 & S^T \end{pmatrix}^{-1} \begin{pmatrix} 0 & \Sigma_{12} \\ \Sigma_{21} & 0 \end{pmatrix} \begin{pmatrix} R^{-1} & 0 \\ 0 & S^{-1} \end{pmatrix} \times \begin{pmatrix} R & 0 \\ 0 & S \end{pmatrix} \begin{pmatrix} v_A \\ v_B \end{pmatrix} = \lambda \begin{pmatrix} R & 0 \\ 0 & S \end{pmatrix} \begin{pmatrix} v_A \\ v_B \end{pmatrix} \quad (24)$$

Here by setting $v'_A = R v_A$ and $v'_B = S v_B$,

$$\begin{pmatrix} R^T & 0 \\ 0 & S^T \end{pmatrix}^{-1} \begin{pmatrix} 0 & \Sigma_{12} \\ \Sigma_{21} & 0 \end{pmatrix} \begin{pmatrix} R^{-1} & 0 \\ 0 & S^{-1} \end{pmatrix} \times \begin{pmatrix} v'_A \\ v'_B \end{pmatrix} = \lambda \begin{pmatrix} v'_A \\ v'_B \end{pmatrix} \quad (25)$$

By simplifying Eq.(25),

$$\begin{pmatrix} 0 & R^{T^{-1}} \Sigma_{12} S^{-1} \\ S^{T^{-1}} \Sigma_{21} R^{-1} & 0 \end{pmatrix} \begin{pmatrix} v'_A \\ v'_B \end{pmatrix} = \lambda \begin{pmatrix} v'_A \\ v'_B \end{pmatrix} \quad (26)$$

Eq.(26) is further simplified by setting $\Omega = R^{T^{-1}} \Sigma_{12} S^{-1}$,

$$\begin{pmatrix} 0 & \Omega \\ \Omega^T & 0 \end{pmatrix} \begin{pmatrix} v'_A \\ v'_B \end{pmatrix} = \lambda \begin{pmatrix} v'_A \\ v'_B \end{pmatrix} \quad (27)$$

Ω is computed as follows by QR-decomposition of X_A and X_B as $X_A = QR$, $X_B = PS$,

$$\begin{aligned} \Omega &= R^{T^{-1}} \Sigma_{12} S^{-1} = R^{T^{-1}} X_A^T X_B S^{-1} \\ &= R^{T^{-1}} (R^T Q^T P S) S^{-1} = Q^T P \end{aligned} \quad (28)$$

v'_A and v'_B are computed by eigenvalue decomposition of Eq.(27). Finally v_A and v_B are computed as $v_A = R^{-1} v'_A$ and $v_B = S^{-1} v'_B$.

(B) CLAFIC Canonical Correlation Analysis

After the axis v_A is computed by a CLAFIC method which is the most typical in the subspace method[2][4], the axis v_B is computed to maximize the cross-correlation $v_A^T \Sigma_{12} v_B$ under the following constraints;

$$v_A^T \Sigma_{11} v_A = c \quad (29)$$

$$v_B^T \Sigma_{22} v_B = c \quad (30)$$

By Lagrange's method of indeterminate multiplier,

$$\varphi(v_B) = v_A^T \Sigma_{12} v_B - \frac{\mu}{2} v_B^T \Sigma_{22} v_B \quad (31)$$

By partial differentiation in terms of v_B,

$$\begin{aligned} \frac{\partial \varphi}{\partial v_B} &= (v_A^T \Sigma_{12})^T - \mu \Sigma_{22} v_B \\ &= \Sigma_{21} v_A - \mu \Sigma_{22} v_B = 0 \end{aligned} \quad (32)$$

Then

$$\mu v_B = \Sigma_{22}^{-1} \Sigma_{21} v_A \quad (33)$$

By substituting v_B into Eq.(30), following equation is obtained;

$$\begin{aligned} & \frac{1}{\mu} (\Sigma_{22}^{-1} \Sigma_{21} v_A)^T \Sigma_{22} \frac{1}{\mu} (\Sigma_{22}^{-1} \Sigma_{21} v_A) \\ = \; & \frac{1}{\mu^2} (v_A^T \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} v_A) = c \end{aligned} \quad (34)$$

Then μ is obtained as follows;

$$\mu = \frac{\sqrt{v_A^T \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} v_A}}{\sqrt{c}} \quad (35)$$

By substituting the above μ into Eq.(33), the following v_B is obtained;

$$v_B = \frac{1}{\mu} \Sigma_{22}^{-1} \Sigma_{21} v_A = \frac{\sqrt{C} \Sigma_{22}^{-1} \Sigma_{21} v_A}{\sqrt{v_A^T \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} v_A}} \quad (36)$$