# UNSUPERVISED SPEAKER NORMALIZATION USING CANONICAL CORRELATION ANALYSIS

*Yasuo Ariki and Miharu Sakuragi*

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 the output probability distribution of the HMMs becomes vague so that it deteriorates the recognition accuracy. To solve this problem, we construct the speaker subspace for an individual speaker and correlate them by o-space canonical correlation analysis between the standard speaker and input speaker. In order to remove the constraint that input speakers have to speak the same sentences as the standard speaker in the supervised normalization, we propose in this paper an unsupervised speaker normalization method which automatically segments the speech data into phoneme data by Viterbi decoding algorithm and then associates the mean feature vectors of phoneme data by o-space canonical correlation analysis. We show the phoneme recognition rate by this unsupervised method is equivalent with that of the supervised normalization method we already proposed.

## 1. INTRODUCTION

Speaker-independent HMMs are widely used in large vocabulary continuous speech recognition. The HMMs are usually constructed using various kinds of speech data 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. The speech data of the speaker $A$ locates in his own subspace where his phoneme characteristics are well represented. On the other hand, the speech data of speaker $B$ locates in his subspace different from the speaker $A$. However, conventional speaker-independent HMMs ignore the speaker subspaces and collect speech data of phonemes in an observation space.

To solve this problem, the individual speaker subspace should be constructed using his own speech data and consequently speaker normalized phoneme data should be produced by projecting the speech data to his own subspace. Speaker-independent HMMs can be trained by collecting the speaker normalized phoneme data.

From this view point of speaker normalization, we have proposed the method to correlate two subspaces of different speakers based on canonical correlation analysis [1], [2]. This method was originally proposed by K.Choukri [3] as speaker adaptation method in word recognition using DP matching. Their problem is that the subspaces of two different speakers are newly created depending on the pair of speakers. This is inconvenient in extending their method to the speaker-independent HMMs because the HMMs have to be newly created depending on the pair of speakers.

To solve this problem, we have proposed CLAFIC canonical correlation analysis [2]. It creates the subspace of the standard speaker $A$ by CLAFIC (Class featuring information compression) method at first and then creates the subspace of the input speaker $B$ by canonical correlation analysis. In this way, speech data of many speakers can be normalized to the subspace of the standard speaker $A$.

The essential problem of the canonical correlation analysis is that it requires the speech data of the same sentences between the standard speaker $A$ and the input speaker $B$ to correlate their subspaces. In this sense, it can be called a supervised speaker normalization method. If we can get rid of this requirement, the input speaker $B$ can be normalized to the standard speaker $A$ by talking any sentences or words to a recognition system. In this sense, this method can be called an unsupervised speaker normalization method.

In this paper, we propose the unsupervised speaker normalization method which allows any speakers to talk any sentences or words in normalization process. We show the experimental results of the unsupervised normalization in phoneme recognition by comparing with the supervised normalization method. As a surprising result, we found that the accuracy of our new method is almost equivalent with that of the supervised method.

## 2. SUPERVISED NORMALIZATION

### 2.1. Canonical Correlation Analysis

As shown in Fig.1, we observe speech data $X_A$ of the standard speaker A and speech data $X_B$ of the input speaker B in an observation space. The speech data is a sequence

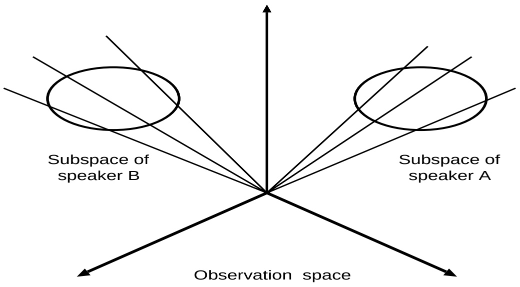

Figure 1: Observation space and speaker subspace

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 $i$, $(1 \leq i \leq N)$.

A well known method of speaker normalization and adaptation is canonical correlation analysis[2][3]. The step of the canonical correlation analysis is summarized as follows;

**STEP(1)** Feature vectors in spoken sentences are matched by dynamic programming (DP) between the standard speaker $A$ and the input speaker $B$, then 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 the standard speaker $A$ and $v_{Bi} = S^{-1}v'_{Bi}$ of the input speaker $B$ are computed. In this way, the second and the more higher order axes of two speaker's subspaces are obtained.

### 2.2. O-space Canonical Correlation Analysis

The canonical correlation analysis has a 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 this problem, we have proposed O-space canonical correlation analysis in which the subspace of the standard speaker $A$ is fixed to an observation space. Then the subspace of the input speaker $B$ is produced as to maximize the correlation of the subspace axes between speaker $A$ and $B$. The step of the O-space canonical correlation analysis is summarized as follows;

**STEP(1)** Feature vectors in spoken sentences are matched by dynamic programming (DP) between the standard speaker $A$ and the input speaker $B$, then the matched speech data $X_A$ and $X_B$ are obtained.

**STEP(2)** Orthonormal bases $V_A$ of the standard speaker $A$ is fixed to the axes of the observation space.

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

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

where $\Sigma_{12}$ and $\Sigma_{21}$ are the cross-covariance matrices between matched speech data $X_A$ and $X_B$ in an observation space. The $\Sigma_{22}$ and $C$ are the auto-covariance matrix of the matched speech data $X_B$ and the variance on the axis $v_A$ respectively.

## 3. UNSUPERVISED NORMALIZATION

### 3.1. Problems in O-space Canonical Correlation Analysis

The O-space canonical correlation analysis described in the previous section has the following two problems;

**(1)** Association of two spoken sentences by DP frequently fails, if the duration difference between them is longer than some threshold. This causes the degradation of speaker normalization accuracy of O-space canonical correlation analysis. To solve this problem, phonemes in two spoken sentences should be well associated. The easiest way is to use the word association in stead of sentence association. But it causes the degradation of speaker normalization accuracy because the normalization is not carried out using continuous speech. We need association of two spoken sentences without DP.

**(2)** Input speaker $B$ has to speak the same sentences as the standard speaker $A$ spoke because the dynamic programming (DP) is required to associate two spoken sentences in O-space canonical correlation analysis. This causes so much troublesome for the input speaker $B$. To solve this problem, canonical correlation analysis without DP should be required. If achieved, the input speaker $B$ can speak any kinds of sentences. This type of speaker normalization can be called unsupervised speaker normalization.

### 3.2. Principal of Unsupervised Normalization

As one of the methods to avoid the dynamic programming, we apply Viterbi segmentation to the speech data spoken by the input speaker $B$ using speaker independent phoneme HMMs. The segmentation of the speech data spoken by the standard speaker $A$ is carried out by hands because the standard speaker is fixed so that hand segmentation is done only once and causes no hard work.

After the segmentation, the phoneme representative vectors are computed by averaging the segmented phoneme data. Then the canonical correlation analysis is carried out using the phoneme representative vectors which are associated between the standard speaker $A$ and the input speaker $B$. We call this method unsupervised speaker normalization because the input speaker $B$ can speak any sentences and the dynamic programming is no more used.

### 3.3. Procedure

The procedure of unsupervised speaker normalization which we propose in this paper can be summarized as follows;

**STEP(1)** Speech data of the standard speaker $A$ and the input speaker $B$ are prepared. The sentences spoken by the input speaker $B$ is called normalization data. The sentences spoken by speakers $A$ and $B$ are different.

**STEP(2)** The normalization data spoken by speaker $B$ is segmented into phoneme sequence by Viterbi algorithm using speaker independent HMMs. The speech data of the standard speaker $A$ is segmented by hands into phoneme sequence. Then the representative (mean) vector of each phoneme data is computed. Using the associated pair of mean vectors of phoneme data between the speaker $A$ and $B$, the covariance matrices $\Sigma_{12}$, $\Sigma_{21}$ and $\Sigma_{22}$ are computed.

**STEP(3)** The subspace of the standard speaker $A$ is fixed to the observation space. The subspace axis $v_B$ of the input speaker $B$ is computed by Eq.(1) defined at O-space canonical correlation analysis using $\Sigma_{12}$, $\Sigma_{21}$ and $\Sigma_{22}$ computed at STEP(2). The normalized speech data of the input speaker $B$ is obtained by projecting the speech data into his subspace.

**STEP(4)** $HMM_A$s are computed using the speech data of the standard speaker $A$. The normalized speech data of the input speaker $B$ is recognized using the $HMM_A$s.

In this paper, we show experimentally the effectiveness of this unsupervised speaker normalization method using o-space canonical correlation analysis, by comparing with the supervised normalization method.

## 4. NORMALIZATION RESULT

### 4.1. Analysis and Database

We carried out phoneme recognition experiments for multiple speakers by supervised or unsupervised normalization using O-space canonical correlation analysis. The number of phonemes is 46 kinds. The experimental condition in the analysis and training is shown in Table1.

Table 1: Experimental condition
(AA:Acoustical Analysis)

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

Table 2: Database used for speaker normalization

| | |
|---|---|
| Standard speaker | MTK |
| Input speaker | MHO, MMY, MHT, MSH,<br>MYI (male) |
| | FYM, FTK, FKS, FKN (female) |
| Normalization data | Even numbered 75 sentences from ATR phoneme balanced set a,h,i (150 sentences) |
| HMM training data<br>Initial training: | Normalization data of MTK |
| Concatenated :<br>training | 500 sentences of MTK from ATR phoneme balanced set a-j |
| Recognition data | Odd numbered 75 sentences from ATR phoneme balanced set a,h,i of input speakers |

Table2 shows the database used for speaker normalization. The speech data used is ATR phoneme balanced sentence set which includes 10 speakers and 500 spoken sentences for each speaker. Standard speaker is fixed to MTK. Input speakers are five males and four females. Speaker normalization was carried out using 75 sentences. The phoneme HMMs were constructed using 500 sentences spoken by MTK.

75 sentences were projected to the speaker subspace and resulted in speaker normalized data. They were recognized by the phoneme HMMs of MTK.

### 4.2. Experimental Specification

We carried out six experiments shown in Table3. *No-norm* indicates that phoneme data included in the 75 sentences spoken by 9 speakers were recognized using phoneme HMMs of the speaker MTK without speaker normalization. *Sv-norm* indicates that supervised normalization proposed in [2] was carried out for the recognition data.

The experiments of *Usv1-same, Usv2-same, Usv1-diff* and *Usv2-diff* are the unsupervised normalization proposed in this paper. *Usv1-same* and *Usv2-same* indicate that speaker normalization was carried out using the same sentences spoken by the standard speaker and the input speakers. On the other hand, *Usv1-diff* and *Usv2-diff* indicate that speaker normalization was carried out using different sentences.

Table 3: Experimental specification

| | No normalization | Normalization | | | | |
|---|---|---|---|---|---|---|
| | | Supervised | Unsupervised | | | |
| | | | Labeling | | Sentences | |
| | | | Manual | Viterbi | Same | Different |
| *No-norm* | ○ | | | | | |
| *Sv-norm* | | ○ | | | | |
| *Usv1-same* | | | ○ | | ○ | |
| *Usv2-same* | | | | ○ | ○ | |
| *Usv1-diff* | | | ○ | | | ○ |
| *Usv2-diff* | | | | ○ | | ○ |

In this experiment, even numbered 75 sentences from seven kinds of three combination sets such as b,c,d set and c,d,e set were used as normalization data for nine input speakers.

In the experiments of *Usv1-same* and *Usv1-diff*, speech data spoken by the standard speaker was manually segmented into phoneme data. On the other hand, in the experiments of *Usv2-same* and *Usv2-diff*, speech data spoken by the standard speaker was automatically segmented into phoneme data using Viterbi decoding algorithm. In this case, conventional speaker independent HMM was employed for segmentation.

### 4.3. Experimental Results

The experimental results are shown in Table4 and Table5. From the Table4, we can say that unsupervised normalization shows almost the same recognition rate as the supervised normalization. The reason is that all the feature vectors do not contribute to correlate two subspaces between the standard speaker and the input speaker because the speech data between successive two phonemes are unstable. On the other hand, the representative (mean) vectors are stable so that it contributes to correlate two subspaces. As a result, *Usv2-same* improved the phoneme recognition rate by 12.7% in total compared with no normalization.

From the Table5, we can say that different sentences used as normalization data decreases the phoneme recognition rate a little. As a result, *Usv2-diff* improved the phoneme recognition rate by 11.6% in total compared with no normalization.

Table 4: Averaged phoneme recognition result(%) (using same sentences as normalization data)

| | Male | Female | Total |
|---|---|---|---|
| *No-norm* | 50.2 | 36.9 | 44.3 |
| *Sv-norm* | 59.0 | 60.0 | 59.6 |
| *Usv1-same* | **57.7** | **60.0** | **58.7** |
| *Usv2-same* | **56.2** | **58.1** | **57.0** |

Table 5: Averaged phoneme recognition result(%) (using different sentences as normalization data)

| | Male | Female | Total |
|---|---|---|---|
| *Usv1-same* | 57.7 | 60.0 | 58.7 |
| *Usv2-same* | 56.2 | 58.1 | 57.0 |
| *Usv1-diff* | **54.7** | **57.5** | **56.0** |
| *Usv2-diff* | **54.9** | **57.2** | **55.9** |

## 5. CONCLUSION

We described the unsupervised speaker normalization method based on the o-space canonical correlation analysis of the normalization data between standard speaker and input speaker. In stead of dynamic programming for association of all the feature vectors, we employed HMM based Viterbi decoding algorithm and associated the mean feature vectors of phoneme data. The phoneme recognition result showed this unsupervised normalization is almost equivalent with the supervised normalization we already proposed. The future works are to increase the number of representative vectors of phoneme data and to improve the HMM based Viterbi decoding.

## 6. REFERENCES

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

[2] Y.Ariki and S.Tagashira, "Effectiveness of Speaker Normalized HMM by Projection to Speaker Subspace", ICASSP97, SPCH8L.10, pp.1051-1054, 1997.

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