Intra-speaker Variability Suppression Method for Robust Speaker Recognition

July 2015

LU HAOZE

Graduate School of

Advanced Integration Science

CHIBA UNIVERSITY

（千葉大学学位申請論文）

Intra-speaker Variability Suppression Method for Robust Speaker Recognition

July 2015

LU HAOZE

Graduate School of

Advanced Integration Science

CHIBA UNIVERSITY

Contents

List of Tables

List of Figures

Acknowledgments

This thesis is carried out under the direction and guidance of Professor Shingo Kuroiwa and Associate Professor Yasuo Horiuchi from Chiba University, Japan. Naturally, I wish to express my sincere gratitude to them. Without their constant encouragement and guidance, this work would not have been possible.

I owe a great deal of thanks to the members of my thesis committee, Professor Josaphat Tetuko Sri Sumantyo, Professor Shingo Kuroiwa, and Associate Professor Yasuo Horiuchi from Chiba University, for their many helpful comments and suggestions.

I would like to appreciate to Associate professor Masafumi Nishida from Nagoya University, Associate professor Takahiro Shinozaki from Tokyo Institute of Technology, for there many helpful and significant comments related to this work.

I wish to thank all the members in the Kuroiwa & Horiuchi Laboratory. As a member of the laboratory, I benefit a lot from their friendship.

I am deeply grateful to my family for their encouragement and love during my life and studies.

Abstract

Speaker recognition is referred to the task of establishing identity of an individual using his or her voice characteristics. In practice, the following desirable properties of speech characteristics are the guarantee to achieve a higher accurate recognition rate: uniqueness that means no two persons should be the same in terms of the characteristic; permanence requires the characteristic should be invariant over a long period; acceptability indicates people are willing to accept the particular biometric system.

According to the speech content match or mismatch, speaker recognition is divided into text-dependent task and text-independent task. Text-independent speaker recognition is the much more challenging of the two tasks. In a text-independent speaker recognition system, results from the difference between the utterance for training and the utterance for test, phonetic variability have been considered to have a negative effect on performance of the system. These are due to changes of the acoustic environment, there are also some other undesirable factors that represents the inter-session variability, including health status, mood and aging.

In general, the same speaker's variation between training and test conditions is termed as intra-speaker variability, and it remains to be the most challenging problem in speaker recognition. If the intra-speaker variability (phonetic variability and the inter-session variability) can be successfully suppressed, a robust speaker recognition system will be realized.

From the different application of identity of a person, speaker recognition is categorized fundamentally into two distinct types of problems, which are identification and verification. Many previous studies have thrown light on the Gaussian mixture model (GMM). This has become the standardized method which has contributed to the leading edge of performance for speaker identification. Recently the support vector machine (SVM) is applied in the field of speaker verification research resulting in better performance. SVM is an efficient tool which can quickly and accurately classify the unknown speech samples into one category or the other based on margin maximization. Speaker verification focuses on whether a voice sample really being produced by supposed speaker or not.

GMM and SVM are statistically constructed using features of speech data in speaker recognition. However, the conventional features of speech data called mel-frequency cepstral coefficients (MFCC) used by GMM and SVM are not based on the removal of the phonetic variability and inter-session variability. Therefore, the conventional method here, cannot obtain a high accuracy level because speech features vary depending on the intra-speaker variability.

The first theme of this thesis, a subspace method based on principal component analysis (PCA) is proposed to remove phonetic variability. The proposed method constructs a subspace where the variance of data is maximized, under the assumption that phonetic variability is large in a lower order subspace called the phonetic-dependent subspace. An orthogonal higher order subspace obtained by PCA represents speaker information and this thesis called the subspace phonetic-independent subspace. A new speech feature is proposed based on a projection onto the phonetic-independent subspace where the phonetic variability is suppressed. The proposed method was shown to be effective in speaker identification experiments. As a result, the identification error rate was reduced by 21% by the proposed method compared with the conventional speech data based on MFCC.

On the other hand, the second theme adopts and extends the previous PCA-based method to reduce the inter-session variability. The basic idea is to transform input speech feature vectors to another subspace, where the inter-session variability is separated into a different subspace. Additionally, the inter-session is reduced in the original speech data subspace. The proposed method was shown to be effective in speaker identification experiments. As a result, identification error rate was reduced by 37% by the proposed method compared with the conventional speech features MFCC, and the proposed method was shown to be robust with the respect to the inter-session variation.

The third theme of this thesis is to propose a speaker verification method, by integrating an innovative phoneme-dependent method using a speech recognition technique. This technique selects the phonemes with a high contribution for speaker verification so as to overcome the shortcoming of inter-session variability. A speaker’s model can be represented by several various phoneme GMMs. The inter-session variability of phoneme GMMs can be constrained in an inter-session independent subspace constructed by a reduction method, which is termed as nuisance attribute subtraction (NAS) in this thesis. SVM-based speaker verification experiments demonstrate the improvements gained from the proposed method. The equal error rate was reduced by 19.4% by the proposed method compared with the conventional MFCC.

The last theme of this thesis proposes an application of speaker diariztion using a speaker verification technique to extract one desired speaker’s utterances from conversational speech. As a result of speaker diariztion experiments，the equal error rate was reduced by up to 43.7% compared with the conventional target speaker model, so that the system was shown to be effective.

Chapter 1

Introduction

1.1 Background

Biometrics generally refers to using personal traits or human physical characteristics to identify an individual. In theory, the biometrics authentication technology is expected, not only to protect the information safety, but also to guard against the threats of terrorism.

An ideal biological measurement is qualified to be a biometric. A biometric is supposed to have the following properties. It is no doubt that the most indispensable property is uniqueness. Uniqueness means no two persons should be the same in terms of the characteristic. In practice, however, permanence and acceptability also play decisive roles. Permanence requires the characteristic should be invariant with a long time, and acceptability indicates people are willing to accept the biometric system [1].

It occurs to our minds that face, iris, fingerprint, hand geometry and voice meet the aforementioned requirements. These biometrics have been proposed, researched, and evaluated. There is no single biometrics can effectively satisfy the needs of all authentication applications. Each biometrics appeals to a particular authentication application. The biometrics are used to identify an individual in roughly the same way in which Biometrics-based systems provide automatic, nearly instantaneous identity of a person by converting the biometrics into digital form and then comparing it against a computerized database [2].

With the continued rise of the needs by more and more companies to securely access information as rapidly as possible, voice biometrics has emerged as an effective solution to satisfy these challenges. A more accurate description for voice biometrics is called speaker recognition that refers to recognizing persons using their voice [3]. Given the situation that humans’ vocal tract shapes, larynx sizes and other parts of the voice production organs are different, namely, no two persons sound absolutely identical. State-of-the-art speaker recognition systems can be integrated as a part of a two-factor authentication process. It is combined with something like a password or PIN code to provide an extra layer of security to achieve more accurate authentication for confidential information and sensitive transactions.

Forensics is one of the important applications for speaker recognition technology. There is a lot of information exchanged between two parties including criminals in telephone conversations. Meanwhile, besides forensics, it has been predicted that ordinary customer will benefit from telephone-based services with integrated speaker recognition in the near future. For example, automatic password authentication or reset via the telephone. It is obviously that the advantages of such automatic authentication system can deal with thousands of telephone calls simultaneously. In addition to speech data in telephone, other spoken documents like teleconference meeting, TV broadcasts, and video clips from vacations are continually increasing and filled with our daily life. The process of extracting metadata like topic of discussion or speaker genders from these spoken documents would make information searching and indexing automated. Speaker diariztion, a typical example, means extracting speaker sections of the different participants from recordings using speaker recognition techniques [4].

1.2 Speaker recognition

Speaker recognition can be divided into text-dependent and text-independent systems. In text-dependent system, the recognition phrases are known beforehand or fixed. For example, the speaker is prompted to read a selected sequence of numbers. In text-independent system, on the other hand, there are no limitations on the content of speeches which the speakers are allowed to use. Namely, the utterances for training and the utterances for test may have different content completely. Consider the mismatch of phonetic, text-independent speaker recognition is the much more challenging task [3].

From different kind of application of identity of a person, speaker recognition is also categorized into two distinct types of problems fundamentally, that is identification and verification. The identification task refers to an unknown speaker is compared with the database of a set of speakers, and the best matching speaker is taken to be the identification result. On the other hand, the verification task refers to the process of determining whether a given sample of speech originated from the target speaker or not.

When the test utterance does not belong to any of the known speakers, it is called open-set speaker recognition, vice versa, a closed-set task refers to the provided speech sample to be determined from among a closed group of known speakers. The open-set nature of the process means much more challenge compared to the closed-set.

General speaking, in a text-independent speaker recognition system, phonetic variability has been considered to have the negative effect on performance. Due to the changes of the acoustic environment, there are also some other undesirable factors that represents the inter-session variability including health status, mood and aging [5]. In general, the same speaker's variation between training and test condition is termed as intra-speaker variability, and it remains to be the most challenging problem in speaker recognition. If the intra-speaker variability (phonetic variability and the inter-session variability) can be successfully suppressed, a robust speaker recognition system will be realized.

1.3 Thesis Structure

The remaining chapters of this thesis are composed as follows:

Chapter 2 provides an overview of current speaker recognition technologies and describes the fundamentals of feature extraction, speaker modeling and score normalization techniques.

Chapter 3 discusses the adverse effect on the accuracy of speaker identification by the phonetic variability and then proposes a phonetic variability compressed feature extraction method. The integration of GMM-based speaker identification system improves the identification accuracy.

Chapter 4 discusses the adverse effect on the accuracy of speaker identification rates by the inter-session variability and describes Nuisance Attribute Projection (NAP). NAP is utilized to remove the inter-session variability by the proposed feature extraction method. The integration of GMM-based speaker identification system further improves identification accuracy.

Chapter 5 inherits the thought of chapter 4 and analysis from the view of some selected phonemes with high contribution for speaker verification. The integration of GMM-SVM based speaker verification in the proposed phoneme-dependent system presents the improvements for suppressing inter-session.

In chapter 6, a new speaker indexing method using speaker verification technique is proposed to extract one desired speaker's utterances from the overlapped speech. The proposed method detected other speakers' speech from the observed speech itself. And then the computer has target speaker's speech overlapped with other speakers' speech to generate the overlapped speech model in order to improve the system.

Chapter 7 concludes the dissertation with a summary of the contributions of this research and suggests further directions for continuing research in robust speaker recognition.

Chapter 2

An overview of speaker recognition technology

2.1 Introduction

Speaker recognition is mainly defined as two tasks respectively by the requirement of different decisions. In speaker identification task, a speaker's speech sample is compared with a set of labeled speaker models. The label of the best matching is taken to be the identification result. In speaker verification task, a claimer utters the speech sample along with his/her ID and system needs to determine whether the given speech sample originated from the target speaker or not. According to the contents of speech, speaker recognition can be divided into two categories: text-dependent and text-independent. In a text-independent context, the system expects a pre-defined phrase to be spoken by the user. This approach allows very high accuracy to be achieved through the analysis of particular phrase and intonation characteristics of the speech. However, increased interaction between the user and the system is required as clients may need to produce a particular set of key-words or be prompted with a required phrase for the verification process. The text-independent case on the other hand, allows the speaker to use unrestricted speech for the recognition process. This is an inherently difficult task and it is applicable to forensic-based applications in which speaker-unaware verification is to be performed. In general, the development of robust speaker recognition aims to achieve these tasks through the following process: speech feature extraction (section 2.2), speaker modeling (section 2.3) and score normalization (section 2.4).

2.2 Speech feature extraction

In the context of automatic speaker recognition, speech processing refers to those operations applied to the raw auditory speech signal to produce a set of features suitable for use in a classifier. This feature extraction process is used to produce feature vectors holding the speaker information from the speech frames. These feature vectors are used to train speaker models and to perform classification. The selection of a feature set is critical for speaker recognition as it influences factors such as accuracy and robustness [6]. A significant benefit of analyzing speech in the cepstral domain is that linear time-invariant channel effects can be conveniently represented as mean offsets from the cepstral coefficients [12]. As a very common and efficient technique for speech processing, Mel Frequency Ceptral Coefficient (MFCC) [16] is based on human hearing perceptions. Some minor variations exits in the process steps of MFCC but the essential details are as below [8, 9, 10, 11].

2.2.1 Pre-emphasis

Pre-emphasis is traditionally applied before the process of windowing. First order high pass FIR filter is used to pre-emphasize the higher frequency components. This process serves to flatten the signal so that the spectrum consists of formants of similar heights.

2.2.2 Windowing

The hamming window is the most commonly used window shape in speech process. The feature extraction is performed on 20-30ms windows with 5-10ms shift overlapped between two consecutive windows. The speech signal is split into several frames such that each frame can be analyzed in the short time instead of analyzing the entire signal at once. Windowing is performed to avoid unnatural discontinuities in the speech segment and distortion in the underlying spectrum.

2.2.3 Fourier transform

The basis of performing Fourier transform is to convert the convolution of the glottal pulse and the vocal tract impulse response in the time domain into multiplication in the frequency domain. Spectral analyses signify that different timbres in speech signals corresponds to different energy distribution over frequencies. Therefore, Fourier transform is executed to obtain the magnitude frequency response of each frame and to prepare the signal for the next stage. In the practical application of speech process, Fast Fourier Transform (FFT) is commonly used.

2.2.4 Mel-frequency warping

The power spectrum is warped according to the mel-scale in order to adapt the frequency resolution to the properties of the human ear. Then the spectrum is segmented into a number of critical bands by means of a filterbank. The filterbank typically consists of overlapping triangular filters. The logarithmic mel-scale is estimated by

$$Mel(f) = 2595log_{10}\left(1 + \frac{f}{700}\right), \tag{2.1}$$

$$c_{lmfb}(l) = \log m(l). \tag{2.2}$$

The role of logarithm here is to separate the convoluted components of glottal pulse and the vocal tract impulse response.

2.2.5 Cepstrum

Mel-Cepstral coefficients are derived by transforming the log-energies of the filterbank outputs using a discrete cosine transform (DCT). DCT encodes the mel logarithmic magnitude spectrum to the mel-frequency cepstral coefficients MFCC.

$$c_{mfcc}(i) = \sqrt{\frac{2}{N}\sum_{l=1}^{L} c_{lmfb}(l)\, cos\left\{\left(l - \frac{1}{2}\right)\frac{i\pi}{L}\right\}} \qquad (2.3)$$

Delta coefficients are generally appended to each feature to capture the dynamic properties of the speech signal. These coefficients approximate the instantaneous derivative of each of the cepstral coefficients by finding the slope coefficient when performing a least-squares linear regression over a window of consecutive frames with a window length of 5-10 frames.

A common method of improving the robustness of a feature set is cepstral mean subtraction (CMS) [20]. This process reduces the effects of channel distortion by removing the mean from cepstral coefficients [16]. Essentially, CMS can be viewed as a high-pass filter applied to a set of feature vectors. Although the technique is effective at reducing the effects of channel distortion, it has been shown to also remove beneficial speaker-specific information from the speaker recognition system [19]. In order to alleviate the effect of additive noise, cepstral mean and variance normalization (CMVN) [20] was proposed as an extension to CMS.

2.3 Speaker Modeling

In recent years, significant changes have been made to the way in which speakers' characteristic is robustly modeled in speaker recognition systems. Approaches of significant influence include vector quantization (VQ) codebooks [25], Gaussian mixture models (GMM) [31], Gaussian Mixture Model-Universal Background Model (GMM-UBM) [47] and Support Vector Machine (SVM) using GMM super-vectors [48, 58].

2.3.1 GMM-based speaker identification

Significant progress was achieved in speaker recognition technology with the introduction of Gaussian mixture models (GMM) such that they are now classical in text-independent speaker identification configurations [29]. A number of factors have contributed to the acceptance of GMMs as the standard in speaker identification including their high accuracy, ability to scale training algorithms for large data sets, and their probabilistic framework [31].

A single Gaussian mixture model can be viewed as several overlapping Gaussian distributions having the ability to reflect the short-term spectral density of a speaker's speech. The density of a sample $\boldsymbol{x}_t$ from a *D*-dimensional multivariate Gaussian distribution is given by

$$\begin{aligned} &N(\boldsymbol{x}_t|\boldsymbol{\mu}_{sm}, \ \Sigma_{sm}) \\ &\qquad = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} exp\left\{-\frac{1}{2}(\boldsymbol{x}_t-\boldsymbol{\mu}_{sm})^T{\Sigma_{sm}}^{-1}(\boldsymbol{x}_t-\boldsymbol{\mu}_{sm})\right\}, \end{aligned} \tag{2.4}$$

with the distribution means of $\boldsymbol{\mu}_{sm}$, covariances defined by Σ_{sm} and T represents the number of samples in the utterance.

Due to the nature of the speech signal to continually change in spectral density, a number of Gaussian components (typically 256 components) are necessary to model the speaker-dependent features over the length of an utterance [31]. The collection of these Gaussian components results in the complete Gaussian mixture model. Central to the idea of Gaussian mixture speaker modeling is the assumption that each feature vector extracted from a test speech segment was produced by only one of the GMM components.

$$\sum_{m=1}^{M} \omega_{sm} = 1. \tag{2.5}$$

In order to place emphasis on those GMM components that better represent the more commonly observed characteristics of the speaker, each component is assigned a weight during model training. The weights ω_{sm} are then utilized in the classification process where the utterance $X = \{\boldsymbol{x_1} \ldots \boldsymbol{x_T}\}$ is compared to GMM of *M* components by the joint density,

$$p(X|\lambda_s) = \prod_{t=1}^{T}\sum_{m=1}^{M} \omega_{sm} N\big(x_t\big|\boldsymbol{\mu_{sm}},\ \Sigma_{sm}\big). \tag{2.6}$$

An identification result is obtained as speaker s with maximum log likelihood shown in Eq. (2.7)

$$s = \arg\max_{i} \sum_{t=1}^{T} \log P(x_t \mid \lambda_i)\,. \tag{2.7}$$

2.3.1.1 Maximum Likelihood Estimation

The expectation-maximization (E-M) algorithm [33] is a common way to train GMMs [32]. The motivation of the E-M algorithm is to estimate a new and improved model λ from the current model $\hat{\lambda}$ using the training utterance X such that the probability $p(X|\lambda) \geq\ p\big(X|\hat{\lambda}\big)$. This is an iterative technique whereby the new model becomes the current model for the following iteration.

As the name suggests, expectation-maximization involves two steps; expectation and maximization. The expectation step, or E-step, calculates the expected value of the model from the training utterance X in order to estimate the information that is missing from the model. The maximization step, or M-step, uses this information to adjust and improve the current model parameters.

Specifically, the E-M algorithm attempts to maximize the auxiliary function $Q\big(\lambda;\hat{\lambda}\big)$. This is generally implemented using Jensen's inequality ensuring $p(X|\lambda) \geq$ $p\big(X|\hat{\lambda}\big)$. The auxiliary function can be formulated as

$$Q\big(\lambda;\hat{\lambda}\big) = \sum_{t=1}^{T}\sum_{m=1}^{M} P(m|\boldsymbol{x}_t) log w_m g(\boldsymbol{x}_t|\boldsymbol{\mu}_m, \Sigma_m), \tag{2.8}$$

where $P(m|\boldsymbol{x})$ forms the E-step or expected probability of component m being responsible for producing observation $\boldsymbol{x}$ using

$$P(m|\boldsymbol{x}) = \frac{\widehat{\omega}_m g(\widehat{\omega}|\widehat{\boldsymbol{\mu}}_m, \widehat{\sum}_m)}{p(\boldsymbol{x}|\hat{\lambda})}. \tag{2.9}$$

The M-step then sees the auxiliary function $Q(\lambda; \hat{\lambda})$ maximized using (2.8). This maximization results in the GMM parameters being estimated as

$$\omega_m = \frac{n_m}{T} \sum_{t=1}^{T} P\left(m|\boldsymbol{x}_t, \hat{\lambda}\right), \tag{2.10}$$

$$\boldsymbol{\mu}_m = \frac{1}{n_c} \sum_{t=1}^{T} P\ (m|\boldsymbol{x}_t, \hat{\lambda})\boldsymbol{x}_{t,} \tag{2.11}$$

$$\Sigma_m = \frac{1}{n_m} \sum_{t=1}^{T} P\left(m|\boldsymbol{x}_t, \hat{\lambda}\right)\boldsymbol{x}_t \boldsymbol{x}_t^T - \boldsymbol{\mu}_t \boldsymbol{\mu}_t^T, \tag{2.12}$$

where n_m is the component occupancy count from all the observations of the utterance X.

A suitable technique is desired to initialize models as it defines the final model parameters and how rapidly the E-M process will converge. The initial GMM is typically defined using the k-means algorithm often used in the vector quantization (VQ) approach [34]. The k-means algorithm is also based on an iterative approach in which the clustering of training feature vectors is performed through the estimation of cluster means or code vectors [35].

2.3.1.2 Maximum A Posteriori Estimation

A more recent and commonly used approach for GMM training is based on Bayesian estimation theory and is termed maximum a-posteriori (MAP) adaptation [36]. MAP adaptation incorporates prior knowledge regarding the nature of speech into the speaker model prior to training, or adapting, the model parameters to exhibit the speaker-specific characteristics. This is accomplished by initializing each new speaker model with the parameters of the universal back-ground model (UBM).

The UBM is a GMM trained from a large selection of representative speech, often using the maximum likelihood (ML) approach described in section (2.3.2). In using a large cohort of training data, the true characteristics of speech can be more reliably modeled in the UBM. MAP adaptation can then exploit this wealth of information to produce more robust speaker models, particularly when subject to limited training data [21]. Further, the proposal of MAP adaptation saw the appearance of more detailed speaker models. The number of components is expanded from around 64 using the ML approach to over 2048 with MAP adaptation.

During the training of speaker models, common practice is to adapt only the means of the mixture components of the UBM to match the target speaker's characteristics. The NIST SREs have demonstrated the benefits of allowing GMM speaker models to maintain the same weights and covariances as the UBM as the mean GMM parameters contain the most speaker information [37].

During training, the model parameters λ are constrained to satisfy the prior distribution of speaker model parameters using the criterion:

$$\lambda^{MAP} = arg_{\lambda} maxp(\lambda|X), \qquad (2.13)$$

where $P(X|\lambda)$ is the posteriori probability of the model parameters after observing the training set X . Applying Bayes theroem, the MAP problem is solved by

$$\lambda^{MAP} = arg_{\lambda} maxp(\lambda)p(X|\lambda), \qquad (2.14)$$

which is the prior distribution $p(\lambda)$ multiplied by the likelihood of the training data for the given model parameters. The joint likelihood of this equation is solved using the E-M algorithm.

The MAP-adapted means $\boldsymbol{\mu}_m^{MAP}$ of Gaussian component m can be adapted from the prior distribution means $\boldsymbol{s}_m$ using

$$\boldsymbol{\mu}_m^{MAP} = \alpha_m \boldsymbol{s}_m + (1 - \alpha_m)\boldsymbol{\mu}_m^{ML}, \tag{2.15}$$

where $\boldsymbol{\mu}_m^{ML}$ are the means estimated using maximum likelihood estimation and α_m is the mean adaptation coefficient defined as

$$\alpha_m = \frac{n_m}{n_m + \tau_m}, \tag{2.16}$$

where n_m is the component occupancy count for the training data and τ_m is the relevance factor, typically set between 8 and 32. Based on these equations, it can be seen that this MAP adaptation is essentially a combination of the prior distribution means and the ML estimated means given the training data whose relative weighting are controlled by α_m.

2.3.2 SVM-based speaker verification

Recently, machine learning techniques have been adapted to the task of pattern recognition. These modelling techniques are trained to differentiate between classes by learning from examples of both the target and non-target. Support vector machine (SVM) has received significant focus in pattern recognition literature [60]. The discriminative nature of the SVM has been successfully applied to speaker verification [21, 39]. The training of Speaker SVM requires example from both the target speaker and a selection of impostor speaker. The following contents in this section provides an overview of the speech feature for SVM-based speaker verification in section 2.3.2.1 and section 2.3.2. SVM is described in detail in section 2.3.2.3 and section 2.3.2.4.

2.3.2.1 GMM-UBM

The most common speaker verification system architecture employed is the GMM-UBM configuration, first proposed by Reynolds [10]. This approach represents state-of-the-art technology when combined with robust modeling techniques [38, 39]. As its name suggests, the universal background model models the characteristics of speech from a representative population of speakers. For this task, a large amount of data is required to train a UBM, which in turn, allows the model to consist of a large number of components to better represent the speaker characteristics in the training data.

As mentioned previously, the main task of the UBM is to provide the prior distribution when employing MAP adaptation for the training of some speaker GMMs. There are several benefits in using this training approach that have accounted for significant performance improvements in GMM-based systems.

Firstly, when training data is not available for the adaptation of components in the UBM, the speaker model parameters revert to those in the UBM to provide a more robust speaker model. In contrast, when a lot of training data is available for a given GMM component, the speaker model parameters approach those of the ML estimate.

The final purpose of the UBM is to represent the null hypothesis or background speaker population when using expected log-likelihood ratio (ELLR) scoring. In ELLR scoring, classification scores are represented as a log-ratio of the probability of a target trial and an impostor trial. Given the speaker model λ and the background model λ_{UBM}, the ELLR can be calculated using

$$\Lambda(s) = E\left[log\frac{p(\boldsymbol{x}_t|\lambda)}{p(\boldsymbol{x}_t|\lambda_{UBM})}\right]. \tag{2.17}$$

In this configuration, the universal background model (UBM) is a reference speaker model to which the target speaker model is compared during the classification process to produce a log-likelihood ratio [47]. During classification, the log-likelihood-ratio (LLR) can be calculated from the target speaker model and background model. In essence, this configuration can be viewed as the UBM normalizing for the characteristics of the impostor population that have potential to affect the classification score.

2.3.2.2 GMM-UBM super-vector

In the context of the GMM-UBM configuration for speaker recognition, GMMs are trained using features extracted from a speech sample of a speaker. GMMs are generated by mapping the parameters of the UBM through MAP adaptation to represent the speaker using the corresponding training data.

The parameters of adaptable GMM include the component mixture weights ω_m the means $\boldsymbol{\mu}_m$ and the covariances Σ_m of the Gaussians. Accordingly, the GMM likelihood function is given as

$$g(x) = \sum_{m=1}^{M} \omega_m N(\boldsymbol{x}; \boldsymbol{\mu}_m, \ \Sigma_m). \tag{2.18}$$

Typically, only the means of the UBM are adapted as they possess the majority of speaker-dependent information within the model [47]. Consequently, the majority of speaker dependent characteristics from a trained speaker GMM can be represented by the adapted component model mean offsets from the UBM model means.

A GMM-UBM super-vector can be obtained by concatenating each of the mean vectors, $\boldsymbol{\mu} = [\boldsymbol{\mu}_1^T \ldots \boldsymbol{\mu}_m^T]$ of an adapted GMM. A GMM-UBM super-vector can be formed from this adapted model as shown in Fig. 2.1

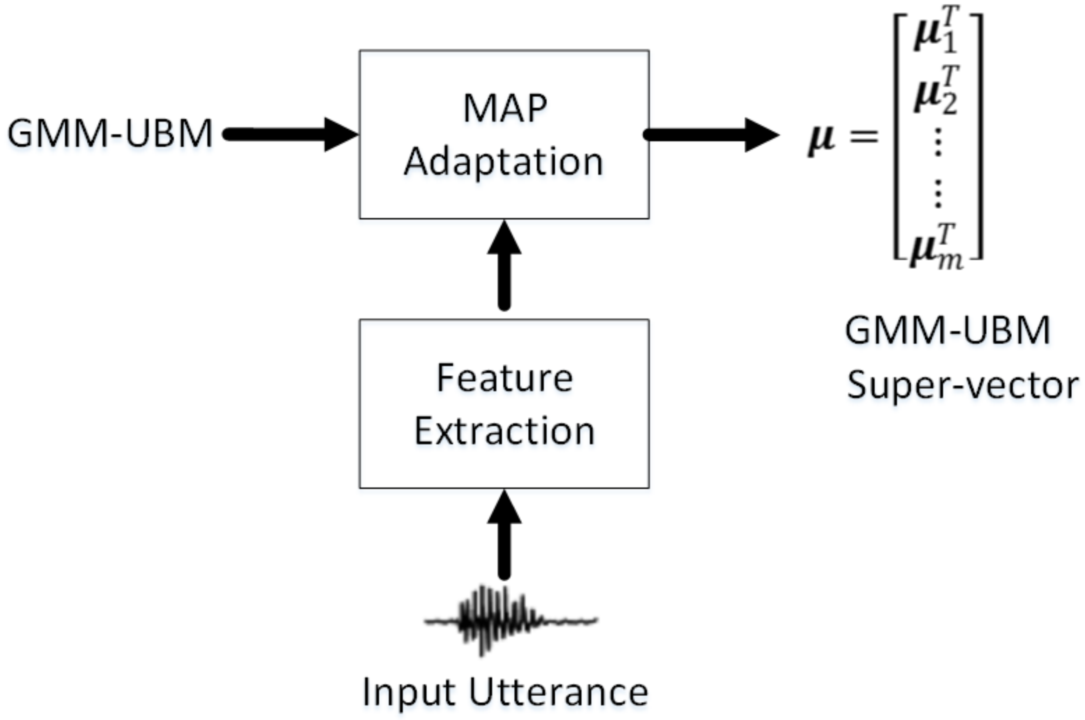

GMM-UBM
MAP Adaptation
Feature Extraction
Input Utterance
$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1^T \\ \boldsymbol{\mu}_2^T \\ \vdots \\ \vdots \\ \boldsymbol{\mu}_m^T \end{bmatrix}$
GMM-UBM Super-vector

Figure 2.1 Concept of GMM-UBM super-vector

2.3.2.3 Support Vector Machines

Guyon, et al. developed Supporter Vector Machines (SVM) [100] based on the theory of structural risk minimization that allowed a specific capacity point to be found which minimizes generalization error in turn and provides the basis for the development of the SVM. Guyon discovered that linear classifiers require the capacity control to maximize potential to generalize from the training data to the classification of unknown data. The capacity of a classifier means the number of adjustable or free parameters. When having a large generalization error for test data, a classifier with a large capacity is more possible to over fit the training data. A small capacity of a classifier may prevent the classifier from adapting to the task at all.

SVM is a linear classifier to minimize the generalization error. As mentioned, SVM is based on the theory of margin maximization. SVM based classification lies a kernel function. The purpose of the kernel is to convert input feature vectors to a higher dimensional space. Actually it is possible to specify a linear kernel in (2.19) that operates in the input feature space.

$$f(x) = \sum_{i=1}^{N} \alpha_i t_i K(x, x_i) + d. \qquad (2.19)$$

Where x represents the input vector, x_i the support vectors, t_i the ideal outputs (±1) and α_i the respective weights, $\sum_{i=1}^{N} \alpha_i t_i = 0$, and $\alpha_i > 0$

In a linearly separable instance, the kernel function is given as $K(\cdot,\cdot) = x \cdot x_i$ which is known as a linear kernel as it simply find the dot product of input vectors in al linear space. In a non-linearly separable instance, the purpose of a non-linear SVM kernel is to allow non-linear separation to be applied to a data set by mapping the input vectors to a high-dimensional space where linearly separable can be achieved. So the kernel can be more stated generally as,

$$K(\cdot,\cdot) = \phi(x_i) \cdot \phi(x_j), \qquad (2.20)$$

where ϕ is mapping function employed to convert input vectors to a desired higher dimensional space. The mapping function is chosen on a basis that satisfies the higher dimensional space where linear separation is to be performed.

A 2-D plot of several linearly separable observations from x is depicted in Figure 2.2. Separation of the positive and negative classes is performed using a “separating” hyperplane as indicated by the solid line in the plot. The points that reside on the hyperplane are given by $\mathrm{\omega \cdot x + b = 0}$ where ω is the normal to the hyperplane and |b| / |ω| defines the distance between the hyperplane and the origin.

A margin exists on either side of the hyperplane (depicted as dashed lines in Figure 2.2) to define the boundaries of each class such that

$$y_i(\omega \cdot x + b) \geq 1. \qquad (2.21)$$

The training objective of SVM training is to maximize this margin through the optimal positioning of the hyperplane. Given the width of the margin is $2/\|\omega\|$, the hyperplane margin can be maximized by minimizing $||\omega||^2$ subject to the constraints of (2.21).

Those training examples that are located on the class boundaries, such that the equality in (2.21) is satisfied, are termed as support vectors and are usually only a small subset of the training data. As their name suggests, support vectors are the training examples that support or define the hyperplane. All other training examples that are not selected as support vectors provide no information in the training of the SVM such that the removal of these examples from the training set would result in the same hyperplane being found.

The SVM training algorithm is often represented in terms of Lagrange multipliers αi as it facilitates the explanation of training from non-separable data [60]. The optimal position of the hyperplane (the normal of ω can be determined by maximizing

$$\sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j \, y_i y_j \, x_i \cdot x_{j,} \qquad (2.22)$$

subject to the following constraints:

$$\omega = \sum_i \alpha_i y_i \, x_{i,} \qquad (2.23)$$

$$\sum_i \alpha_i y_i = 0. \qquad (2.24)$$

In this form, the training examples allocated $\alpha_i > 0$ are support vectors and lie on the hyperplane margin. For this reason, the Lagrangian multiplier α_i is often referred to as a support vector weighting or coefficient.

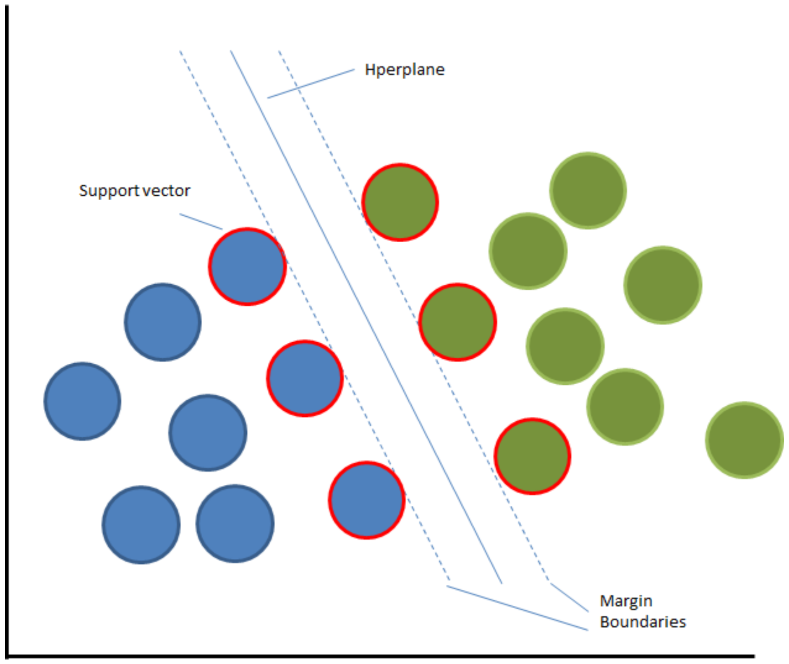


Figure 2.2 Concept of SVM training

2.3.2.4 Support vector machines for speaker verification

Support vector machine (SVM) classification is particularly suited to the task of speaker verification as the objective is to determine whether a given speech sample belongs to the target speaker or not. In this sense, the discrimination of the target speaker from other speaker is performed by the hyperplane of a trained SVM.

Being a discriminative classifier, the training of an SVM requires examples from both the target and impostor speaker classes; this imposter role is fulfilled by a set of observation from a representative background speakers.

As described in the previous section, the SVM relies on the negative examples of the background dataset to provide discriminatory information against client data during the training process. In the context of speaker verification, it is common for the number of impostor observations used in SVM training to significantly outweigh the number of positive speaker examples. Consequently, the SVM get most of its discriminatory information from the background dataset. The background dataset must, therefore, consist of suitable impostor examples to ensure good classification performance.

2.4 Score Normalization techniques

Speaker verification systems developed for practical and real-world applications are often faced with several factors that contribute to significant accuracy loss. These factors include additive noise, channel distortion, handset mismatch and human changes due to health and age [46, 21].

Normalization techniques are employed to model and counteract the effects of these adverse factors [6, 42, 48]. Normalization in speaker verification systems is the process of removing statistical error from features, models or scores in order to allow for a more direct comparison of the true speaker characteristics being modeled.

Score normalization refers to the scaling of the classification score distribution based on a set of parameters obtained using a set of impostor trials [42]. The cohort of impostor utterances used to perform these trials is from dataset held out from the evaluation data. Cohort normalization [49] attempts to normalize classification scores in the same manner as the UBM. Rather than using a world model, an impostor speaker model with similar characteristics to the target speaker is dynamically selected from a small cohort of similar impostor speakers. This cohort is selected based on a distance metric between models. Higgins, et al. [50] showed that using a cohort of speakers ‘close’ to the target speaker left the speaker verification system vulnerable to very dissimilar speakers. Reynolds resolved this issue by introducing a method of selecting a wider range of speakers to make up the cohort model set [51]. Restricting the degree of similarity between the models in the cohort set provided a much more robust cohort model. The performance advantages provided through score-based normalization have made it a commonly employed technique in both GMM and SVM-based speaker verification systems in recent NIST SREs [54]. Auckenthaler, et al. [42] presented a study on two most common forms of score normalization: Zero and Test normalization (Z-norm and T-norm, respectively).

2.4.1 Zero-normalization

Zero-normalization (Z-norm) attempts to compensate for training variations that exist between speakers in the verification process [42]. Z-norm is performed after the training of a speaker model but prior to the testing phase.

The technique firstly calculates the mean μ_Z and variance σ_Z from the impostor score distribution which is estimated by trialing a set of impostor utterances against a given speaker model. During testing time, the unnormalized score distribution $\rho(x)$ is normalized by,

$$\rho z(x) = \frac{\rho(x) - \mu_z}{\sigma_z}. \tag{2.25}$$

2.4.2 Test-normalization

Z-norm is the approach to compensate for the training conditions of the target model. However, the score distribution from verification trials can also be effected by the conditions exhibited by the test utterance. Test-normalization (T-norm) was proposed by Auckenthaler, et al. [42] to address this issue. In contrast to Z-norm, T-Norm is employed during the testing phase using the encountered test utterance.

The encountered test utterance is firstly trialed against a set of impostor models which are trained from a cohort of impostor utterances. The mean μz and variance σ_Z from resulting impostor score distribution is then calculated. The classification score of the utterance against the target speaker model is finally normalized using equation (2.25). The advantage of T-norm is that acoustic mismatch is avoided due to the same utterance being used to estimate impostor parameters.

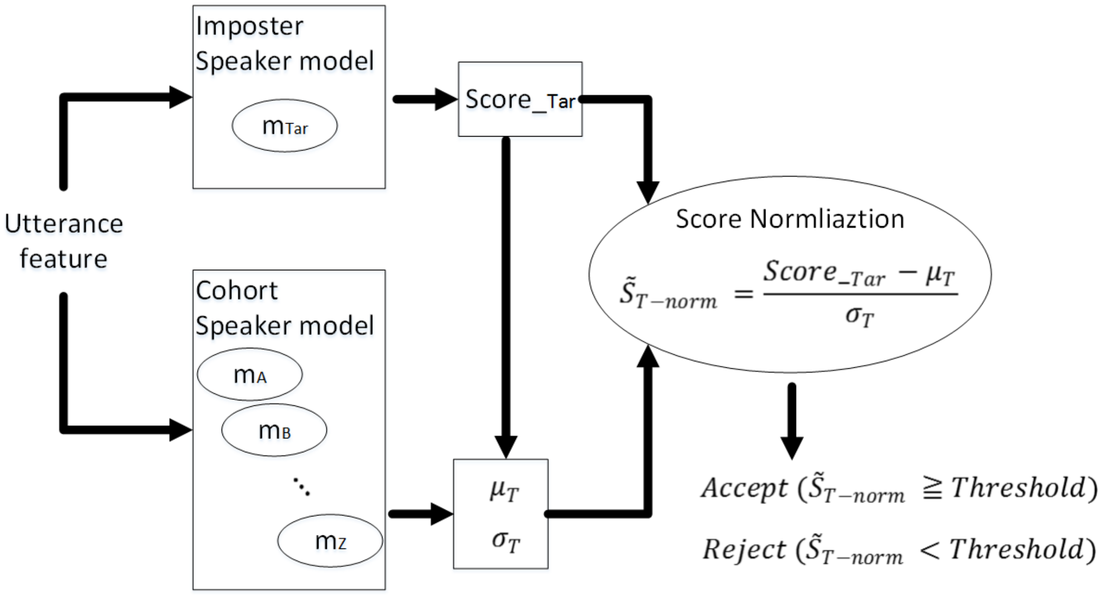

Imposter
Speaker model
mTar
Utterance
feature
Cohort
Speaker model
mA
mB
mZ
Score_Tar
μ_T
σ_T
Score Normliaztion
$\tilde{S}_{T-norm} = \frac{Score_{\_Tar} - \mu_T}{\sigma_T}$
$Accept\ (\tilde{S}_{T-norm} \geqq Threshold)$
$Reject\ (\tilde{S}_{T-norm} < Threshold)$

Figure 2.3 Concept of T-norm-based speaker verification

2.5 Summary

Speaker modeling and GMM-based speaker identification is discussed regarding the classical technology offered through Gaussian mixture models. Then SVM-based speaker verification using GMM-UBM super-vectors was described in detail to highlight the benefits offered through this generative model that had led to its widespread acceptance.

Score normalization was detailed as the successful method used to counteract statistical variations in the GMM-UBM configuration. Zero normalization (Z-norm) and test normalization (T-norm) were highlighted as the most commonly employed score normalization techniques in speaker recognition systems.

In the chapter, the typical cepstral-based speech feature extraction process (i.e., MFCC) for text-independent speaker recognition is also described. However, result from the difference between the utterance for training and the utterance for test, phonetic variability has the negative effect on accuracy. In additional, due to the changes of the acoustic environment, there are also some other undesirable factors that represents the inter-session variability including health status, mood and aging. Phonetic variability and inter-session variability in MFCC are merged as the intra-speaker variability which is needed to be reduced to improve the accuracy.

The focus of the following chapters is given to the techniques developed to increase the robustness of the feature extraction process. The techniques are highlighted as effective approaches to help construct robust speaker recognition system.

Chapter 3

New speech feature with less phonetic variability

3.1 Introduction

Speaker recognition technique makes it possible to use the speaker's voice to verify their identity and control access to services such as voice dialing, banking by telephone, telephone shopping, database access services, information services, voice mail, security control for confidential information areas, and remote access to computers, etc. Speaker identification has potentially more applications than verification, which is mostly limited to security field. For instance, in speaker tracking the task is to locate the segments of given speakers in an audio stream [87, 88, 89, 90]. However, there are still many problems to solve in this technology. The most essential one among them is to find a suitable feature to discriminate the speaker. This chapter focuses on the method of suppressing phonetic variability in speech feature which influences the performance for Text-independent (TI) speaker identification.

The process of speaker identification depends on the feature information extracted from his/her speech data. Many research shows that the coefficients based on frequency domains are efficient in speaker recognition. Most of the feature representations that are often used in the literatures of speaker recognition such as Mel Frequency Ceptrum Coefficient (MFCC) [16], Liner Prediction Ceptrum Coefficient (LPCC) [16; 91], Delta Coefficient [16], etc, are all short term spectral based features. Among them, the MFCC is undoubtedly the most widely used and successful feature. In this decade, MFCC has been often used in speaker identification. It is a problem that speech feature varies depending on sentences and time difference for TI speaker identification task [10]. Especially, the variation of phonetic variability strongly affects the performance of speaker identification. If this phonetic variability in his/her speech data can be suppressed, a robust speaker identification system will be realized by using speech data having less phonetic variability. However, it is difficult to suppress phonetic variability included in speech data completely at present.

In TI speaker identification, GMM (Gaussian Mixture Model) [86] has been conventionally used and is statistically constructed using features [95] (e.g. MFCC) of speech data. However, MFCC is used as a common feature vector for speech recognition and speaker recognition and it doesn’t consider the influence of phonetic variability to the performance of TI speaker identification. Thus, prior researches about speaker recognition based on a speaker space method [96, 97], pointed out that the speech feature variation is mainly caused by the variation of the phonetic variability in speech data. Although some works have been used Principal Component Analysis (PCA) in speaker recognition [5, 92, 93］, they did not discuss the meanings of each axis of PCA. Hence, we consider the meanings of the axes and propose a new feature vector that does not use axes which mainly represent phonetic variability. In the method, it is assumed that some axes with larger dispersion in the speech feature space represent phonetic variability when having many speech data prepared. PCA is utilized in order to confine phonetic variability into some axes. The phonetic variability is reduced by the projection using PCA eigenvectors without some principal components that have larger contribution rates, i.e. eigenvalues. That is, a subspace constructed with the some principal components is considered as "phoneme-dependent space" and a subspace constructed without these components is considered as "phoneme-independent space" in this paper.

The remainder of this chapter is organized as follows. Section 3.2 introduces a new feature vector of our proposed method with PCA. In section 3.3, comparative experiments and discussion is carried out to show an effectiveness of the proposed method with the conventional MFCC. Finally, section 3.4 gives the summary of this chapter.

3.2 PCA-based phonetic variability suppression for speaker identification

Figure 3.1 shows a block diagram of our proposed TI speaker identification system. It consists of a training module and a recognizing module.

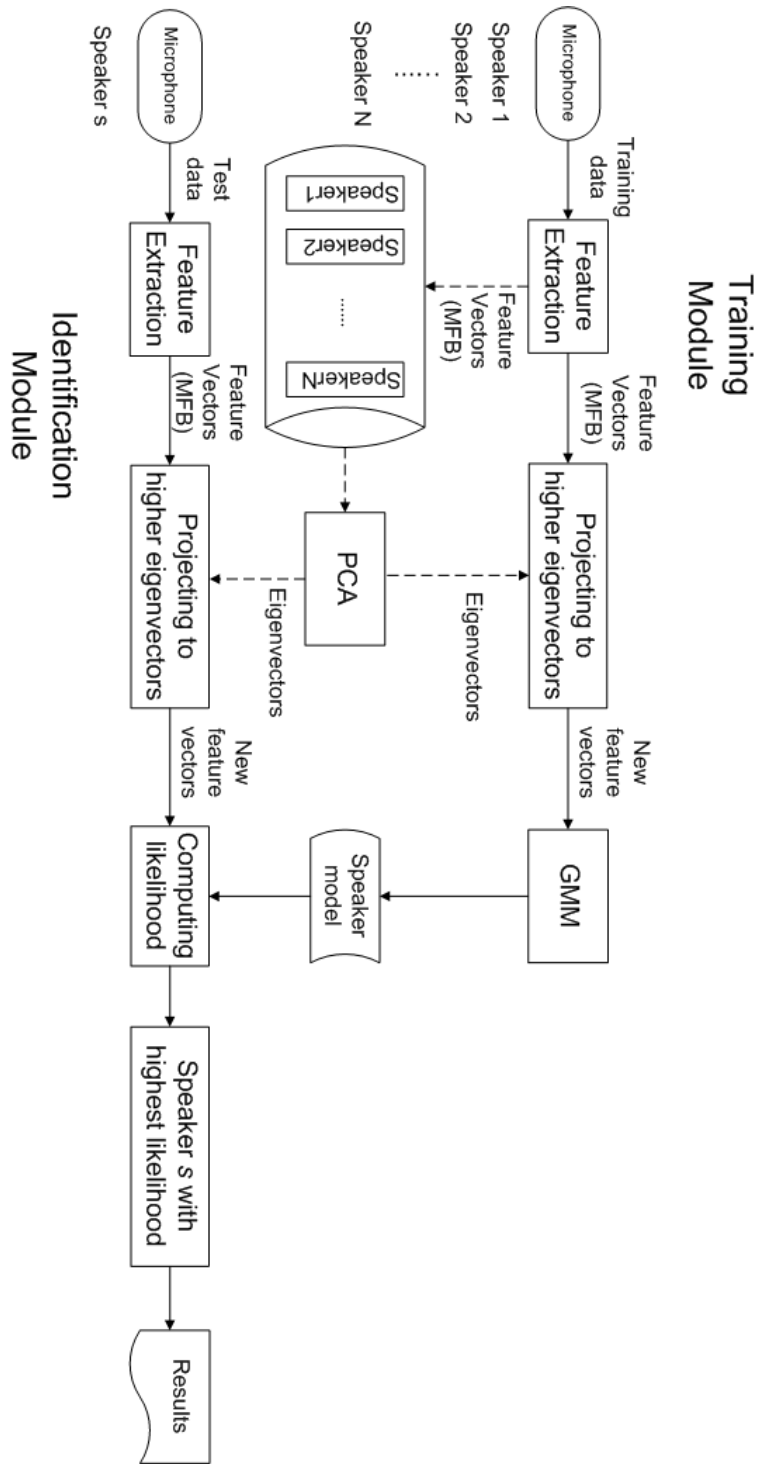


Figure 3.1 the proposed speaker identification system

In the proposed system, it is assumed that phonetic variability has larger dispersion in speech feature space and the information is mainly represented with the some principal components calculated using many speech data.

At the process of training module, a robust TI speaker model can be realized for phoneme differences by projecting feature vectors of each speaker's speech data into a "phoneme-independent subspace". In order to realize the projection to the phoneme-independent subspace, PCA is employed. First, PCA obtains eigenvectors from many speech data that include many phonetically-rich texts by many speakers. Next, each speaker’s data is projected into a subspace using the eigenvectors that do not have several largest contribution rates and several smallest contribution rates. The detail of the projection is described in the following subsection. As original feature vectors, the Mel-Frequency Filter Banks (MFB) is used instead of MFCC. 24-channel MFB is used in this paper. Finally, after projection, the new feature vectors having less phonetic variability are used to train speaker model and the trained model is stored to a database with ID of each speaker.

3.2.1 Projection algorithm with PCA

MFCC is commonly used in speaker identification. MFCC is obtained from the MFB using discrete Cosine Transform (DCT). However DCT is not designed to transform a space by considering a data distribution as well as correlation of feature parameters. In general, PCA is used to diagonalize a data covariance matrix and to decorrelate each dimension of the feature parameters. In this study, PCA is utilized to separate the “phoneme-dependent subspace”.

A set of observed n-dimensional training feature vectors of all speakers in an observation space can be denoted by $\{x_t\}(t=1,2\ldots,N)$ .The mean vector of $\sum_{t=1}^{N} x_t$ can be computed by

$$\bar{x} = \frac{1}{N}\sum_{t=1}^{N} x_t \ , \tag{3.1}$$

here, N denotes frames of all speakers’ speech data. And a covariance matrix by

$$R_X = \frac{1}{L}\sum_{t=1}^{N}(x_t - \bar{x})(x_t - \bar{x})^T \ . \qquad (3.2)$$

The covariance matrix R_X (n by n) can be composed into eigenvectors and eigenvalues as follows:

$$R_X = \Phi\Lambda\Phi^T , \qquad (3.3)$$

where Λ is a diagonal matrix whose diagonal components are eigenvalues $\lambda_i = (i = 1,\cdots,k,\cdots,n)$ of R_X . Φ is a matrix whose columns are eigenvectors $\varphi_i(i = 1,\cdots,k\cdots,n)$ of R_X corresponding to the eigenvalues in the matrix Λ.

In this theme, It is assumed that a space constructed with some principal components having larger contribution rates corresponding to the large eigenvalues up to k is considered as "phoneme-dependent subspace" and a subspace constructed without these components is considered as "phoneme-independent subspace". This means that the phoneme-dependent subspace can separated by means of projecting the input speech data to the eigenvectors corresponding to the higher eigenvalues from k+1 to m $(k+1 \le m \le n)$ obtained by performing PCA for feature data.

So the projection of the t-th frame data can be expressed as follow:

$$\begin{bmatrix} \varphi_{k+1,1} & \cdots & \varphi_{k+1,n} \\ \vdots & \ddots & \vdots \\ \varphi_{m,1} & \cdots & \varphi_{m,n} \end{bmatrix} \begin{bmatrix} x_{t,1}^{(s)} \\ x_{t,2}^{(s)} \\ \vdots \\ \vdots \\ x_{t,n}^{(s)} \end{bmatrix} = \begin{bmatrix} \hat{x}_{t,1}^{(s)} \\ \hat{x}_{t,2}^{(s)} \\ \vdots \\ \vdots \\ \hat{x}_{t,m-k}^{(s)} \end{bmatrix} , \qquad (3.4)$$

where $x_{t,j}^{(s)}$ $(j = 1,\cdots,n)$ are n dimensional feature vectors （i.e. 24-channel MFB was used in the experiment of this paper and n=24) of speaker s observed in an n-dimensional observation space. PCA is applied to all the speech data of all the speakers and n dimensional eigenvectors $\varphi_{i,j}(i=k+1,\cdots,m;\ j=1,\cdots,n)$ are obtained by the eigenvalue decomposition. $\hat{x}_{i,j}{}^{(s)}(j=1,\cdots,m-k)$ is the (m-k) dimensional new feature vectors that have less phonetic variability after projection. Each speaker's GMM is trained using these new feature vectors.

As well as in speaker model training, after the process of projecting again by (4) with test data $\hat{x}$, a log likelihood of each customer c is computed by GMM. An identification result is obtained as customer c with maximum log likelihood shown in Eq. (3.5)

$$c = \arg\max \log P(\hat{x} \mid \lambda^{(s)})\ . \qquad (3.5)$$

The proposed method is termed as "MFB-PCA".

3.3 Experiments

3.3.1 Conditions

The evaluation of the proposed method was conducted by text-independent closed-set speaker identification experiments using two databases.

One is large-scale Japanese speaker recognition evaluation corpus constructed by National Research Institute of Police Science (NRIPS), which contains 283 Japanese males from 18 to 76 years old, recorded at two time session over 3 months. Each speaker uttered 50 ATR phoneme balanced sentences and each utterance having a length of about 5 seconds on average was recorded twice at each session via 4 kinds of channel , a bone-conduction microphone, an air-conduction microphone, and two kinds of speeches referred over cell phone respectively. In this study, only the utterances recorded via an air-conduction microphone at first recording session is used. A full description of the corpus can be referenced at [74]. After down-sampling at 16 kHz, the first 5 utterances were defined for each speaker as the training set and remaining 45 utterances as the test set.

The other is NTT database made up of 2110 utterances by 23 male speakers at seven time session over 16 months. The length of each speech data is 6 seconds on average. Each utterance is recorded at 16 kHz sampling frequency with 16 bit per sample. In the speaker identification experiment, each five sentences uttered by 23 speakers at first time session are defined as the training set and each 15 sentences uttered at other time session as the test set for evaluation, that is, a total of 1,995 utterances is used.

Compared with the NTT database, the NRIPS has a large scale of speakers so that we can test the performance in a multi speaker environment by the proposed method, but it has only one time session. On the contrary, the NTT database with a much smaller scale of speakers has a multi time session so that the performance at different time session can be tested. Evaluation by using a database that meets the demand of both scale and multi time session is the following work for us.

For both of the two databases, each utterance is divided into 25ms frames with 10ms frame increment and parameterized into 12 cepstral coefficients obtained by 24-channel Mel-frequency filter-bank (MFB) analysis. The first speaker identification experiments were carried out by two kinds of method: the first method is a conventional method based on GMM using 12 dimensional MFCC parameters. The second is the proposed method based on GMM using the 12 dimensional phoneme-independent subspace vectors obtained from 24-channel MFB.

3.3.2 Experimental results

Table 3.1 shows the results of comparative speaker identification experiments between the proposed method (MFB-PCA) and the conventional method with MFCC on the two database using 16, 32 and 64 mixtures. The GMM technique is used for both of methods.

In Table 3.1, “MN” indicates number of mixtures. “MFCC” denotes the IER of using 12 dimensional MFCC feature vector extracted from 24-channel MFB. “MP” denotes the IER by the proposed method (MFB-PCA) based on the same 24-channel

MFB. In order to investigate phoneme-dependency of each eigenvector axis, four subspaces are compared with 1-12th, 2-13th, 3-14th and 4-15th eigenvectors respectively for NTT database first. The best IER 3.56% was achieved by MFB-PCA (3-14), Compared with the conventional method with MFCC (64 mixtures), the IER was reduced by 36%, meanwhile, after PCA dimensionality reduction, new feature vector sets such as MFB-PCA (2-13), MFB-PCA (3-14) and MFB-PCA (4-15) achieved some improvement compared with MFCC except for MFB-PCA (1-12). The reason of this phenomenon might be that MFB-PCA (1-12) had not suppressed any phonetic variability owing to the subspace being projected to the axis of the 1st lower eigenvector. Therefore, that phonetic variability is affirmed powerfully represented in the lower 1st and 2nd principal components. A detailed discussion of the 1st and 2nd eigenvectors is provided in the flowing subsection 3.3.

Table 3.1 Speaker identification error rate on 12 dim.(%)

	NTT				
MN	MFCC(1-12)	MP(1-12)	MP(2-13)	MP(3-14)	MP(4-15)
16	6.70	6.87	5.31	4.76	5.31
32	5.61	5.76	4.41	3.56	4.81
64	5.56	7.07	4.16	4.41	5.21

So the following evaluation for NRIPS database, comparative experiments are directly conducted between MFCC (1-12) and MP (3-14).

Table 3.2 Speaker identification error rate on 13 dim.(%)

	NRIPS	
MN	MFCC(1-12)	MP(1-12)
16	5.73	5.58
32	5.61	5.32
64	6.10	6.87

In table II, the NRIPS database presents a more realistic challenge to speaker identification for its scale of 283 speakers than the NTT database. The best IER was 5.32% by MFB-PCA (3-14) (32 mixtures). Compared with the conventional method with MFCC (32 mixtures), the IER was reduced by 5.17%.

In addition, on the basic of suppressing the first two eigenvectors, some more experiments are carried out by increasing the dimensionality of MFCC and MFB-PCA respectively. Table III, IV, and V show the experimental results with 13, 14, and 15, 16-dimensional feature vectors respectively. Although the performance of IER by MFCC became better and better with dimension increasing, our proposed method (MFB-PCA) still indicated some improvements than MFCC. Especially, in table V, achieved 2.61% IER that was the best performance and has reduced error by 21% compared with the conventional method with MFCC (32 mixtures) for NTT database. On the side of NRIPS database, the best IER was 4.52% by MFB-PCA (3-17) (32 mixtures). Compared with the conventional method with MFCC (32 mixtures), the IER was reduced by 15.8%.

Table 3.3 Speaker identification error rate on 14 dim.(%)

DB	NTT		NRIPS	
MN	MFCC(1-13)	MP(3-15)	MFCC(1-13)	MP(3-15)
16	5.86	4.21	5.59	5.37
32	4.11	3.21	5.19	5.03
64	4.06	4.41	5.71	6.37

Table 3.4 Speaker identification error rate on 15 dim.(%)

DB	NTT		NRIPS	
MN	MFCC(1-14)	MP(3-16)	MFCC(1-14)	MP(3-16)
16	5.51	4.31	5.58	4.84
32	3.96	3.11	5.36	4.75
64	3.51	3.36	6.12	6.39

Table 3.5 Speaker identification error rate on 16 dim.(%)

DB	NTT		NRIPS	
MN	MFCC(1-14)	MP(3-16)	MFCC(1-14)	MP(3-16)

16	5.51	4.31	5.58	4.55
32	3.96	3.11	5.37	4.52
64	3.51	3.36	5.89	6.06

Furthermore, a graph is made to show the trends of the best IER (highlighted in every table) achieved by MFB-PCA corresponding to the different ranges of projected axis.

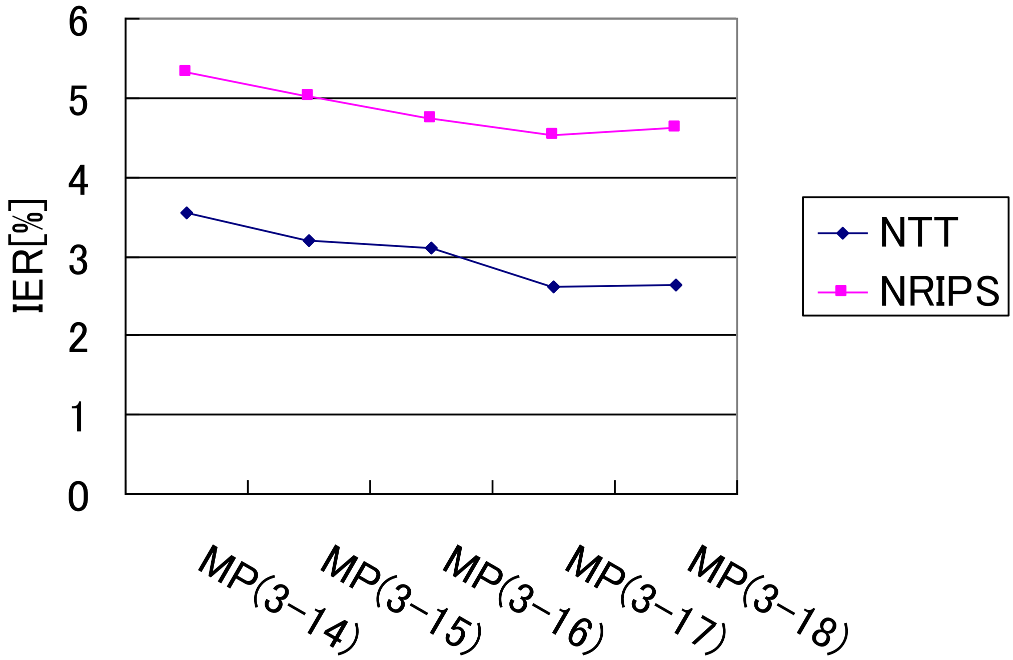


Figure 3.2 Best IER of 32 Mixtures

Obviously, for both of the two databases, when projecting over the 17th eigenvector, the performance of MFB-PCA begins to decrease to some degree. The reason resulting in the phenomenon is that besides phonetic variability and speaker information, there is no doubt some other thimbleful information like time difference and people condition in the very higher eigenvectors. When utterance data projected to them, they are also could be factors that affect the performance of the IER.

As a result, high identification performance can be obtained by the proposed method (MFB-PCA) by suppressing phonetic variability, under the assumption that a subspace constructed with some principal components having larger contribution rate is considered as "phoneme-dependent space" and a subspace constructed without these components is considered as "phoneme-independent space" in this chapter.

3.3.3 Discussion

In this section, the phonetic variability is discussed in the lower eigenvectors when a great deal of utterance was prepared for PCA. To confirm the phoneme-dependent space constructed by the first two eigenvectors that contain a large part of phonetic variability, the projection distance is investigated between two vowels respectively in phoneme-dependent and phoneme-independent subspace. Figure.3.2 shows the image of projection distance.

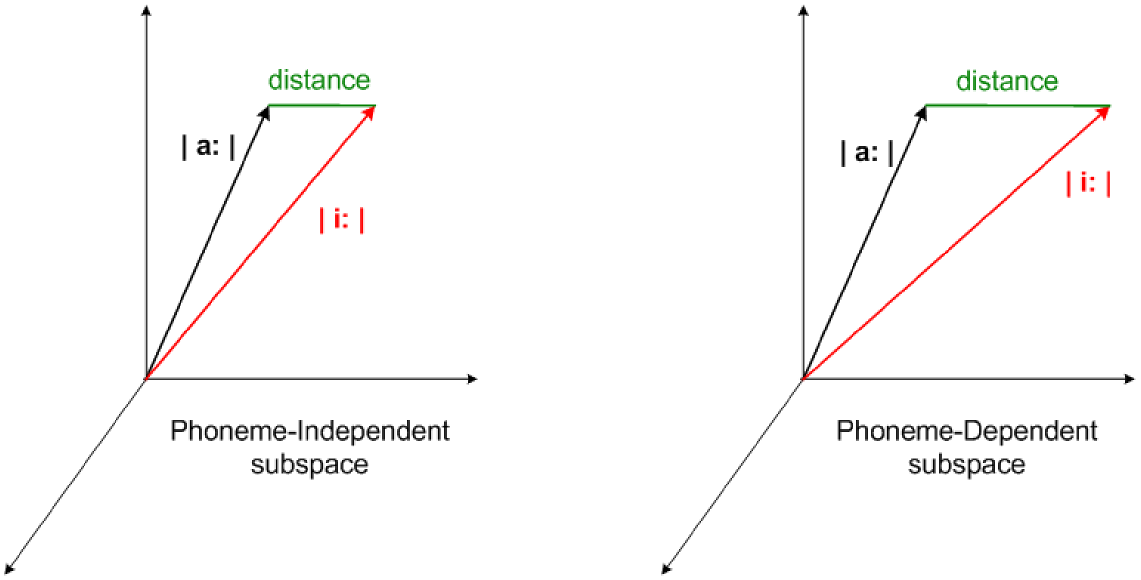


Figure 3.3 the image of the distance between projection of two vowels

In this subsection, Japanese long-vowels |*a:*| and |*i:*| are projected into a subspace constructed with some eigenvectors $\varphi_i (i = 1, \cdots, k \cdots, s \cdots, n)$. The distance between two projections of vowel can be shown as follows:

$$dist = \sum_{i=k}^{s} | \bar{x} \cdot \varphi_i - \bar{y} \cdot \varphi_i |, \tag{3.6}$$

where $\bar{x}, \bar{y}$ are the mean vector of a set of 24-dimensional MFB feature vectors of a: and i: respectively. k and s denotes the range of eigenvectors projected. Because the phonetic variability of these two vowels is absolutely different, the optimal outcome is that the projection distance in the phoneme-independent subspace is much shorter than it in the phone-dependent subspace.

In advance, a series of utterance data of Japanese long vowels are prepared for projection test. They were recorded by a person from 2003 to 2004, once a week, in the morning, afternoon and night, over 16 months. Each long vowel has 204 utterances respectively. Figure 3.3 and 3.4 shows the projection distance by projecting to the lower eigenvectors obtained in subsection 3.3.2.

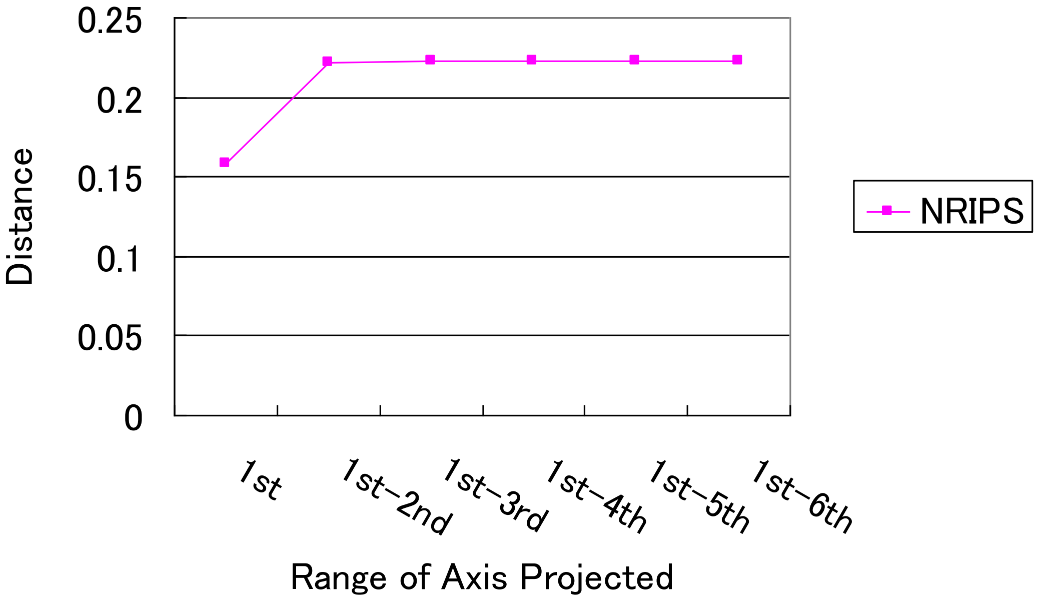


Figure 3.4 Projection distance in phoneme-dependent subspace for NIRPS database

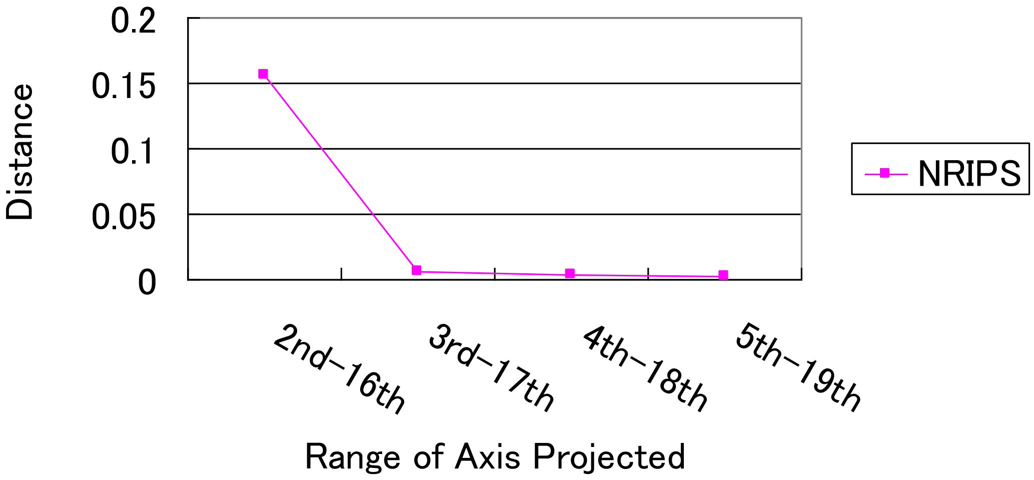


Figure 3.5 Projection distance in phoneme-independent subspace with 15 DIM for NIRPS database

From the Figure 3.4, it is obvious that until projecting to the 3rd eigenvector, the projection distance rises sharply, compared with almost no changes by projecting to the 3rd eigenvectors or later ones. Therefore, it is proved that the very first two lower eigenvectors suppress the phonetic variability when a great deal of utterance data was prepared for PCA. It is also clear in Figure 5 that we can find the projection distance decreases steep until by projecting to the 3rd eigenvectors or the later ones in such a phoneme-independent subspace. This is consistent with the explanations for the changes of projection distance in phoneme-independent subspace.

Then, let us continue to focus on the Figure 3.3, Figure 3.4 to compare the distance which is corresponding to the range of eigenvectors projected after the 3rd. The projection distance in phoneme-independent subspace is much closer than it in phoneme-dependent subspace. The projection for NTT database also shows the characteristic of the distance changes similarly. The MFB feature transformation to phoneme-independent subspace has suppressed a large part of phonetic variability so that the IER can be improved.

In general, the speech data roughly consists of three main parts, i.e. phonetic variability, speaker information, time difference, and others. In this paper, phonetic variability is discussed. From the results of the experiments, TI speaker identification performance is confirmed to be improved by discarding some principal components. It shows these components are strongly affected by phonetic variability. On the other hand, the dimensionality increasing reduced the error. Although it should be studied where the upper limit achievable for higher dimension PCA eigenvector is, speaker information and other information are intermingled in them. Hence, [74] developed new speech database in which time difference has the largest dispersion.

3.4 Summary

This chapter has introduced a new feature vector extraction method using PCA for text-independent speaker identification. In the method, it is assumed that a subspace constructed with some principal components having larger contribution rates is considered as "phoneme-dependent subspace" and a subspace constructed without these components is considered as "phoneme-independent subspace". GMM-based TI speaker identification experiments are conducted using the proposed phoneme-independent feature vector (MFB-PCA) and the conventional MFCC and show how a standard speaker identification system can be significantly improved. The results for two databases are also better than the conventional method. Therefore, a robust speaker model can be constructed by the new feature vector having less phonetic variability in a phoneme-independent subspace and get a better TI speaker identification performance.

Chapter 4

Inter-session variability reduction for MFCCs

4.1 Introduction

A speaker identification system consists of a feature extraction frontend and a classifier. For Text-Independent (TI) speaker identification systems, a popular choice for the features is Mel-Frequency Cepstral Coefficients (MFCCs) [75] and the classifier is the Gaussian Mixture Model (GMM) [76]. MFCCs are extracted from a speech waveform by first obtaining Mel filter bank spectrum (MFB) and then applying Discrete Cosine Transforms (DCT). GMM is a statistical model that can model complex data distribution using multiple Gaussian distributions and their mixing weights.

The performance of speaker identification systems largely depends on features. To achieve higher performance, the frontend should extract features so that speaker information is emphasized while phonetic and inter-session variability are suppressed. While the MFCC features are very popular for speaker identification, they do not equip a mechanism to suppress the phonetic variability because they are originally developed for speech recognition, where the extraction of phone information is important. Therefore, the GMM has to manage phonetic variability all by itself. In order to suppress the phonetic variability helping the GMM to identify speakers regardless of what is said, a feature extraction method that utilizes Principal Component Analysis (PCA) has been previously proposed, and has shown that it is effective to improve speaker identification performance [77, 78].

Inter-session variability refers to variability of speech characteristics that arise for speech sounds recorded in different sessions over a certain time-span. It is known that even in the same recording environment, such as the same microphone and the same room reverberation, the characteristics of recorded speech sounds vary. This variability is due to the fact that characteristics of our voice itself drift over time. Though it is rare that the change is noticed, it largely impacts the performance of automatic speaker identification systems.

Although the problem of the session variability is widely recognized, its mechanism has not been well investigated, and there are only few researches to normalize the effect [79]. However, without addressing the problem, it is not possible to provide a practical speaker identification system as a useful biometrics application since the performance degrades as time passes after the registration of users' voice.

In this chapter, previous PCA based phonetic variability suppression method [78] is adopt and extend to suppress the session variability [72]. The basic idea is to transform observed MFB spectrum to another space, where the session variability and others including speaker information are separated into different subspaces. Then, by discarding the session variability subspace and applying inverse transformation, normalized spectrum is obtained. The question is how to obtain such transformation. For this purpose, speech data that specifically contains session variability is prepared by controlling other factors. PCA is applied to it assuming that the session variability subspace is obtained as a primary subspace by the PCA analysis.

The formulation of our proposed method is similar to the Nuisance Attribute Projection (NAP) method in which a transformation is applied to features to suppress channel differences in SVM expansion space [65]. The NAP has been developed to a popular intersession variability compensation method [80], which estimates and removes the channel information existed in speaker's features from the super-vectors before SVM training. Consider the channel dependent portion is much lower dimension than the speaker dependent portion, so PCA can be employed to estimate the first n largest components from the super-vectors. But NAP is designed only for SVM and the super-vector is computed from performing a MAP adaption to GMM-UBM training with the speech data. Actually, our method can be regarded as a special case of the NAP method, so the PCA is also applied to the controlled recordings of a special database [74] to using the speech features instead of super-vectors. The new contribution of our paper is to apply the framework for the time difference problem that is originated from speaker’s pronunciation itself and the way of estimating the parameters of the transformation is also a new point of this paper.

The remainder of this chapter is organized as follows. In Section 4.2, the proposed session variety suppression method is explained. Experimental setups are described in Section 4.3 and the results are shown in Section 4.4. Finally, summary is given in Section 4.5.

4.2 Nuisance Attribute Projection

Solomonoff, et al proposed nuisance attribute projection (NAP) [56] as a technique to suppress session variations.

NAP constructs a session matrix subspace with low-dimensional session variations observed in a training corpus of speech. NAP uses projection method in the SVM kernel space to project data onto a subspace in which less prone to variations while inter-session variability modelling removes variations in the GMM space via a set of mean offsets. It means to project out the unwanted dimensions of within-class variation.

A sequence of n nuisance directions, defined by the low-rank transformation matrix U, can be reduced from the input data x using the projection.

$$P_n x = (I - U_n U_n^T)x, \tag{4.1}$$

I is the identity matrix in (5.1). The directions are found by the criterion maximization

$$J(u) = u^T S_\omega u, \tag{4.2}$$

S_ω is the training data from within-class scatter matrix. This is same as determining the eigenvectors corresponding to the largest eigenvalues.

$$S_\omega u = \lambda u. \tag{4.3}$$

In practice, PCA is used to solve the problem. With the help of PCA, the eigenvalues and eigenvectors corresponding to the low-dimensional correlation matrix of S_ω are separated from the high-dimensional S_ω through the decomposition. The within-class scatter matrix can be got from

$$S_\omega = (diag(W1) - W)K, \tag{4.4}$$

where K is the matrix, 1 is a column vector, W is the weight matrix that means observations in the same speaker's training data set.

The linear SVM kernel with the integration of NAP compensation can finally be indicated as

$$K(x_{i,}x_{j,}) = P_n x_i \cdot P_n x_j, \tag{4.5}$$

where P_n is the projected out of the input data $x_i$and x_j.

NAP is regularly employed throughout this dissertation because it simplistic application and famous capability to suppress session variability.

4.3 Inter-session variability suppression for GMM based speaker identification

A. General Structure

The overview of our proposed method is shown in Fig.1. The system diagram of our proposed TI speaker identification system consists of a training phase and an identification phase. In the training phase, inter-session variability subspace is first obtained by applying PCA to MFB spectrum vectors of a specific vowel from a single speaker. It is assumed that inter-session variability is mainly located in the first principal component. In other words, it is assumed that it represents the direction of inter-session variability. Given the subspace of inter-session variability, original MFB spectrum vectors from all the training and test set speakers are normalized by removing that component. Using the normalized MFB spectrum, MFCC features are estimated and GMMs are trained and evaluated for speaker identification.

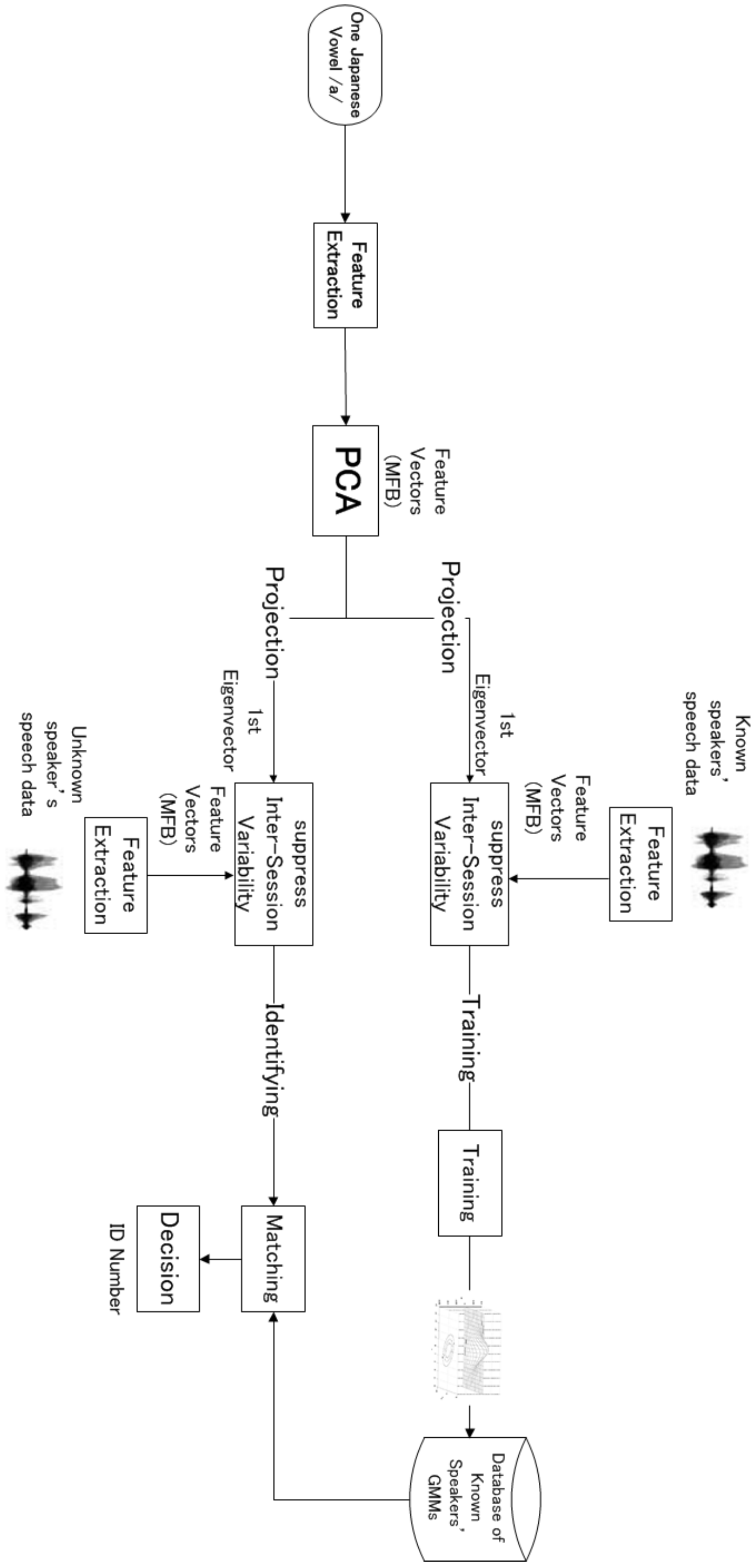


Figure 4.1 Overview of the proposed speaker identification system

B. PCA based Inter-session Variability Suppression

Let $\boldsymbol{x} = \{x_1, x_{2,} \cdots, x_n\}^T$ denote n-dimensional MFB spectrum vectors of a specific vowel of a designated speaker. A mean vector can be computed by

$$\bar{x} = \frac{1}{N}\sum_{t=1}^{N} x_t \ , \tag{4.6}$$

where *L* is the number of samples. Similarly, a covariance matrix is obtained by

$$R_X = \frac{1}{L}\sum_{t=1}^{N}(x_t - \bar{x})(x_t - \bar{x})^T \ , \tag{4.7}$$

The covariance matrix R_x (n by n) can be decomposed to eigenvectors and eigenvalues as follows:

$$R_X = \Phi\Lambda\Phi^T \ . \tag{4.8}$$

where Λ is a diagonal matrix whose diagonal components are eigenvalues $\lambda_i (i = 1, \cdots, k, \cdots, n)$. Φ is a matrix whose columns are eigenvectors $\varphi_i (i = 1, \cdots, k \cdots, n)$ corresponding to the eigenvalues.

In this study, it is assumed that the first eigenvector (i.e. φ_1) spans the subspace that has the largest inter-session variability. This means the most inter-session variability can be removed by projecting the MFB spectrum into the ortho-complementary space of the first eigenvector and then transforming it back to the original space. Or equivalently, the inter-session variability component can be first obtained by projecting the MFB spectrum to the inter-session variability subspace and then subtracting it in the original feature space as shown in Equation (4.9).

$$x' = x - (x \cdot \varphi_1{}^T)\varphi_1 \,. \tag{4.9}$$

C. GMM training and likelihood evaluation

The inter-session variability subspace obtained from the designated speaker is used in common for the inter-session variability normalization for all the training and test set speakers. Then, MFCC features are obtained from the normalized MFB spectrum $\boldsymbol{x}'$ by applying DCT as shown in Equation (4.10).

$$\hat{\boldsymbol{x}}'' = DCT(x') \,. \tag{4.10}$$

For speaker identification, GMM log likelihood for speaker s is obtained for the MFCC features $\hat{\boldsymbol{x}}''$ as $\log P(\hat{\boldsymbol{x}}''|\lambda^{(s)})$, where λ is the GMM parameters. An identification result c is obtained by computing the maximum of the log likelihood as shown in Eq. (4.11)

$$c = \arg\max_s \log P(\hat{\boldsymbol{x}}''|\lambda^{(s)}) \,. \tag{4.11}$$

4.4 Experiments

4.4.1 Conditions

Two databases were used in the experiments. One was used to obtain the session-variability subspace and the other was used as training set of speaker GMMs and test set. For the former, “Specific Speakers’ Speech Corpus over Long and Short

Time Periods" [81] is used including five Japanese vowels /a/, /i/, /u/, /e/, and /o/, uttered by a person once a week for 10 months. For the latter, a subset of the NTT speaker recognition database [3] consisting of 780 phoneme-balanced utterances from 23 male speakers is used. Among the 23 speakers, one had three sessions, another had had 6 sessions, and the ohters had 7 sessions. These sessions are recorded separately during 16 months. The duration of each utterance is 6 seconds on average. The waveforms are recorded at16kHz sampling frequency and 16bit quantization. Five utterances per a speaker were used for his speaker model training. Similarly, five utterances per a speaker from the rest of the sessions are used for the evaluation [82].

4.4.2 Experiment results

The results with various vowels are shown in the following tables where the MN means mixture numbers.

Table 4.1 Baseline speaker identification error rate (IER) using MFCCs without inter-session variability normalization. Rows are number of GMM components of a speaker model. Columns are a dimension of MFCC feature vectors that are obtained from 24 channel Mel-Filter Bank output.

MN	12dim	13dim	14dim	15dim	16dim
16	3.31	3.61	3.01	2.56	1.65
32	1.95	1.95	1.80	1.65	1.20
64	1.95	2.11	1.20	2.11	<u>1.19</u>
96	4.06	2.71	1.35	1.50	1.35
128	3.31	3.46	2.11	1.80	2.26

Table 4.2 IER (%) using MFCCs with proposed PCA based session variability suppression. Session variability subspace is obtained using /a/ sound.

MN	12dim	13dim	14dim	15dim	16dim
16	4.36	3.91	2.56	2.56	2.11

32	2.11	1.50	1.35	0.75	0.90
64	1.50	1.65	1.20	1.20	1.05
96	2.26	1.65	2.26	2.26	1.65
128	2.86	2.26	3.01	3.31	1.95

Table 4.3 IER (%) using MFCCs with proposed PCA based session variability suppression. Session variability subspace is obtained using /i/ sound.

MN	12dim	13dim	14dim	15dim	16dim
16	4.66	4.82	3.91	3.31	3.31
32	2.11	2.11	1.65	1.95	1.65
64	3.01	2.71	1.05	1.50	1.65
96	3.01	2.56	1.95	2.11	1.50
128	4.66	4.21	3.76	3.01	2.41

Table 4.4 IER (%) using MFCCs with proposed PCA based session variability suppression. Session variability subspace is obtained using /u/ sound.

MN	12dim	13dim	14dim	15dim	16dim
16	4.66	4.82	3.91	3.31	3.31
32	2.11	2.11	1.65	1.95	1.65
64	3.01	2.71	1.05	1.50	1.65
96	3.01	2.56	1.95	2.11	1.50
128	4.66	4.21	3.76	3.01	2.41

Table 4.5 IER (%) using MFCCs with proposed PCA based session variability suppression. Session variability subspace is obtained using /e/ sound.

MN	12dim	13dim	14dim	15dim	16dim
16	4.06	4.66	3.91	3.31	2.56
32	2.86	2.71	3.16	2.41	1.50
64	3.31	2.86	2.11	2.11	1.35
96	3.16	1.95	2.11	1.80	2.41

128	3.31	1.95	2.26	2.71	1.65

Table 4.6 IER (%) using MFCCs with proposed PCA based session variability suppression. Session variability subspace is obtained using /o/ sound.

MN	12dim	13dim	14dim	15dim	16dim
16	3.76	4.51	4.21	2.26	2.41
32	2.11	1.50	1.65	1.35	1.05
64	3.61	2.86	1.65	0.90	1.20
96	3.01	2.41	1.50	2.26	1.80
128	3.31	3.01	1.80	1.65	1.95

Table 4.7 IER (%) using MFCCs with proposed PCA based session variability suppression. Session variability subspace is obtained using /N/ sound.

MN	12dim	13dim	14dim	15dim	16dim
16	5.41	4.21	3.01	3.31	2.86
32	2.86	1.95	2.11	1.65	1.05
64	3.01	2.11	1.80	2.11	1.35
96	3.91	2.11	2.26	2.11	1.80
128	3.01	1.95	3.01	2.41	1.65

4.4.3 Discussion

Table 4.1 shows baseline speaker identification error rate (IER) that are obtained by using the MFCC features without the inter-session variability normalization. IER were evaluated using speaker GMMs with varying number of component Gaussians and MFCC features with varying dimensions. These MFCC features are sub-vectors of MFCCs that were obtained from 24 channel Mel-filter bank output. The lowest IER 1.19 is obtained when the MFCC dimension was 16 and the speaker GMM had 64 Gaussian components.

Tables 4.2 to 4.7 show IER using the MFCC features made from the session-variability normalized MFB spectrum using different vowels to estimate the session-variability subspace. The lowest IER based on vowels /a/, /i/, /u/, /e/ and /o/ for the subspace-estimation were 0.75, 1.05, 1.35, 0.90 and 1.05, respectively. The lowest IER of 0.75 was obtained when vowel /a/ was used with a GMM having 32 mixture components and the features having 15 dimensions. The relative IER reduction from the baseline result was 37.0%.

4.5 Summary

A new feature vector extraction method is proposed using PCA for text-independent speaker identification that suppresses inter-session variability. According to the experiment results, by applying PCA to MFB spectrum vectors of a specific vowel from a single speaker recorded over several sessions, session variability sub-space is obtained as the primary eigenvector. Given the eigenvector representing the direction of session variability, the session variability component is obtained as an inner product of the spectrum vector and the eigenvector. By subtracting the obtained session variability term from the original vector, the session variability is normalized. GMM based text-independent speaker identification experiments showed that the proposed session-variability normalization method is effective. Compared to a MFCC based baseline, 37.0% relative reduction of IER was obtained. Future work includes improving the estimation of the session variability subspace.

Chapter 5

Phoneme dependent inter-session variability reduction for speaker verification

5.1 Introduction

The current trend in Text-Independent speaker verification is using GMM-UBM super-vectors to model the speaker features [47]. However, variation between training and test utterances recorded over months can strongly affect the performance. The variation referred to as inter-session variability has become one of the most challenging problems facing speaker verification researchers today. Due to the fact that characteristics of voices change over time, the characteristics of recorded speech sounds vary even if they are recorded in the same environment, such as using the same microphone and a room with the same reverberation. State-of-the-art session variability compensation methods have been proposed to reduce the confusing variability that is generally caused by GMM-UBM super-vectors. Eigenvoice [62], Eigenchannel [63] and joint factor analysis (JFA) [64] are based on separating the lower dimensional speaker-dependent subspace and the channel-dependent subspace in the super-vector frame-work. Nuisance attribute projection (NAP) [65] applies a linear transformation to the GMM-UBM super-vectors to project out nuisance directions. Inspired by JFA, speaker and channel variability are jointly modeled to obtain i-vectors [66] using factor analysis, and then LDA and WCCN are used to reduce dimensionality, while retaining the speaker identity information in the GMM-UBM super-vectors domain. Meanwhile, the enrolment utterances are required to be long enough to contain as many phonemes as possible [67].

Nevertheless, they do not take into account phoneme deficiency. What if just a small amount of training utterances were available? GMM-UBM super-vectors will potentially lead to worse modeling because some phonemes are missing in the enrolment utterances. Several phoneme-based works have tried to constrain the effect, e.g. Margin-Chagnoleau et al. [68] have shown that speaker identification performance depends on the phonetic label of the speech segments used. Matsui and Furui [69] used phoneme-specific HMMs to more accurately model the target speakers. Mohamed Abdel et al. [70] described the phoneme selection of a speaker utterance before recognition much improved the speaker verification accuracy. Gutman and Bistritz [71] proposed a two-stage phoneme adaption method using the TIMIT and NTIMIT database.

However, they did not directly consider the other side that the original acoustic GMM mean vectors contained the inter-session variability. In this study, an effective phoneme dependent method is developed to reduce the inter-session variability. Via speech segmentation using speech recognition technology, a speaker’s model can be represented by several various phoneme Gaussian mixture models. Each of them covers an individual phoneme. Then, the inter-session variability subspace for each individual phoneme can be obtained as a primary subspace constructed with first several principal components. A reduction method similar to NAP, called nuisance attribute subtraction (NAS) was employed to reduce the inter-session variability through using a long-term recorded corpus uttered by a single speaker. That is, a subspace constructed with the some principal components is considered as ‘inter-session variability space’ and a subspace constructed without these components is considered as ‘inter-session-independent space’ in this paper.

The remainder of this chapter is organized as follows. Section 5.2 provides a description of phoneme-based inter-session reduction for speaker verification. In Section 5.3, comparative experiments and discussions are carried out to show the effectiveness of our proposed method. Finally, Section 5.4 gives the summary of this chapter.

5.2 Baseline GMM-SVM speaker verification system

State-of-the-art GMM-SVM speaker verification systems provide a performance reference point for the further development of speaker verification systems and associated techniques. Accordingly, research efforts through this dissertation are compared to a baseline GMM-SVM system in order to evaluate the potential performance gains achieved by the proposed methods.

The GMM-UBM configuration described in Section 2.3.2.1 represents state-of-the-art technology and forms the fundamental baseline configuration used in this work. This system is also used to produce the GMM-UBM mean super-vectors for the baseline support vector machine (SVM) classifier later described in Section 5.4.

The feature extraction process is depicted in section 2.2 and benefits from12-dimensional feature-warped MFCCs with appended delta coefficients. GMM training utilizes mean-only MAP adaptation. Unless otherwise stated, an adaptation relevance factor of τ=10 and 256-mixture models are used throughout this work. Speaker adaptation takes place after a single E-M MAP adaptation iteration. Gender-dependent UBMs were trained using a diverse selection of 1675 utterances from National Research Institute of Police Science (NRIPS) speech database. GMM-SVM based classification scores are calculated by T-norm score normalization.

5.3 Phoneme-based inter-session reduction for speaker verification

Figure 5.1 shows a block diagram of our proposed phoneme-based inter-session reduction for speaker verification. The process consists of the following modules:

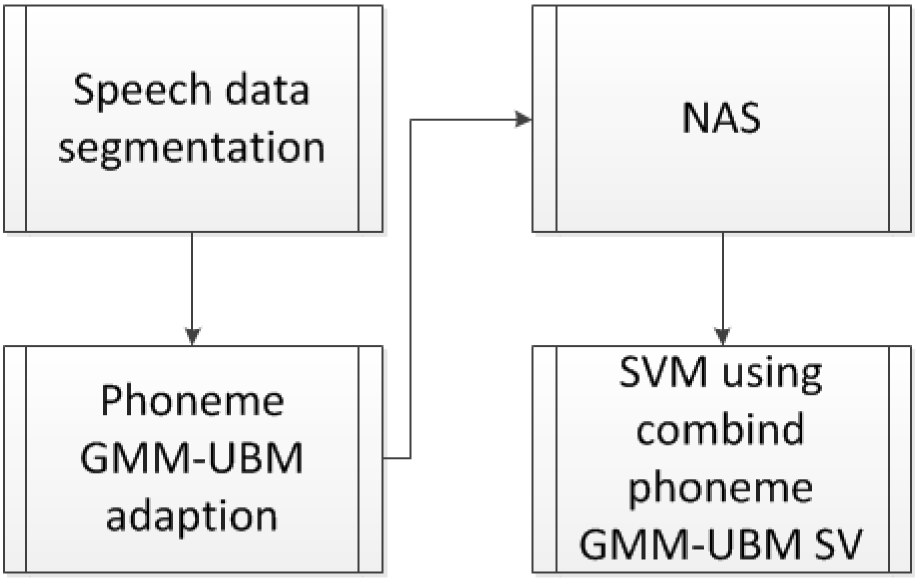


Figure 5.1 Block diagram of our proposed Phoneme-based inter-session reduction for speaker verification

5.3.1 Segmentation

Segmentation of the speech was carried out by 3-state monophonic HMM-based speech recognition. Since the sentences for training are known, segmentation was relatively easy to perform by doing forced alignment. Then, phoneme segmentation for each testing utterance can be processed with the previously constructed HMM phoneme models.

5.3.2 Phoneme-based GMM-UBM adaption

For a given phoneme, a phoneme GMM can be generated by adapting a well-trained phoneme UBM using the maximum a posteriori (MAP) adaptation approach. Supposing a *D*-dimensional phoneme feature vector, $\boldsymbol{x}_p$, the probability density for a phoneme model is defined as the weighted sum of *M* Gaussian densities:

$$p(\boldsymbol{x}_p|\lambda) = \sum_{i=1}^{M} \omega_i g(\boldsymbol{x}_p|\boldsymbol{\mu}_i, \Sigma_i)\,, \qquad (5.1)$$

where $g(\boldsymbol{x}_p|\boldsymbol{\mu}_i, \Sigma_i)$ is the Gaussian probability density function, which is parameterized by mean vector $\boldsymbol{\mu}_i$, covariance matrix Σ_i, and mixture weights ω_i that add to unity. These parameters are collectively represented by the notation: $\lambda = \{\omega_i, \boldsymbol{\mu}_i, \Sigma_i\}$ $i = 1, \dots, M.$ Each component density is a D-variate Gaussian function of the form:

$$\mathrm{g}(\boldsymbol{x}_p|\boldsymbol{\mu}_i, \Sigma_i) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(x_p-\mu_i)^T \Sigma_i^{-1}(\boldsymbol{x}_p - \boldsymbol{\mu}_i)\right\}. \quad (5.2)$$

5.3.3 Inter-session Reduction Procedure

The section focuses on reducing the inter-session variability. Previous PCA-based phonetic variability suppression method [72] is adopted and extended to reduce the inter-session variability. The basic idea is to transform input vectors to another subspace, where the inter-session variability and other factors including speaker information were separated into different subspaces. Then, by discarding the inter-session variability subspace and applying inverse reduction, normalized feature vectors were obtained. The question is how to obtain such reduction. For this purpose, speech database that deliberately contained inter-session variability is prepared by controlling other factors.

The next two subsection introduce how to apply the inter-session reduction, which similar to NAP, in the GMM mean vectors domain and GMM-UBM super-vectors domain respectively

A. Nuisance Attribute Subtraction in GMM Mean Vectors Domain (NAS-m)

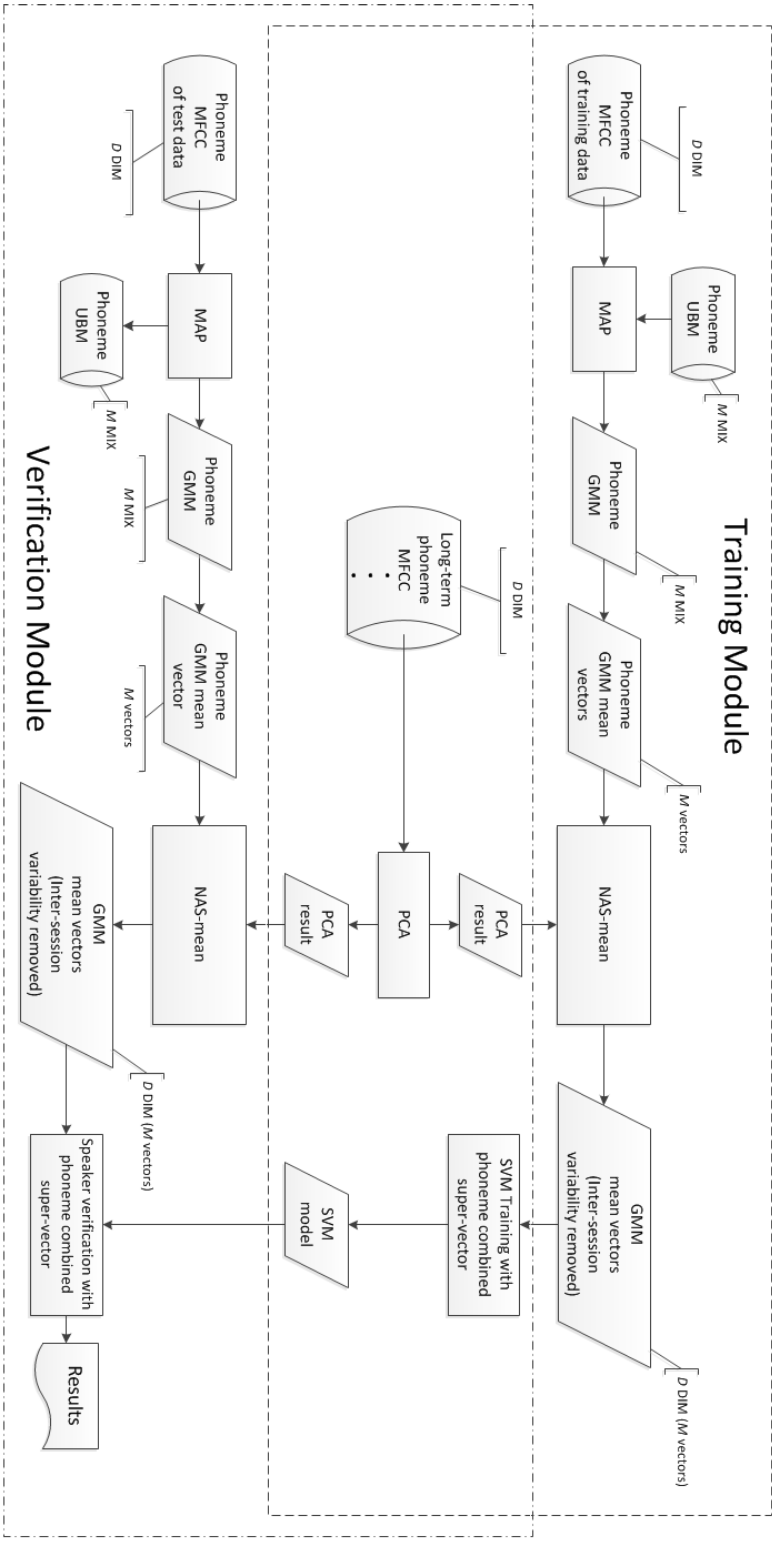


Figure 5.2 NAS in GMM mean vectors domain for a phoneme

Figure 5.2 shows a block diagram of phoneme-dependent using nuisance attribute subtraction in the GMM mean vector domain, when the long-term speech data uttered by a single speaker is segmented into phoneme segments, and then, the sequence data of a phoneme $\{x_t\}(t = 1, 2, \ldots, N)$ was observed in a D-dimensional observation space. The mean vector $\bar{\mathrm{x}}$ and covariance matrix $\mathrm{M_g}$ were computed from the input data as $\bar{x} = \frac{1}{N}\sum_{t=1}^{N} x_t$ and $M_g = \frac{1}{N}\sum_{t=1}^{N}(x_t - \bar{x})(x_t - \bar{x})^T$, respectively. M_g could be decomposed to eigenvectors and eigenvalues as follows:

$$M_g = U_g \Lambda_g U_g^T , \qquad (5.3)$$

where the columns of U_g contain the eigenvectors φ of M_g, and Λ_g contains the corresponding eigenvalues λ in its diagonal. In this study, it is assumed that the first q eigenvectors (i.e.$\varphi_1, \ldots, \varphi_q$) span the subspace that had the largest inter-session variability, which means that the inter-session variability component can be first obtained by projecting each phoneme GMM mean vector m_g onto the inter-session variability subspace and then subtracting it in the original space, namely applying nuisance attribute subtraction (NAS) on the GMM mean vector as shown in Equation (5.4).

$$\boldsymbol{p}_g = \left(I - U_{[q]g} U_{[q]g}^T\right)\boldsymbol{m}_g , \qquad (5.4)$$

where $U_{[q]g}$ means a subspace of U_g, which is constructed with the first q components. Then, the new GMM mean vector $\boldsymbol{p}_\mathrm{g}$ could concatenate an inter-session variability reduced phoneme GMM-UBM super-vector. Thus, a large dimensional GMM-UBM super-vector including the entire selected phoneme could be combined with equal weights for continued SVM-based speaker verification.

B. Nuisance Attribute Subtraction in Super-Vector Domain (NAS-sv)

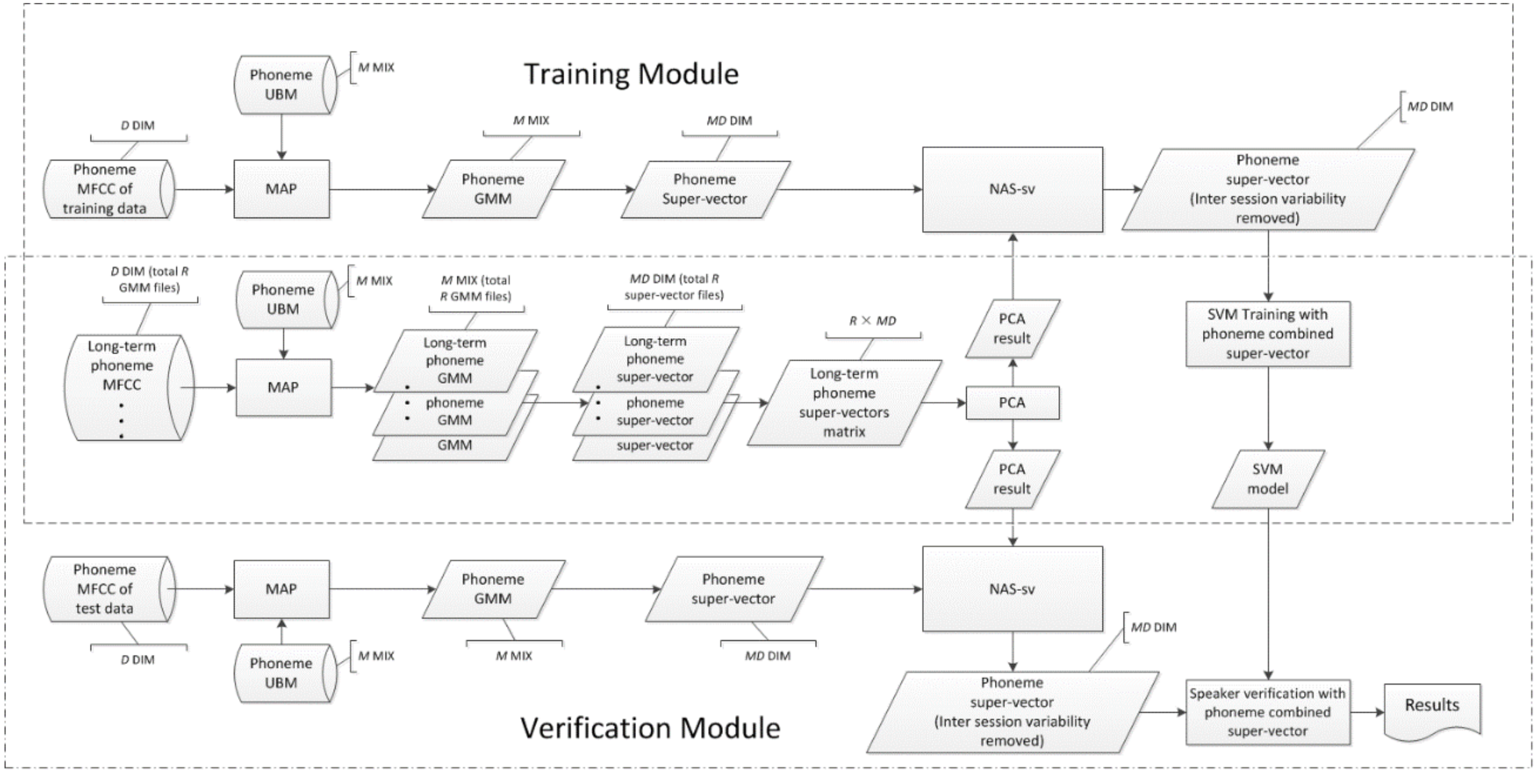

Training Module
Phoneme MFCC of training data
D DIM
Phoneme UBM
M MIX
MAP
Phoneme GMM
M MIX
Phoneme Super-vector
MD DIM
NAS-sv
Phoneme super-vector (Inter session variability removed)
MD DIM
D DIM (total R GMM files)
Long-term phoneme MFCC
Phoneme UBM
M MIX
MAP
M MIX (total R GMM files)
Long-term phoneme GMM
phoneme GMM
GMM
MD DIM (total R super-vector files)
Long-term phoneme super-vector
phoneme super-vector
super-vector
R × MD
Long-term phoneme super-vectors matrix
PCA result
PCA
PCA result
SVM Training with phoneme combined super-vector
SVM model
Phoneme MFCC of test data
D DIM
MAP
Phoneme UBM
M MIX
Phoneme GMM
M MIX
Phoneme super-vector
MD DIM
NAS-sv
MD DIM
Phoneme super-vector (Inter session variability removed)
Speaker verification with phoneme combined super-vector
Results
Verification Module

Figure 5.3 NAS in super-vectors domain for a phoneme

Figure 5.3 shows a block of phoneme-dependent inter-session variability reduction for speaker verification using nuisance attribute subtraction in the super-vector domain. When segmentation of the long-term speech was completed, every *M*-mixture phoneme GMM generated from one short utterance could be concatenated into a *MD*-dimensional phoneme GMM-UBM super-vector. Then, for *R* short utterances, there is a sequence of the phoneme GMM-UBM super-vector $\{\boldsymbol{s}_j\}(j = 1, 2, \dots, R)$ observed in a *MD*-dimensional observation space. The mean vector $\bar{\boldsymbol{s}}$ and covariance matrix M_{sv} were computed from the input data as $\bar{\boldsymbol{s}} = \frac{1}{R}\sum_{t=1}^{R} \boldsymbol{s}_j$ and $M_{sv} = \frac{1}{R}\sum_{t=1}^{R}(\boldsymbol{s}_j - \bar{\boldsymbol{s}})(\boldsymbol{s}_j - \bar{\boldsymbol{s}})^T$, respectively. M_{sv} could be decomposed to eigenvectors and eigenvalues as follows:

$$M_{sv} = U_{sv}\Lambda_{sv}U_{sv}^T \ . \qquad (5.5)$$

the inter-session variability subtraction applied to the GMM-UBM super-vector can be given by Equation(5.6).

$$\boldsymbol{p}_{sv} = (I - U_{[k]sv}U_{[k]sv}^T)\, \boldsymbol{m}_{sv} \, , \qquad (5.6)$$

where m_{sv} is the phoneme GMM-UBM super-vector and $U_{[k]sv}$ is a subspace of U_{sv} constructed with the first k components. Thus, a large dimensional GMM-UBM super-vector including the entire selected phoneme could be combined with equal weights for SVM-based speaker verification.

In this paper, these selected phonemes refer to vowels and nasals in Japanese, i.e., /a/, /i/, /u/, /e/, /o/, /m/, /n/ and /s/&/sh/ [70].

5.4 SVM-based Speaker Verification using phoneme GMM-UBM super-vectors

Support vector machines (SVMs) have proven to be a novel effective method for speaker verification because of their inherent property of discriminative training, and the ease of combining feature vectors.

The SVM two-class classifier was constructed using weighted sum of a kernel function $K(\cdot,\cdot)$ as follows:

$$f(\boldsymbol{x}_{ps}) = \sum_{i=1}^{N} \propto_i t_i K(\boldsymbol{x}_{ps,}\boldsymbol{v}_i) + d\,, \quad (5.7)$$

where $\boldsymbol{x}_{ps}$ are the GMM-UBM super-vectors of phoneme combined with equal weights; N is the number of support vectors, $\boldsymbol{v}_i$ are the support vectors obtained via an optimization process, and t_i are ideal outputs, with values of ±1 depending on whether the accompanying support vectors belonged to class 0 or 1. Overall, the parameters were subject to the constraint:$\sum_{i=1}^{N} \alpha_i t_i = 0$.

In this study, the linear kernel was used, which is given by Equation 5.8.

$$K(\boldsymbol{x}_{ps}, \boldsymbol{v}_i) = \boldsymbol{x}_{ps} \cdot \boldsymbol{v}_i\,, \qquad (5.8)$$

For classification, a class decision was based upon whether the value $f(\boldsymbol{x}_{ps})$, was above or below a threshold.

5.5 T-norm

The SVM score distribution from verification trials, however, can also be affected by the conditions exhibited by the test utterance. T-norm shares similarities with model-based cohort normalization while additionally incorporating score variance to better model the impostor cohort score distribution to address this issue. It is commonly applied to SVM-based classification scores as was seen in the recent NIST SREs to provide improved robustness to statistical errors encountered during the speaker modeling process [73].

T-norm is utilized to enhance the performance of GMM-SVM systems in this study. The normalization parameters were estimated using scores derived at test time from a set of imposter speaker models. A fixed set of imposter speaker models were scored in parallel with the target speaker model. The mean μ_{tnorm} and standard deviation σ_{tnorm} of the imposter scores were then used to adjust the target speaker score as

$$S_{tgt|tnorm}(O) = \frac{S_{tgt}(O) - \mu_{tnorm}}{\sigma_{tnorm}}\,, \qquad (5.9)$$

where $S_{tgt}(O)$ is the target speaker score for observations O.

5.6 Experiments

5.6.1 Conditions

The evaluation experiments were processed based on a SVM speaker verification system. It was conducted by text-independent open-set speaker verification experiments with a large Japanese speaker verification evaluation corpus separated into four cross-validations. The corpus was developed by the National Research Institute of Police Science (NRIPS). It contains speech data of 335 Japanese males aged 18 to 76 years, and was recorded in two time sessions over three months. Each ATR phoneme-balanced sentence has a length of about 2.5 seconds on average. In this study, the utterances recorded at the first session are used for training, and the phoneme models are constructed and the utterances recorded at the second session are used for testing. A full description of the corpus can be referenced in [74].

179 male speakers are used to train the UBM, and 78 male speakers were treated as target speakers, the remaining 78 male speakers were treaded as imposters.

After down sampling to 16 kHz, each 5 phoneme-balanced utterance was defined as training UBM, training set, and test set. The speech signals were analyzed with 25 ms window width and 10 ms shift. The feature vector size was 26-dimensional, which consisted of 12 mel-cepstrum coefficients, a log energy coefficient, and these delta mel-cepstrum coefficients. Since the utterance was not long enough to keep the mixture count at a large number, the mixture number was set at 32 for phoneme segments. Thus, each phoneme GMM could be concatenated into a 26*32=832 dimensional phoneme GMM-UBM super-vector. Then, several various phoneme GMM-UBM super-vectors continued to concatenate i.e., /a/, /i/, /u/, /e/, /o/, /m/, /n/ and, /s/&/sh/ into an 832*8 = 6656 dimensional GMM-UBM super-vector with equal weights. This was precisely corresponds to the baseline based 26*256 = 6656 dimensional GMM-UBM super-vector.

5.6.2 Experiment results

The results of comparative experiments are evaluated by equal error rate (EER) and detection error trade-off (DET) curves, which were also used for the performance criterion.

Table 5.1 Results comparison with respect to EER%

System		EER (%)
BASELINE		3.61
PHONE COMB (no NAS)		3.46
PHONE COMB	Parameter	
NAS-m	q = 1	3.34
	q = 2	3.52
	q = 3	3.60
NAS-sv	k = 1	3.48
	k = 26	3.93
	k = 52	3.85
	k = 78	4.63

In Table 5.1, BASELINE indicates the conventional method based on short phoneme-balanced utterances. PHONE COMB means the proposed method based on general phoneme combination without NAS. First, a TI speaker verification experiment is carried out with the GMM-UBM super-vectors concatenated by special phonemes to compare the system performance with that of BASELINE. Note that PHONE COMB led to an improvement in the system performance from 3.61% (achieved by BASELINE) to 3.46% (4.16% relative reduction of EER). The results reveal the fact that using some special phonemes extracted from phoneme-balanced utterance to model the GMM-UBM super-vectors can improve the performance when the enrolment utterances are relative short.

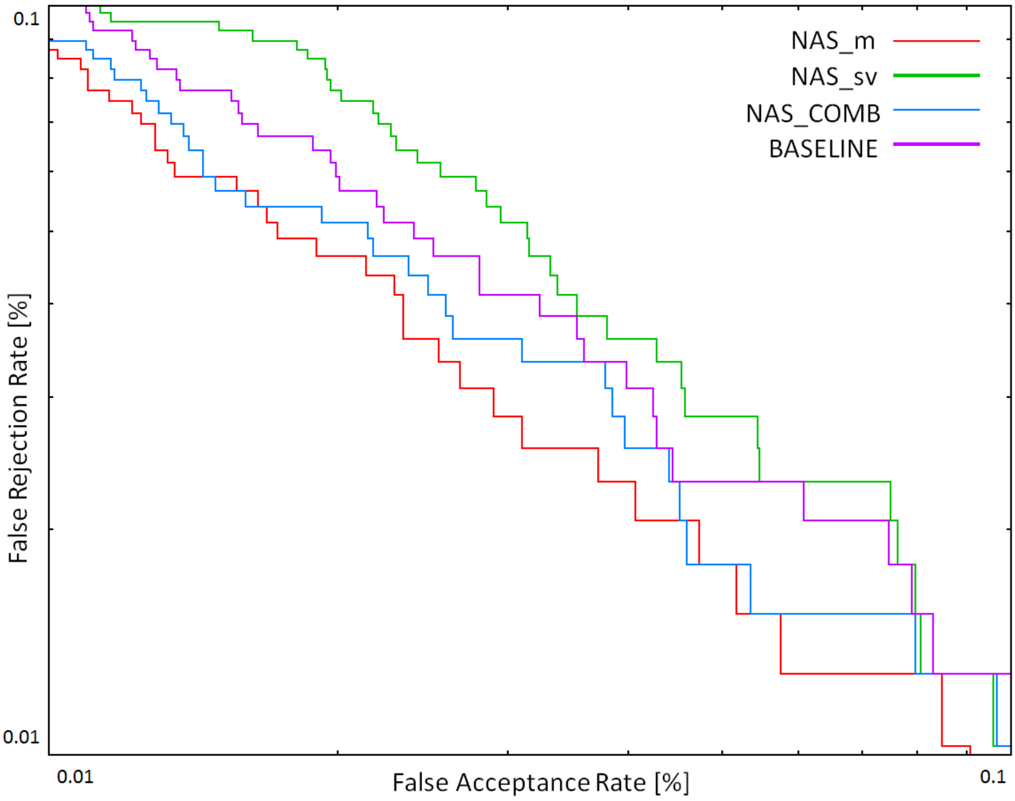


Figure 5.4 Detection Error Trade-off curves

To investigate the effect of inter-session reduction, several inter-session independent subspaces are compared by applying various parameters for NAS. From the results in Table 5.1, the best EER of 3.34% was achieved with NAS[1]-m, which can greatly improve the system performance up to 7.48% relative to the BASELINE. It is also superior to the ERR of 3.48% achieved by NAS[1]-sv, and the EER was reduced by 3.47% compared with the PHONE COMB. However, there was no improvement achieved with NAS-sv compared to PHONE COMB. The reason for this might be that NAS-sv had not successfully reduced any inter-session variability in the super-vectors domain owing to its phoneme GMM-UBM super-vectors being generated directly from short utterances.

For EER, given the simple transformation used, the performance of NAS[1]-m clearly shows the effectiveness of the proposed inter-session reduction procedure. Performance comparison of the BASELINE, PHONE COMB, NAS[1]-m, and

NAS[1]-sv is also shown in Figure 5.4 using DET curves. Here, it is observed that the superiority of the proposed method NAS[1]-m over the BASELINE system while it consistently provides further improvement compared to PHONE COMB in the DET range.

5.6.3 Discussion

Besides the above comparison experiments, table 5.2 also shows the EER of SVM-based TI speaker verification using various single phonemes in Table 5.2.

Table 5.2 EER% obtained using various phonemes

Phoneme	EER (%)
/a/	12.33
/i/	19.78
/u/	24.34
/e/	20.05
/o/	16.32
/m/	26.93
/N/	27.25
/s/	14.68

As illustrated in Table 5.2 and Table 5.1, it has been shown and confirmed that speaker identify information can be characterized by some phonemes and their combination, such as vowels and nasals, which are hard to affect by inter-session variability.

Furthermore, forced alignment is also conducted for test utterance segmentation to get a more accurate speech recognition rate and carry out the comparative experiments again.

Table 5.3 Results comparison with respect to EER% using forced alignment for test utterance segmentation

System		EER (%)
BASELINE		3.61
PHONE COMB (no NAS)		3.34
PHONE COMB	Parameter	
NAS-m	q = 1	2.91
	q = 2	3.34
	q = 3	3.70
NAS-sv	k = 1	3.48
	k = 26	3.79
	k = 52	4.35
	k = 78	4.42

With a more accurate speech recognition rate for phoneme segmentation, table 5.3, shows the EER% of PHONE COMB has been improved from 3.34% in Table 5.1 to 3.46%, also the best EER% achieved by NAS[1]-m has been improved from 3.34% in Table 5.1 to 2.91%. Compared with the conventional BASELINE method, it led to the improvement of system performance from 3.61% to 2.91% (a relative reduction of EER of 19.39%).

Obviously, for both experiments, when the reduced components were over the 2nd eigenvector, the performance of NAS began to decrease to some degree. The reason for this is that besides inter-session variability, there is no doubt that useful phonetic variability and especially speaker information also exist in the very low eigenvectors. When utterances were projected on them, they could be the factors that affected the performance of the EER.

5.7 Summary

In this chapter, an effective phoneme-based inter-session reduction method is developed including a new reduction called nuisance attribute subtraction (NAS) for TI speaker verification. In the method, for each phoneme GMM mean vector, given the eigenvector representing the direction of inter-session variability subspace obtained by applying PCA to a long-term corpus, the inter-session variability component was obtained as an inner product of the input vector and the eigenvector. By subtracting the obtained inter-session variability term from the original GMM mean vector, the inter-session variability was normalized. SVM-based TI speaker verification experiments are conducted using the proposed NAS in the phoneme GMM mean vectors domain and showed how a standard speaker verification system can be significantly improved. Therefore, a robust speaker model could be constructed by the new GMM-UBM super-vectors with less inter-session variability obtained in an inter-session independent subspace to get a better TI speaker verification performance.

Chapter 6

Robust extraction of desired speaker's utterance in overlapped speech

6.1 Introduction

By the popularization of the smart phone equipped with voice recording function like IC recorder these recent years, the voice recording becomes easier and more convenient at quite a long time meeting. If you want to detect a voice section of a desired speaker from the recording speech, it obviously costs a lot of effort only through listening and searching manually by human being. Therefore, if it is possible to automatically label the desired speaker's voice section from the recording speech by speaker recognition technique, a quick and accurate system for extracting the desired speaker's voice section will be realized, which can also be expected as a useful information search tool.

For this purpose, in recent years, the study of speaker diarization has been actively developed [2, 101, 102]. In a general way, speaker diariztion is asked to detect who and when speaking the voice under the condition that the number of conversation participants and the length of voice are unknown [2]. As mentioned of the process, the input speeches for evaluation are divided into multiple clusters in the first place. Second, analyze the features of speech clusters, and the partition not similar will be further divided into more clusters and then the similar partition will be combined. Finally, the repeated dividing and combining processes will keep running until the predetermined stop criterion is met. It's the popular way to estimate who and when is speaking by dividing the speech into segments [2, 101, 103, 104, 105].

There also are other researches for speaker diariztion like using the video and audio information, detecting who is speaking to whom with fisheye lens [106], identifying who is photographed with face recognition [107] detecting speaker with audio and mouth information [108].

Of course, speaker diarization are still facing a lot of problems most of which is how to accurately extract the desired speaker's utterance in an overlapped conversational speech [2, 109]. Of course, it is possible to do conduct speaker diariztion correctly when a speech section is uttered by just only one speaker, owing to the likelihood of evaluation speech that uttered by the same speaker is relatively higher than the likelihood of the overlapped speech. The likelihood, however, probably decreases due to the noise of conversation itself in a relative complicated overlapped conversational speech. As a result, a speech should have been identified as a section of target speaker might be false rejected. Vice versa, it also probably comes higher for the target speaker's model's likelihood which should have been lower due to other speakers' overlapped speech.

In order to resolve the problem of performance decreasing resulted from the overlapped speech, it is necessary to detect the section of overlapped speech in the first place. There exists several detection methods based on the HMM/GMM using the silence model, single speech model and overlapped speech model [2, 110, 111]. On the other hand, some researches focus on detecting the ratio of the silence section since only few silence section exists in an opinion exchange frequent conversational speech [111].

If the identified overlapped speech section is labeled, it is possible to divide the whole section into each speaker's clusters using the single speaker model. In addition, it is the general way that two speakers who get the highest scores can be regarded as the corresponding conversation speakers through calculating the scores between the labeled overlapped speech and each speaker's model [109, 110].

This paper is conditional on detecting a desired speaker (afterwards also called target speaker). A speaker indexing method is proposed using speaker verification technique to extract target speaker's utterances from conversational speech. Although the target of this study is easier than the speaker diariztion mentioned above, it still has considerably practical significance in many occasions, such as detecting the speech section of a target speaker's himself introduction at the beginning of a meeting in which is usually labeled manually. In this study, with extracting the unknown speaker's (afterwards also called cohort speaker) speech from the observed speech, a method is proposed using two kinds of overlapped speeches that are target and cohort speaker's overlapped speech, other cohort speakers' overlapped speech. The study is also based on the following requirements: there are three speakers in the conversational speech, and two speakers' speeches are overlapped. Except for the target speaker, other two cohort speakers' speech are unknown. The effectiveness of constructing the model of undetected speaker's speech and overlapped speech from the observed speech is presented. Besides two kinds of overlapped speech models mentioned earlier, there are seven kinds of speech models in total (single speech UBM, overlapped speech UBM, target speaker model, target and cohort speaker overlapped speech model, other cohort speakers' model, target and other cohort speakers' overlapped speech model, other two speakers' overlapped speech model). After calculating their scores, the indexing experiments are carried out using Support Vector Machine (SVM) [100] and then the evaluation experiment results showed the effectiveness of our proposed method.

The remainder of this paper is organized as follows. In section 6.2 introduces process of detection of target speaker's speech section in detail. In section 6.3, experiments are carried out using the proposed method to show the effectiveness of our proposed method. Finally, section 6.4 and section 6.5 gives the discussion and summary, future work of this chapter.

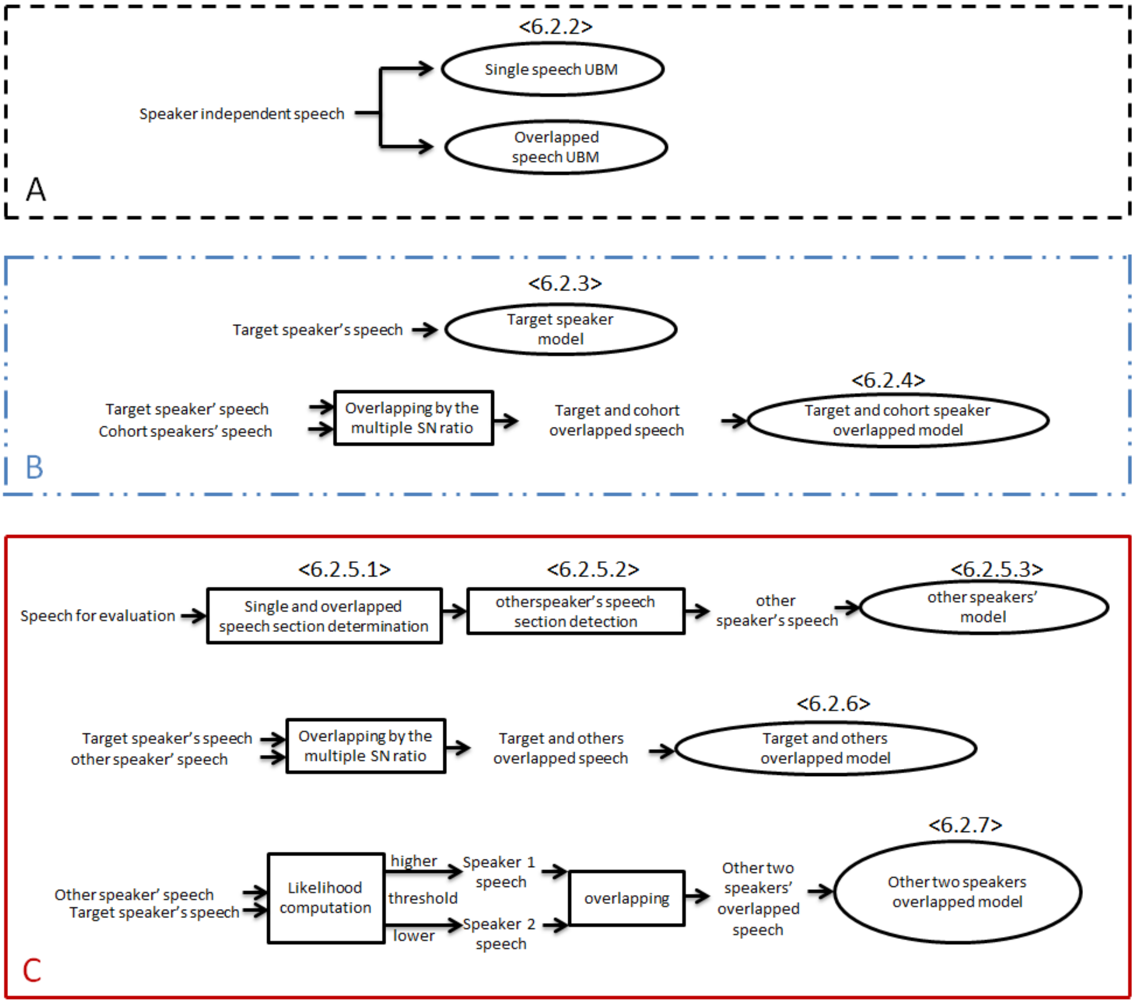


Figure 6.1 A block diagram of the proposed method of the model construction

6.2 Detection of target speaker's speech section

6.2.1 Process of detecting the target speaker's speech section

This section introduces the process of detecting the target speaker's speech section which means the detection of the registered speaker's speech section from the input speeches. For detecting the target speaker's speech section, speaker verification is performed by every section of a certain length (one second in this paper) from the beginning of the conversational speech. The process is started with that the result of speaker verification in this section is different from the previous section, and then end in the same way.

In this study, the detection of speaker’s speech section is performed by every second and shift by every 0.5 second. That means the result of speaker verification is output by every 0.5 second.

In this study, cohort speaker’s speech not overlapped is extracted from the conversational speech itself. It is utilized to construct the speaker model for speaker verification. Fig.1 shows the image of constructing each speech model i.e. single speech UBM, overlapped speech UBM of section 6.2.2, target speaker model of section 6.2.3, target and cohort speaker overlapped speech model of section 6.2.4, other cohort speakers’ model of section 6.2.5, target and other cohort speakers’ overlapped speech model of section 6.2.6, other two speakers’ overlapped speech model of section 6.2.7. Section 6.2.2 of A is the process of constructing UBM with large scale of corpus. Section 6.2.3~6.2.4 is the process of constructing the model with labeling speech section manually. Section 6.2.5~6.2.7 of B is the process of constructing the model using the speech extracted from the conversation by proposed method.

6.2.2 UBM

The process consists of constructing single speech UBM and overlapped speech UBM with large scale of corpus. The following is the detail of construction for the two kinds of UBM.

All the undesignated speakers’ non-overlapped speech is utilized to construct the single speech UBM through maximum likelihood estimation.

On the other hand, for constructing overlapped speech UBM, the computer has two speakers’ speech overlapped when the signal-noise ratio of speech that constructs the single speech UBM equals 0. Multiple speaker is reading the same sentence in general, actually, however, given the situation where it is rare to read the same sentence, the different sentence is used to pile up. And then the overlapped speech is utilized to construct the overlapped speech UBM. The large scale of corpus was developed by the National Research Institute of Police Science (NRIPS).

6.2.3 Target speaker model

The speech section of a target speaker's himself introduction is detected at the beginning of a meeting in which is labeled manually and then extract the target speaker's speech. Target speaker's model can be generated by adapting a well-trained UBM using the maximum a posteriori (MAP) adaptation approach. The process can be indicated by function (6.1) as below:

$$\widehat{\mu_1} = \frac{\tau m_i + \sum_{t=1}^{T} c_{it} x_t}{\tau + \sum_{t=1}^{T} c_{it}}, \qquad (6.1)$$

where m_i is the i-th component's mean vector of UBM, x_t is the adapted speech vector, c_{it} is the Gaussian probability of each mixture, T is the number of Frames and the t is the frame. Obviously, how much $\widehat{\mu_1}$ close to UBM is defended by τ. Namely, a mixture-dependent adaptation of parameters is allowed by using a data-dependent adaptation coefficient τ. If τ → 0, the function (6.1) is causing the use of the new target speaker-dependent parameters, otherwise, if τ → ∞, $\widehat{\mu_1}$ is closer to the UBM.

6.2.4 Target and cohort speaker overlapped model

The computer has cohort speakers' speech overlapped with the target speaker's speech to generate the target and cohort speaker overlapped model. The cohort speakers' speech is random extracted and overlapped in the signal-noise ratio equals -6dB, -3dB, 0dB and 6dB.

6.2.5 Other speakers' model

Other speaker model is generated by other speakers' speech extracted from the conservation speech. The detail process is followed by the step as below.

6.2.5.1 Single and overlapped speech section determination

The log likelihood of conservation speech and the two kinds of UBM generated in (6.2.2) is computed respectively and then carried out the examination of likelihood ratio. The log likelihood L of segment i is computed by the following function (6.2):

$$L = log\frac{P(x_i|H_1)}{P(x_i|H_0)}, \qquad (6.2)$$

x_i indicates the observed segment i while H_0 is the single speech and H_1 is the overlapped speech. The speech is determined as single speech if L is greater than the threshold and conversely it is determined as overlapped speech. Fig 6.2 is a sample of determination.

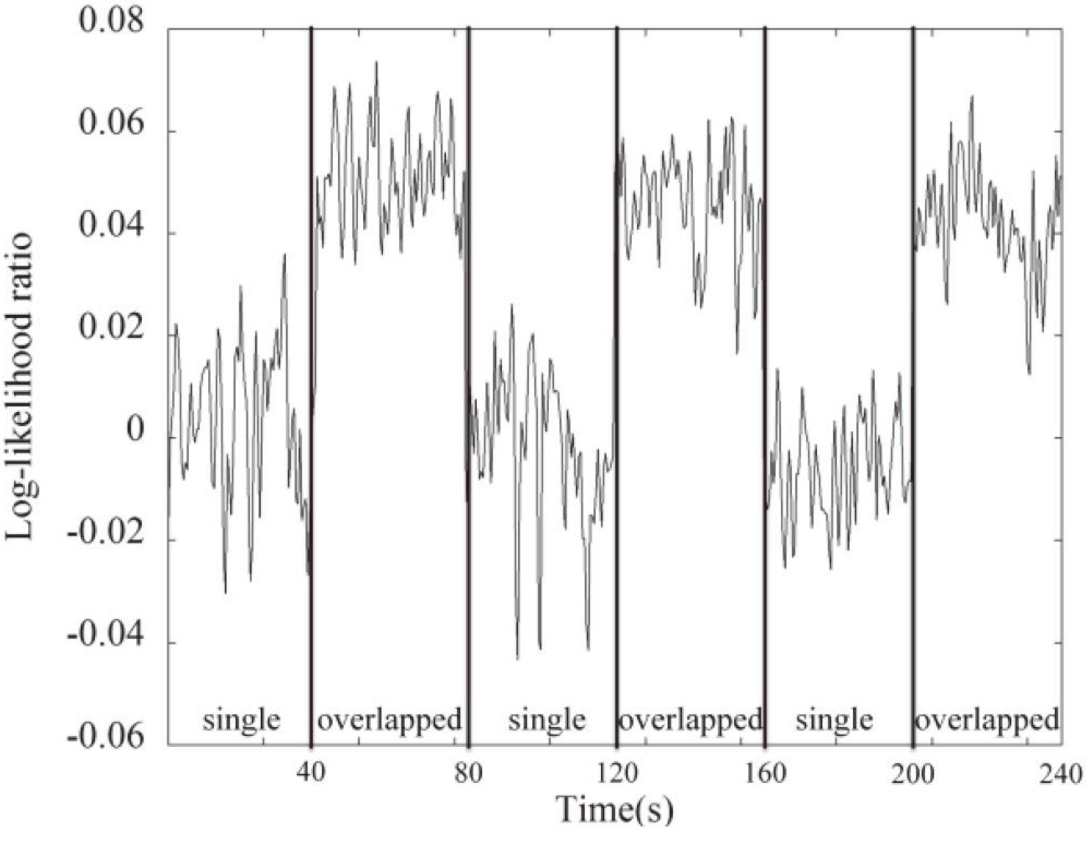


Figure 6.2 Single/overlapped speech section determination

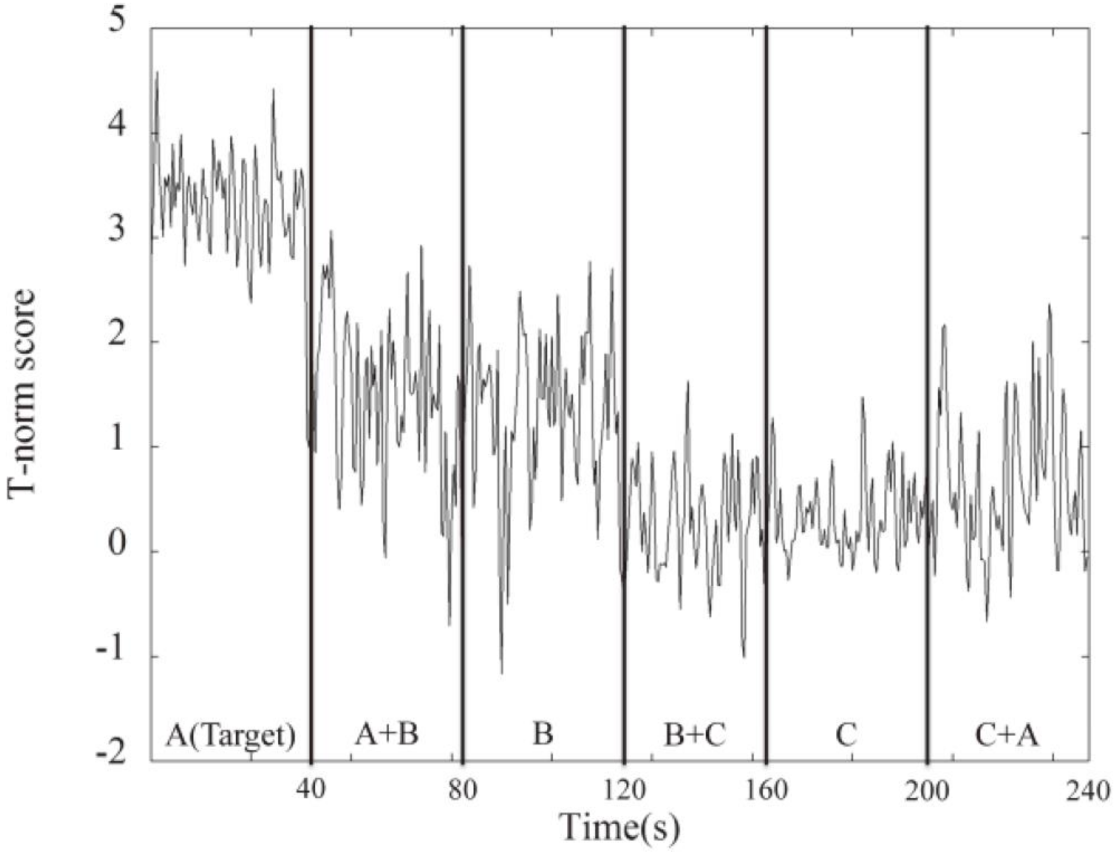


Figure 6.3 T-norm score of each section

6.2.5.2 Other speakers' speech section detection

T-norm shares similarities with model-based cohort normalization while additionally incorporating score variance to better model the cohort score distribution to address the issue caused by the conditions exhibited. T-norm score $\widetilde{S_c}(X_i)$ of target speaker model is computed from the single speech detected in (6.2.5.1). The mean μ_λ and standard deviation σ_λ of the imposter scores were then used to adjust the target score as function (6.3):

$$\widetilde{S_c}(X_i) = \frac{S_c(x_i)-\mu_\lambda}{\sigma_\lambda}. \qquad (6.3)$$

A sample of T-norm score of single speech and overlapped speech is given by Fig. 6.3 gives where the x-axis is time and y-axis is T-norm score. And then the other speakers' speech section is detected if the T-norm score is below the threshold.

6.2.5.3 Other speakers' model

Other speakers model can be generated by adapting a well-trained single speech UBM of (6.2.2) using the maximum a posteriori (MAP) adaptation approach.

6.2.6 Target and others overlapped model

The computer has other speakers' speech overlapped with the target speaker's speech to generate the target and others overlapped model. Other speakers' speech is random extracted and overlapped in the same way as described in (6.2.4) in which signal-noise ratio equals -6dB, -3dB, 0dB and 6dB. Target and others overlapped model can be generated by adapting a well-trained single speech UBM of (6.2.2) using the maximum a posteriori (MAP) adaptation approach.

6.2.7 Other two speakers overlapped model

The mean of T-Norm score is computed using the other speakers' speech and target speech. The mean can be seen as a threshold to determine whose speech belongs to speaker A and whose speech belongs to speaker B. After overlapping the other speakers' speech, other speakers overlapped model can be generated by adapting a well-trained single speech UBM of (6.2.2) using the maximum a posteriori (MAP) adaptation approach.

6.2.8 SVM

Extraction of desired speaker's Utterance in overlapped speech is identified segment by segment using support vector machine (SVM). The seven kinds of parameters, that is the likelihood for evaluation speech and single speech UBM, overlapped speech UBM, target speaker model, target and cohort speaker overlapped model, other speaker model, Target and others overlapped model, other speakers overlapped model is utilized as the input 7-dim vector of SVM.

6.3 Experiments

The evaluation experiments were processed based on a SVM system. In order to show the effectiveness of the proposed model, the experiments is also conducted using three conventional methods as the input parameter of SVM, which are ① T-norm score of target speaker model described in (6.2.5.2), ② Likelihood for single speech UBM, Likelihood for overlapped speech UBM + ①, ③ Likelihood for target and cohort speaker overlapped model + ②, Table 5.1 shows the detail of each method.

Table 6.1 Input vector of SVM

①	T-norm score of target speaker model
②	T-norm score of target speaker model Likelihood for single speech UBM Likelihood for overlapped speech UBM
③	T-norm score of target speaker model Likelihood for single speech UBM Likelihood for overlapped speech UBM Likelihood for target and cohort speaker overlapped model
④	T-norm score of target speaker model Likelihood for single speech UBM Likelihood for overlapped speech UBM Likelihood for target and cohort speaker overlapped model T-norm score of other speakers' model Likelihood for target and others overlapped model Likelihood for other two speakers overlapped model

6.3.1 Condition

Evaluation Experiments is conducted with a large Japanese corpus separated into four cross-validations. The corpus was developed by the National Research Institute of Police Science (NRIPS). It contains speech data of 336 Japanese males aged 18 to 76 years, and was recorded in two time sessions over three months. Each ATR phoneme-balanced sentence has a length of about 2-5 seconds. In this study, the utterances recorded at the first session are used. A full description of the corpus can be referenced in [74].

150 male speakers are used to train the UBM, and the remaining 186 male speakers were separated into 61 groups to test the performance of extraction of desired Speaker's utterance in overlapped speech. After down sampling to 16 kHz, 5 phoneme-balanced utterance (A01-A05) are defined as training UBM set and the following A06-A10 are adapted to MAP the UBM into GMM. T-norm speakers involve 183 of all 186 speakers who do not attend the corresponding conservation speech.

A11-A50 are defined as evaluation utterance. Fig.6.4 defines the rules of the conversational speech for evaluation, Speaker A is regarded as the target speaker and then Speaker B or C is regarded as other speaker. The silence section is removed in all the 40 speech sentences and then joint the silence section removed speech. The length of each speaker's jointed speech is around 120~160 seconds. The first 40 seconds of the jointed speech is single speech and next 40 seconds speech is overlapped with other speaker's speech. The results of 40 seconds overlapped speech are showed in 6.3.2. Another set of experiments is also carried out by reducing the time of overlapped speech to 20 seconds, 10 seconds, 5 seconds, and 2.5 seconds and the results are showed in 6.3.3.

The detection of speaker's speech section is performed by every second and shift by every 0.5 second. The feature vector size was 26-dimensional, which consisted of 12 mel-cepstrum coefficients, a log energy coefficient, and these delta mel-cepstrum coefficients. A detail of speech analysis is showed in table 6.2. Owing to the results of previous experiments, the RBF ($\gamma = 0.25$) is employed as the kernel function of SVM.

Table 6.2 Acoustic analysis conditions

Sample Frequency	16kHz
Pre-emphasis	$1 - 0.97z^{-1}$
Frame Length	25ms
Frame Periodicity	10ms
Window	hamming
Number of Mel-Filter Bank	24

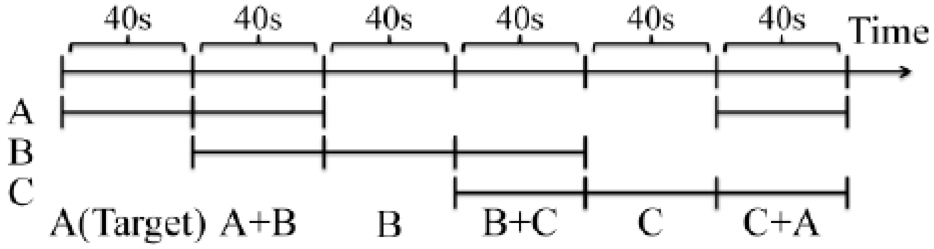


Figure 6.4 Speech for evaluation

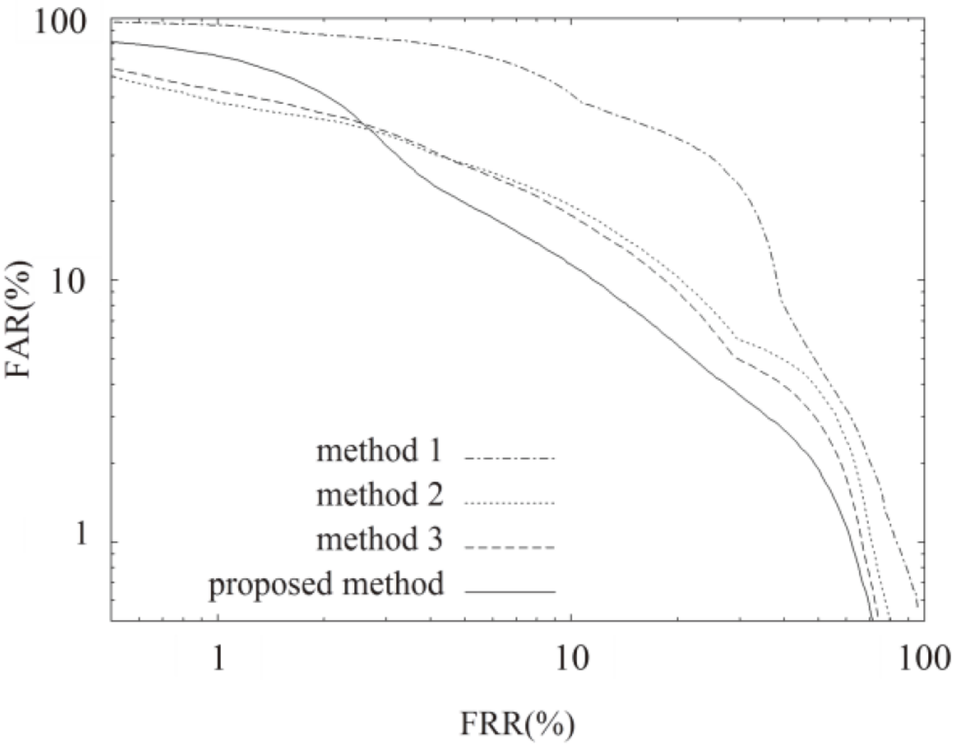
6.3.2 Experiment results 1

The results of extraction of target speaker’s utterance in all speech segments and in overlapped segments are evaluated respectively by detection error trade-off (DET) curves and equal error rate (EER). Fig.6.5 shows the DET curves and Table 6.3 shows the EER.

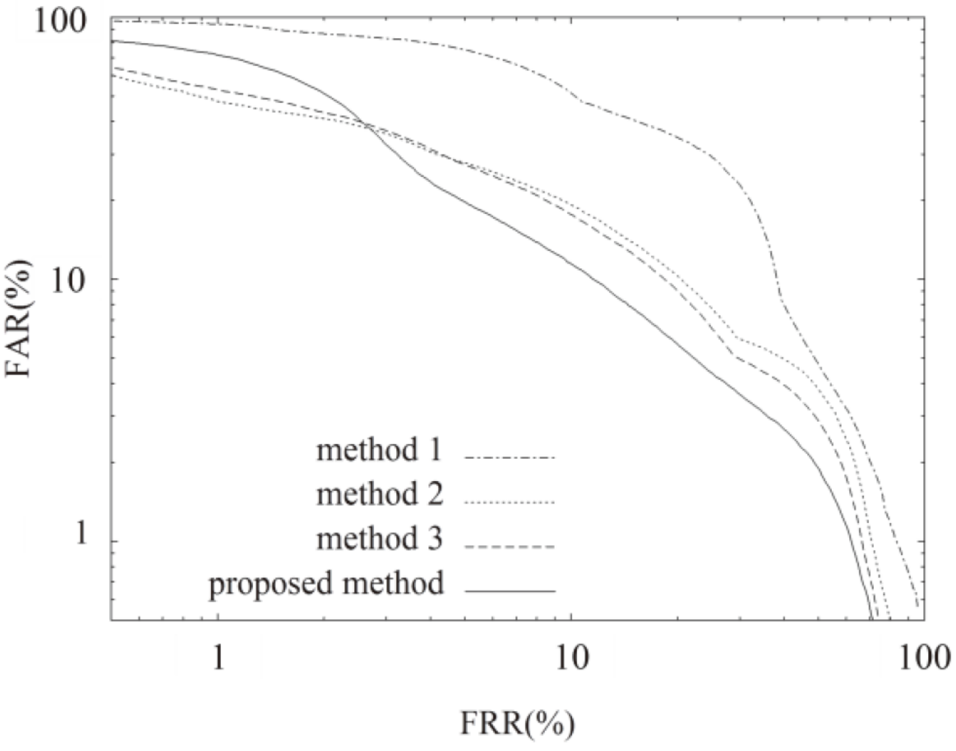
In the Table 6.3, the best extraction result is obtained by the proposed method compared to the other conventional methods. Compare the results obtained by method ① with method ②, we can see the extraction results is improved when the overlapped speech model is used. Compare the results obtained by method ② with method ③, we can see the extraction results is improved in the overlapped speech segments when target speech is included in the overlapped speech model. Furthermore, compare the results obtained by method ③ and method ④, the extraction result is further improved by the proposed overlapped speech model when using the speech extracted from the observed conversation speech.

Table 6.3 Experiment result 1 (EER (%))

	all segments	overlapped segments
Method ①	26.28	32.87
Method ②	14.30	25.99
Method ③	13.60	24.45
Proposed method ④	10.72	18.77



(a) All segments



(a) All segments

Figure 6.5 DET curve of each method

6.3.3 Experiment results 2

Another set of experiments is also carried out by reducing the time of overlapped speech to 20 seconds, 10 seconds, 5 seconds, and 2.5 seconds and the results are showed in table 6.4 (a)~(d). The training condition of SVM is exactly same as the time of overlapped speed 40 seconds in experiment 1.

Table 6.4 Experiment result 1 (EER (%))

(a) Time of overlapped speech 20 Secs

	all segments	overlapped segments
Method ①	21.29	30.89
Method ②	10.25	25.67
Method ③	9.62	24.20
Proposed method ④	7.48	18.42

(b) Time of overlapped speech 10 Secs

	all segments	overlapped segments
Method ①	14.01	31.11
Method ②	6.18	25.73
Method ③	5.66	24.38
Proposed method ④	4.97	18.15

(c) Time of overlapped speech 5 Secs

	all segments	overlapped segments
Method ①	9.59	31.11
Method ②	6.18	25.73
Method ③	5.66	24.38
Proposed method ④	4.97	18.15

(d) Time of overlapped speech 5 Secs

	all segments	overlapped segments
Method ①	9.59	31.19
Method ②	3.96	25.60
Method ③	3.63	24.66
Proposed method ④	2.75	18.12

Even if the time of overlapped speech is reduced, the EER obtained by the proposed method is still better than the EER obtained by other conventional method.

6.4 Discussion

T-norm score of other speakers’ model, likelihood for target and others overlapped model, likelihood for other two speakers’ overlapped model is showed in the Fig.6.6~6.8 respectively. The mean of likelihood is also showed above the every corresponding section.

At first, let us focus on the Fig.6.6 of T-norm score of other speakers’ model. The score of speaker B or C’s single speech section is higher than the score of speaker A’s single speech section. The score of other speakers’ overlapped speech (B+C) is a little higher than the score of target speaker’s overlapped speech (A+B, A+C) even if it is not so much obvious.

Second, let us focus on the Fig.6.7 of likelihood for target and others overlapped model. This is what is proposed in this section. The likelihood for target speaker’s overlapped speech (A+B, A+C) is obviously higher than the likelihood for other speakers’ overlapped speech (B+C). According to the scores of all sections including single speech and overlapped speech, the score of the target speaker’s speech (A, A+B, B+C) is still higher than the score of the speech section excluding target speaker. Compared with the Fig.6.3, the score of speaker B’s single speech is higher than the score of the speakers’ overlapped speech (C+A) while the likelihood for speaker B’s single speech decreases in Fig.6.7.

Finally, continue to focus on the Fig.6.8 of likelihood for other two speakers’ overlapped model. The likelihood for other speakers' overlapped speech (B+C) is higher than the likelihood for target speaker's overlapped speech (A+B, C+A). The likelihood for target speaker's single speech is lower than the likelihood for other speaker B or C's single speech. Therefore, other speakers' overlapped speech model is considered to contribute to the target speaker verification.

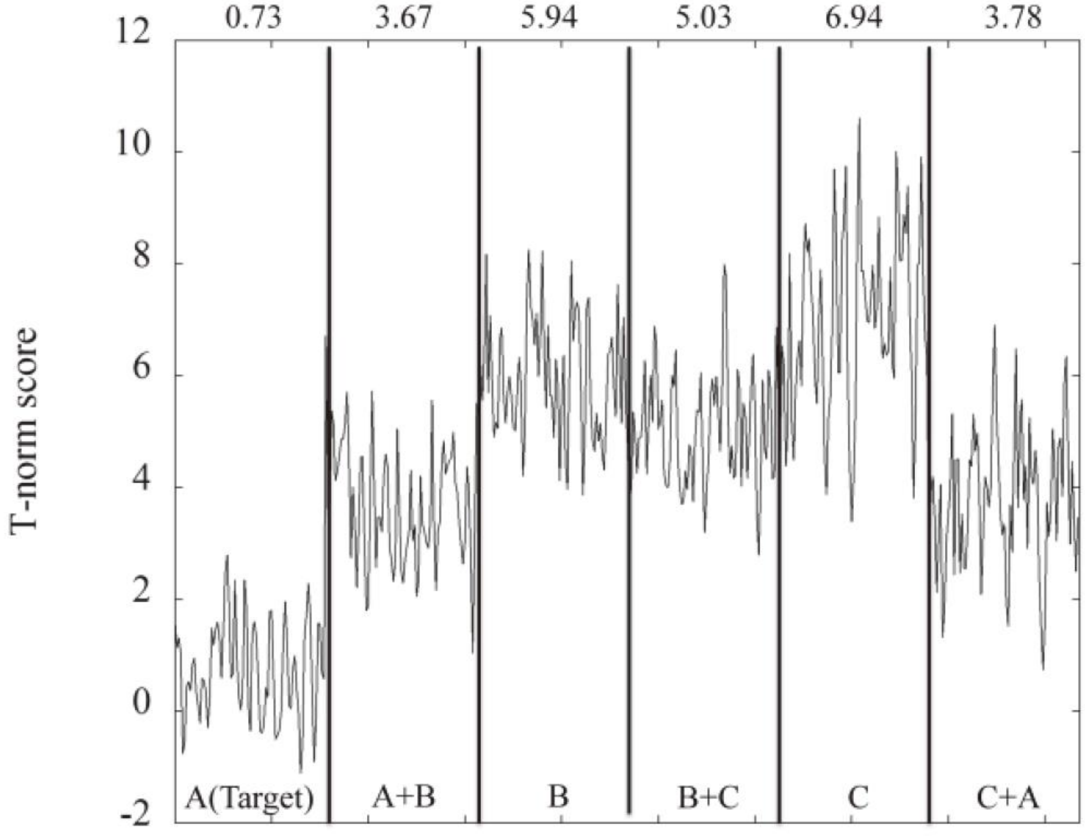


Figure 6.6 T-norm score of other speakers' model

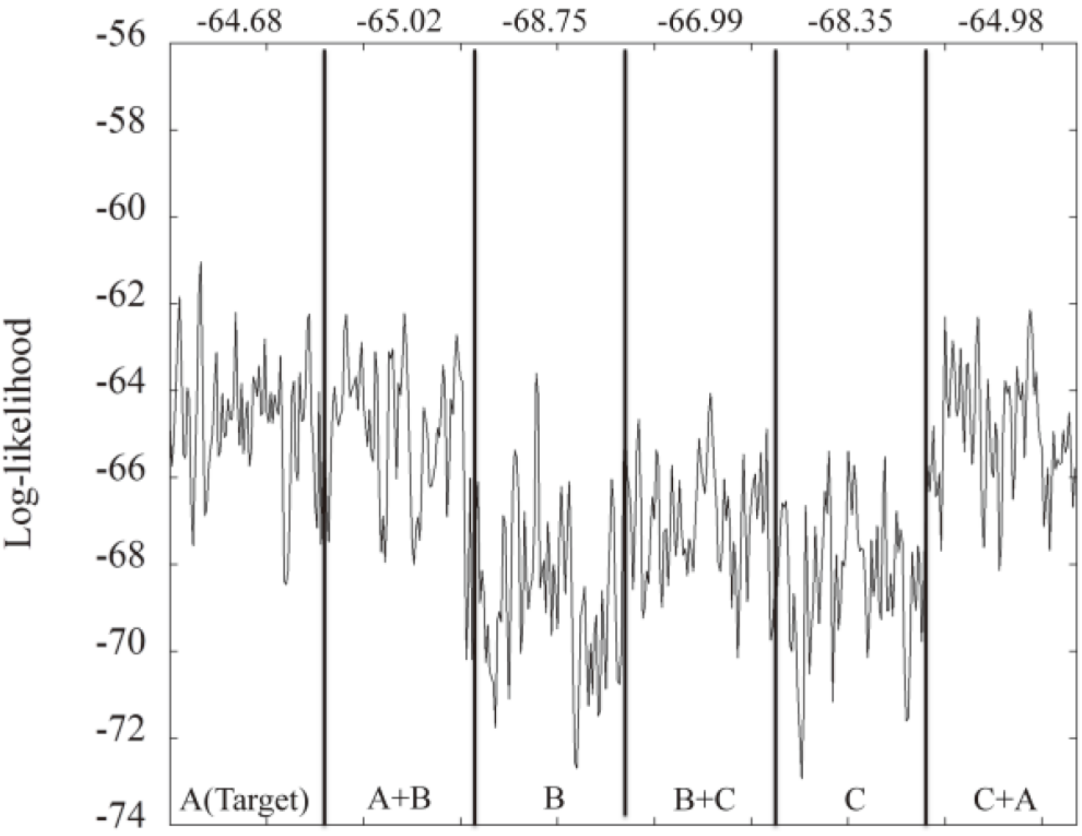


Figure 6.7 Likelihood for target and others overlapped model

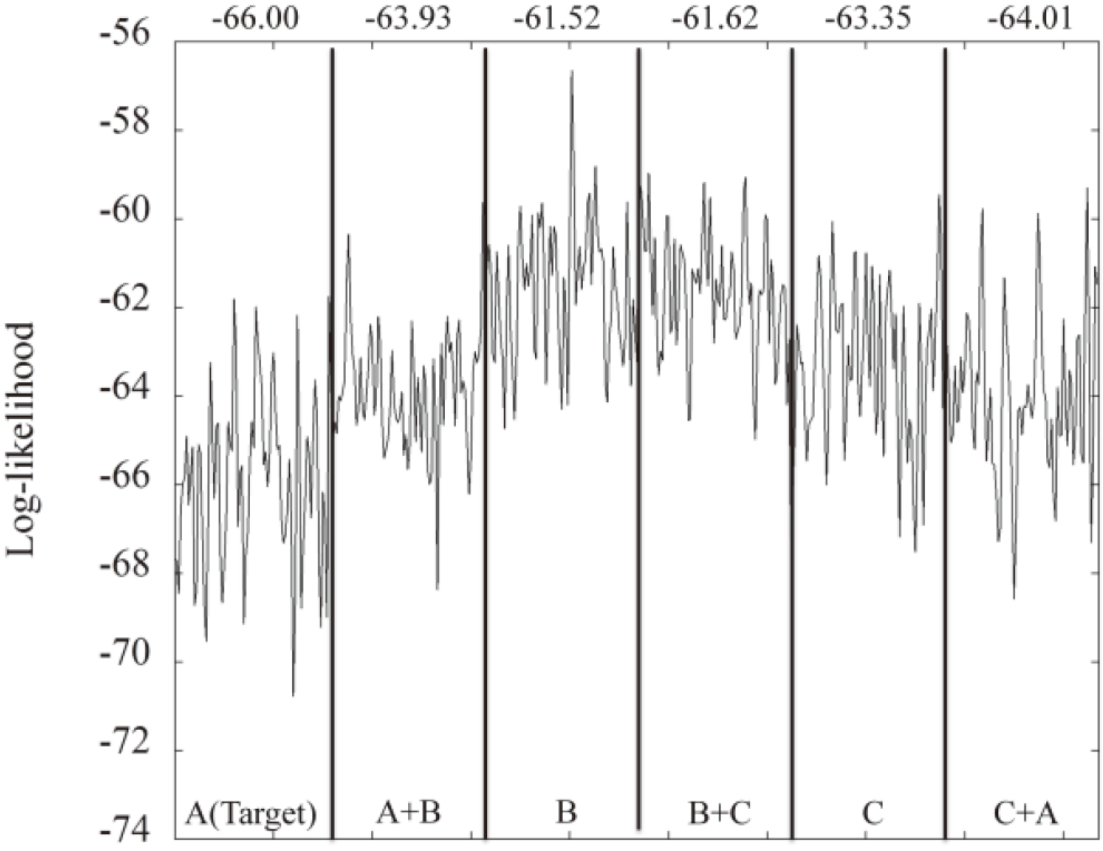


Figure 6.8 Likelihood for other two speakers overlapped model

Furthermore, to confirm the effectiveness of the proposed method, the verification experiments are performed using the additional method ③ + T-norm score of other speakers' model, ③ + likelihood for target and others' overlapped model, ③ + likelihood for other two speakers' overlapped model. The results are showed in Table. 6.5 (the time of overlapped speech is 40 seconds). The best result is obtained by ③ + likelihood for other two speakers' overlapped model that reduced EER by 4~5%.

Table 6.5 Verification experiment result (EER (%))

	all segments	overlapped segments
③ (for reference)	13.60	24.45
③ + T-norm score of other speakers' model	11.41	20.18
③ + likelihood for target and others' overlapped model	11.12	19.12
③ + likelihood for other two speakers' overlapped model	11.48	20.62
Proposed method ③ (for reference)	10.72	18.77

6.5 Summary

In this study, a new speaker indexing method is proposed using speaker verification technique to extract one desired speaker's utterances from conversational speech. It is under the condition that there are 3 speakers, two of whose speech is overlapped. The proposed method detected other speakers' speech from the observed speech itself. And then the computer has target speaker's speech overlapped with other speakers' speech to generate the overlapped speech model. This is also the feature of this study.

Compared with the conventional methods that are T-norm score of target speaker model using the speaker verification technique, Likelihood for single speech UBM, Likelihood for overlapped speech UBM, Likelihood for target and cohort speaker overlapped model, the experiment results shows the performance can be improved by using the proposed method. Furthermore, under the condition of overlapped speech only, the EER was reduced by up to 43.7% compared with the conventional methods that use a target speaker model and overlapped speech model. The EER is reduced by up to 21.2% under the condition of all speech. Therefore, a robust

Therefore, a robust speaker indexing system to extract one desired speaker's utterances from conversational speech can be constructed by the proposed method even in a condition of overlapped speech.

In the future work, it will be discussed how to identify the other two unknown speakers and to generate the overlapped speech model speaker by speaker. There also needs to conduct the desired speaker indexing experiment under a more actual conservation environment without the limit on the number of speakers.

Chapter 7

Conclusions and Future works

7.1 Introduction

This chapter provides a summary of the work presented in this dissertation and the conclusions drawn. The summary follows the four main research themes in aforementioned chapters — new speech feature with less phonetic variability; inter-session variability reduction for MFCCs in a GMM-based speaker identification; innovative phoneme dependent inter-session variability reduction for SVM-based speaker verification; and an application of a multi-speaker diarization using speaker verification technology.

7.2 Speech feature with less phonetic variability

Chapter 3 described the adverse effect of phonetic variability along with conventional speech feature MFCC. A new speech feature for TI speaker identification that suppresses the phonetic variability by a subspace method was proposed, under the assumption that a subspace with large variance in the speech feature space is a ‘phoneme-dependent subspace’ and a complementary subspace of it is a ‘phoneme-independent subspace’. PCA is employed to construct these subspaces. GMM-based speaker identification experiments using both the phonetic variability suppressed feature and the conventional MFCC were carried out. As a result, the proposed method has been proven to be effective for decreasing the identification error rates.

7.3 Inter-session variability reduction for MFCCs

Chapter 4 saw the proposal of a novel speech feature MFCCs for the purpose of inter-session variability reduction in speaker identification systems. It adopted and extended the previous PCA-based phonetic variability suppression method to reduce the inter-session variability. The basic idea is to transform input vectors to another subspace, where the inter-session variability and other factors including speaker information were separated into different subspaces.

7.4 Innovative phoneme dependent inter-session variability reduction for SVM-based speaker verification

Chapter 6 presented that an innovative phoneme-dependent using speech recognition technique was integrated with GMM-SVM-based speaker verification. This technique selects the phonemes with high contribution for speaker verification so as to overcome the shortcoming of inter-session variability along with the traditional GMM-UBM super-vectors. A speaker's model can be represented by several various phoneme Gaussian mixture models. Each of them covers an individual phoneme whose inter-session variability can be constrained in an inter-session independent subspace constructed by a reduction method. The reduction method is termed as nuisance attribute subtraction (NAS). SVM-based experiments performed using a large Japanese speaker recognition evaluation corpus constructed by the National Research Institute of Police Science (NRIPS) demonstrate the improvements gained from the proposed method.

7.5 Robust extraction of desired speaker's Utterance in overlapped speech

A new speaker diarization method using speaker verification technique is proposed in chapter 6. The proposed method detected other speakers' speech from the observed speech itself. And then the computer has target speaker's speech overlapped with other speakers' speech to generate the overlapped speech model to extract one desired speaker’s utterances from the overlapped speech. The experiment results shows the performance can be improved by using the proposed method. Furthermore, under the condition of overlapped speech only, the EER was reduced by up to 43.7% compared with the conventional methods that use a target speaker model and overlapped speech model. The EER is reduced by up to 21.2% under the condition of all speech.

7.6 Future work

The work aims to improve the classification performance and practicality of text-independent, speaker recognition systems via the removal of the intra-speaker variability of speech. The improvement of classification errors was achieved through the subspace–based reduction method. However, there is a factor also needs to be taken into account. It is emotion factor that is supposed to affect the accuracy. It is expected that speaker’s emotional characteristic is to be held via database refinement.

On the other hand, the illustration of differences between the GMM and SVM domain for the reduction of variation provides motivation for further development of robust modeling techniques. Meanwhile, to improve the estimation of the intra-speaker variability subspace with the NIST SRE data is also the future work.

References

[1] R. Clarke, “Human identification in information systems: Management challenges and public policy issues,” *Information Technology & People,* Vol. 7, No. 4, pp. 6-37, 1994.

[2] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, and G. Friedland, “Speaker Diarization: A Review of Recent Research,” *IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING,* Vol. 20, No. 2, Feb 2012.

[3] Tomoko Matsui,Sadaoki Furui, “Comparison of Text-Independent Speaker Recognition Methods Using VQ-Distortion and Discrete/Continuous HMM's,” *IEEE Transactions on Speech and Audio Processing*, Vol. 2, No. 3, pp.456-459, Jul. 1994.

[4] Mathieu Ben, Michael Bester, Frederic Bimbot, and Guillaume Gravier, “Speaker diarization using bottom-up clustering based on a parameter-derived distance between adapted GMMs,” in *Proc. of ICSLP*, 2004.

[5] W. Zhang, Y. Yang, Z. Wu, and L. Sang, “Experimental evaluation of a new speaker identification framework using PCA,” *IEEE International Conference on SMC,* pp. 4147-4152, 2003.

[6] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik, “Feature selection for SVMs,” *Advances in Neural Information Processing Systems,* vol. 13, pp. 668-674, 2000.

[7] D. Reynolds, “Experimental evaluation of features for robust speaker identification,” *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 639-643, 1994.

[8] R. Mammone, X. Zhang, and R. Ramachandran, “Robust speaker recognition: A feature-based approach,” *IEEE Signal Processing Magazine*, vol. 13, no. 5, pp. 58-71, 1996.

[9] L. Rabiner and B. Juang, *Fundamentals of speech recognition*. New Jersey, USA: Prentice Hall, 1993.

[10] D. Reynolds, “Comparison of background normalization methods for text-independent speaker verification,” in *Proc. Eurospeech*, vol. 2, pp. 936-966, 1997.

[11] D. Reynolds, “Channel robust speaker verification via feature mapping,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 53–56, 2003.

[12] J. Campbell, “Speaker recognition: A tutorial,” *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437-1462, 1997.

[13] F. Bimbot, J. F. Bonastre, C. Fredouille, G. Gravier, I. Magrin Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovsk Delacrtaz, and D. Reynolds, “A tutorial on text-independent speaker verification,” *EURASIP Journal on Applied Signal Processing*, vol. 4, pp. 430-451, 2004.

[14] R. Mammone, X. Zhang, and R. Ramachandran, “Robust speaker recogni- tion: A feature-based approach,” *IEEE Signal Processing Magazine*, vol. 13, no. 5, pp. 58-71, 1996.

[15] A. Stolcke, L. Ferrer, S. Kajarekar, E. Shriberg, and A. Venkataraman, “MLLR transforms as features in speaker recognition,” in *Proc. Eurospeech*, pp. 2425-2428, 2005.

[16] S. Furui, “Cepstral analysis technique for automatic speaker verification,” *IEEE*

Transactions on Acoustics, Speech and Signal Processing, vol. 29, no. 2, pp. 254-272, 1981.

[17] H. Hermansky and N. Morgan, “RASTA processing of speech,” *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578-589, 1994.

[18] J. Pelecanos and S. Sridharan, “Feature warping for robust speaker verification,” *A Speaker Odyssey, The Speaker Recognition Workshop*, vol. 2001, pp. 213-218, 2001.

[19] K. Yiu, M. Mak, and S. Kung, “Environment adaptation for robust speaker verification,” in *Proc. of Eurospeech*, pp. 2973-2976, 2003.

[20] O. Viikki and K. Laurila, “Cepstral domain segmental feature vector normalization for noise robust speech recognition,” *Speech Communication*, vol. 25, no. 1-3, pp. 133–147, 1998.

[21] R. Vogt, Automatic Speaker Recognition Under Adverse Conditions. *PhD thesis*, Queensland University of Technology, Brisbane, Queensland, 2006.

[22] J. Haigh and J. Mason, “Robust voice activity detection using cepstral features,” *IEEE Region 10 Conference on Computer, Communication, Control and Power Engineering (TENCON’93)*, pp. 321-324, 1993.

[23] R. Tucker, “Voice activity detection using a periodicity measure,” *IEEE Proceedings I. Communications, Speech and Vision*, vol. 139, no. 4, pp. 377-80, 1992.

[24] K. S¨onmez, L. Heck, and M. Weintraub, “Multiple speaker tracking and detection: Handset normalization and duration scoring,” *Digital Signal Processing*, vol. 10, no. 1-3, pp. 133-142, 2000.

[25] F. Soong, A. Rosenberg, L. Rabiner, and B. Juang, “A vector quantization approach to speaker recognition,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 10, pp. 387-390, 1985.

[26] C. Che, Q. Lin, and D. Yuk, “An hmm approach to text-prompted speaker verification,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 673-676, 1996.

[27] A. Lodi, M. Toma, and R. Guerrieri, “Very low complexity prompted speaker verification system based on hmm-modeling,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, pp. 3912-3915, 2002.

[28] J. Oglesby and J. Mason, “Radial basis function networks for speaker recognition,” in *Proc. International Conference on Acoustics, Speech, and Signal Processing*, pp. 393–396, 1991.

[29] D. Mashao and M. Skosan, “Combining classifier decisions for robust speaker identification,” *Pattern recognition*, vol. 39, no. 1, pp. 147-155, 2006.

[30] W. Campbell, J. Campbell, D. Reynolds, D. Jones, and T. Leek, “Phonetic speaker recognition with support vector machines,” *Advances in Neural Information Processing Systems*, vol. 16, pp. 1377-1384, 2004.

[31] D. Reynolds and R. Rose, “Robust text-independent speaker identification using Gaussian mixture speaker models,” *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 72-83, January 1995.

[32] J. Bilmes, “A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models,” *tech. rep., International Computer Science Institute*, 1998.

[33] A. Dempster, N. Laird, D. Rubin, et al., “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1-38, 1977.

[34] J. Pelecanos, S. Myers, S. Sridharan, and V. Chandran, “Vector quantization based Gaussian modeling for speaker verification,” in *Proc. International Conference on Pattern Recognition*, vol. 15, pp. 294-297, 2000.

[35] R. Duda, P. Hart, and D. Stork, Pattern classification. Wiley New York, 2001.

[36] J. Gauvain and C. Lee, “Bayesian adaptive learning and MAP estimation of HMM,” *Advanced Topics in Automatic Speech and Speaker Recognition*, pp. 83-107, 1996.

[37] National Institute of Standards and Technology, “Speaker recognition workshop,” June 1999.

[38] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, “Joint factor analysis versus eigenchannels in speaker recognition,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435-1447, 2007.

[39] R. Vogt and S. Sridharan, “Experiments in session variability modeling for speaker verification,” *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. 897-900, May 2006.

[40] B. Baker, R. Vogt, M. Mason, and S. Sridharan, “Improved phonetic and lexical speaker recognition through MAP adaptation,” *in ODYSSEY04-The Speaker and Language Recognition Workshop*, pp. 91-96, 2004.

[41] P. Kenny and P. Dumouchel, “Disentangling speaker and channel effects in speaker verification,” *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 37-40, 2004.

[42] R. Auckenthaler, M. Carey, and H.L. Thomas, “Score normalization for text independent speaker verification system,” *Digital Signal Processing*, vol.10, no.1-3,

pp.42-54, Jan. 2000.

[43] National Institute of Standards and Technology, “The NIST year 2005 speaker recognition evaluation plan,” 2005.

[44] National Institute of Standards and Technology, “The NIST year 2006 speaker recognition evaluation plan,” 2006.

[45] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, “Factor analysis simplified,” *in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 637-640, 2005.

[46] D. Reynolds, “HTIMIT and LLHDB: speech corpora for the study of hand- set transducer effects,” *in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 1535-1538, 1997.

[47] D. Reynolds, T. Quatieri, and R. Dunn, “Speaker verification using adapted Gaussian mixture models,” *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19-41, 2000

[48] V. Wan and S. Renals, “Speaker verification using sequence discriminant support vector machines,” *IEEE Transactions on Speech and Audio Process ing*, vol. 13, pp. 203-210, March 2005.

[49] A. Rosenberg, J. DeLong, C. Lee, B. Juang, and F. Soong, “The use of cohort normalized scores for speaker verification,” in *Second International Conference on Spoken Language Processing*, pp. 599-602, 1992.

[50] A. Higgins and L. Bahler, “Text-independent speaker verification by discriminator counting,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 405-408, 1991.

[51] D. Reynolds, “Speaker identification and verification using Gaussian mixture speaker models,” *Speech Communication*, vol. 17, pp.91-108, August 1995.

[52] T. Isobe and J. Takahashi, “A new cohort normalization using local acoustic information for speaker verification,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 841-844, 1999.

[53] R. Teunen, B. Shahshahani, and L. Heck, “A model-based transformational approach to robust speaker recognition,” in *Proc. International Conference on Spoken Language Processing*, vol. 2, pp. 495-498, 2000.

[54] National Institute of Standards and Technology, “NIST speech group website,” 2009. http://www.itl.nist.gov/iad/mig/tests/sre/index.html.

[55] A. Hatch, S. Kajarekar, and A. Stolcke, “Within-class covariance normalization for SVM-based speaker recognition,” in *Ninth International Conference on Spoken Language Processing*, pp. 1471-1474, 2006.

[56] A. Solomonoff, C. Quillen, and W. Campbell, “Channel compensation for SVM speaker recognition,” in *Proc. Odyssey*, pp. 57-62, 2004.

[57] R. Vogt, S. Kajarekar, and S. Sridharan, “Discriminant NAP for SVM speaker recognition,” in *Proc. IEEE Odyssey Workshop*, pp. 629-632, 2008.

[58] W. Campbell, D. Sturim, D. Reynolds, and A. Solomonoff, “SVM based speaker verification using a GMM supervector kernel and NAP variability compensation,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. 97-100, May 2006.

[59] D. Matrouf, N. Scheffer, B. Fauve, and J. F. Bonastre, “A straightforward and efficient implementation of the factor analysis model for speaker verification,” in

Interspeech, pp. 1242-1245, 2007.

[60] C. Burges, “A tutorial on support vector machines for pattern recognition,” *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121-167, 1998.

[61] S. Lucey and T. Chen, “Improved speaker verification through probabilistic subspace adaptation,” in *Eighth European Conference on Speech Communication and Technology*, pp. 2021-2024, 2003.

[62] R. Kuhn, J. Junqua, P. Nguyen, and N. Niedzielski, “Rapid speaker adaptation in Eigenvoice space,” *IEEE Trans.Audio, Speech, Lang. Process.*, vol.8 no.6, pp. 695-707, Nov. 2000.

[63] P. Kenny, M. Mihoubi, and P. Dumouchel, “New MAP estimators for speaker recognition,” in *Proc. Eurospeech*, Geneva, Switzerland, Sept. 2003.

[64] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, “A study of interspeaker variability in speaker verification,” *IEEE Trans. Audio,Speech, Lang. Process.*, vol. 16, no. 5, pp. 980-988, Jul. 2008.

[65] A.Solomonoff, W.Campbell, and I.Boardman, “Advances in channel compensation for SVM speaker recognition,” in *Proc. ICASSP*, vol. 1, 2005, pp.629-632.

[66] N. Dehak, P.Kenny, R. Dehak, P.Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” *IEEE Trans.Audio, Speech, Lang. Process.*, vol9.8 no.4, pp. 788-798, May. 2010.

[67] Anthony Larcher, Pierre-Michel Bousquet, Kong-Aik Lee, Driss Matrouf, Haizhou Li, Jean-Fran çois Bonastre, “I-vectors in the context of phonetically-constrained short utterances for speaker verification,” in *ICASSP 2012* pp.4773-4776

[68] I. Magrin-Chagnoleau, J. Bonastre and F. Bimbot, “Effect of utterance duration and phonetic content on speaker identification using second-order statistical methods,” in *Proc. Eurospeech*, vol. 1, pp. 337-340, Madrid, Spain, Sept. 1995.

[69] Matsui, T., Furui,S., “Concatenated Phoneme Models for Text-Variable Speaker Recognition,”, in *Proc. ICASSP 1993*, Minneapoils MN.

[70] A. Mohamed, F. Ren and S.Kuroiwa, “Effects of Phoneme Type and Frequency on Distributed Speaker Identification and Verification,” *IEICE Trans. INF. & SYST.*, vol. E89-D, no. 5, pp. 1712-1719, May. 2006.

[71] Gutman, D., Bistritz Y., “Speaker Verification Using Phoneme-Adapted Gaussian Mixture Models,” in *Proc. EUSIPCO 2002*, Toulouse France.

[72] Wenbin Zhang, Haoze Lu, Yasuo Horiuchi, Satoru Tsuge, Kenji Kita, Shingo Kuroiwa, “Text-Independent Speaker Identification Based on Reducing Inter-Session Variability of Speech Feature Using PCA Transformation,” *2011 International Workshop on Nonlinear Circuits, Communication and Signal Processing*, Tianjin, China, pp.421-424, Mar. 2011.

[73] W. Campbell, D. Reynolds, and J. Campbell, “Fusing discriminative and generative methods for speaker recognition: Experiments on Switchboard and NFI/TNO field data,” *in Odyssey: The Speaker and Language Recognition Workshop*, pp. 41-44, 2004.

[74] Tsuge, S., Shishibori, M., Kita, K., Ren, F. and Kuroiwa, S, “Study of intra-speaker’s speech variability over long and short time periods for speech recognition”, *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2006),* Toulouse, France, pp.397-400, 2006.

[75] B.A.Dautrich, L.R.Rabiner, and T.B.Martin, “On the effects of varying filter bank

parameters on isolated word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. 31, pp. 793-897, 1983.

[76] D.A.Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication*, Vol. 17, pp. 91-108, 1995.

[77] M.Nishida and Y.Ariki, "Speaker recognition by separating phonetic space and speaker space," *EUROSPEECH-2001*, pp. 1381-1384, 2001.

[78] H.Lu, M.Nishida, Y.Horiuchi and S.Kuroiwa, "Text-independent speaker identification in phoneme-independent subspace using PCA transformation," *International Journal of Biometrics*, Vol.2, pp. 379-390, 2010.

[79] Stolcke, A. S. Kajarekar, L. Ferrer, and E. Shriberg, "Speaker recognition with session variability normalization based on MLLR adaptation transforms", *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 15, pp. 1987-1998, Sep. 2007.

[80] S. Hyunson, C. S. Jung, and H.G. Kang, "Robust Session Variability Compensation for SVM Speaker Verification," *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 19, NO. 6, 2011, pp. 1631-1641.

[81] S.Tsuge, M.Fukumi and S.Kuroiwa, "Specific speakers' speech corpus over long and short time periods," *Oriental COCOSDA 2008*, Kyoto, Nov. 2008.

[82] S.Young, and etc, The HTK Book, Cambridge University Engineering Deparment, 2005.

[83] S. Furui, "Recent advances in speaker recognition," *Pattern Recognition Letters*, vol. 18, no. 9, pp. 859-872, 1997.

[84] R. A. Cole and colleagues, "Survey of the State of the Art in Human Language

Technology," *National Science Foundation European Commission*, http://cslu.cse.ogi.edu/HLTsurvey/ch1node47.html, 1996.

[85] J. P. Cambell, JR. (1997) 'Speaker Recognition: A Tutorial', *Proceeding of the IEEE*, Vol. 85, No. 9, pp. 1436-1462.

[86] D. A. Reynolds, "An overview of automatic speaker recognition technology," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2002)*, Orlando, FL, pp. 4072-4075, 2002.

[87] A. Martin and M. Przybocki, "Speaker Recognition in a multi-speaker environment," in *Proc. 7th Eur. Conf. Speech Communication and Technology (Eurospeech 2001)*, Aalborg. Denmark, PP. 787-790, 2001.

[88] I. Lapidot, H. Guteman, and A. Cohen, "Unsupervised speaker recognition based on competition between self-organizing maps," *IEEE Trans. Neural Networks*, vol. 13, pp. 877-887, 2002.

[89] D. Liu and F. Kubala, "Fast speaker change detection for broadcast news transcription and indexing," in *Proc. 6th European Conf. Speech Communication and Technology (Eurospeech 1999)*, Budapest, Hungary, pp. 1031-1034, 1999.

[90] S. Kwon and S. Narayanan, "Speaker change detection using a new weighted distance measure," in *Proc. Int. Conf. on Spoken Language Processing (ICSLP 2002)*, Denver, CO, pp. 2537-2540, 2002.

[91] S.Furui, "An overview of speaker recognition technology," *Workshop on automatic speaker recognition, identification and verification*, Martigny, Switzerland, pp.289-292, 1994.

[92] C.B.Lima, A.Alcain, and J.A.Apolinario Jr, "On the Use of PCA in GMM and

AR-Vectors Models for Text Independent Speaker Verification,” *DSP-2002*, Vol. 2, pp. 595-598, 2002

[93] C. Seo and K. Y. Lee, “GMM based on local PCA for speaker identification,” *Electronics Letters*, Vol. 37, pp. 1486-1488, 2002.

[94] D.A. Reynolds, “Speaker Identification and Verification using Gaussian Mixture Speaker Models,” *Speech Communication*, Vol. 17, pp. 91-108, 1995.

[95] B. A. Dautrich, L.R. Rabiner and T. B. Martin., “on the effects of varying filter bank parameters on isolated word recognition,” *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. 31, pp. 793-897, 1983.

[96] M. Nishida and Y. Ariki, “Speaker Verification by Intergrating Dynamic and Static Features using Subspace Method,” in *Proc. Int. Conf. on Spoken Language Processing (ICSLP 2000)*, Vol 3, Beijing, China, pp.1013-1016, 2000.

[97] M. Nishida and Y. Ariki, “Speaker recognition by separating phonetic space and speaker space,” in *Proc. 7th Eur. Conf. Speech Communication and Technology (Eurospeech 2001)*, Aalborg. Denmark, pp. 1381-1384, 2001

[98] W. M. Campbell, “Generalized linear discriminant sequence kernels for speaker recognition,” in *Proceedings of the International Conference on Acoustics Speech and Signal Processing*, 2002, pp. 161-164.

[99] Nello Cristianini and John Shawe-Taylor, “Support Vector Machines,” *Cambridge University Press*, Cambridge, 2000.

[100] I. Guyon, V. Vapnik, B. Boser, L. Bottou, and S. Solla, “Capacity control in linear classifiers for pattern recognition,” in *Pattern Recognition, 1992. Vol. II. Conference B: Pattern Recognition Methodology and Systems, Proceedings, 11th*

IAPR International Conference, pp. 385-388, 1992.

[101] G. Dupuy, M. Rouvier, S. Meignier, and Y. Estève, “I-vectors and ILP clustering adapted to cross-show speaker diarization,” in *Proc. Interspeech* 2012.

[102] E. Newham, “The Biometric Report,” http://www.sjb.com/: SJB Services, New York, 1995.

[103] T. Nguyen et al., “The IIR-NTU speaker diarization systems for RT 2009,” in *Proc. RT’09, NIST Rich Transcription Workshop, Melbourne, FL*, 2009.

[104] X. Anguera, C. Wooters, and J. Hernando, “Friends and enemies: A novel initialization for speaker diarization,” in Proc. ICSLP, Pittsburgh, PA, 2006

[105] S. Araki, M. Fujimoto, K. Ishizuka, H. Sawada, and S. Makino, "A DOA based speaker diarization system for real meetings," *HSCMA2008*, pp.29-32, 2008

[106] Pavel Campr, “Audio-Video Speaker Diarization for Unsupervised Speaker and Face Model Creation,” *Text, Speech and Dialogue Lecture Notes in Computer Science Volume 8655, 2014*, pp 465-472

[107] Yu HORII, Hiroaki KAWASHIMA, and Takashi MATSUYAMA, “Speaker Detection Using the Timing Structure between Lip Motion and Speech Signal,” *MIRU2008*, pp.193-200, 2008.7.30.

[108] K. Boakye, O. Vinyals, G. Friedland, “Improved Overlapped Speech Handling for Speaker Diarization,” in *Proc. Interspeech*, 2011.

[109] K. Boakye, B. Trueba-Hornero, O. Vinyals, and G.Friedland, “Overlapped speech detection for improved speaker diarization in multiparty meetings,” in *Proc. ICASSP*, 2008, pp. 4353-4356.

[110] B. Trueba-Hornero, “Handling overlapped speech in speaker diarization” *M.S. thesis, Univ. Politecnica de Catalunya, Barcelona, Spain,* 2008.

[111] S. H. Yella, F. Valente, “Speaker Diarization of Overlapping Speech based on Silence Distribution in Meeting Recordings,” in *Proc. Interspeech, 2012*.

[112] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, “Support Vector Machines Using GMM Supervectors for Speaker Verification,” *IEEE SIGNAL PROCESSING LETTERS,* Vol. 13, No. 5, May 2006.

[113] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum Likelihood from Incomplete Data via the EM algorithm,” *Journal of the Royal Statistical Society，Series B (Methodological),* vol.39， no.1， pp.1-38， 1977.

[114] J. L. Gauvain, and C. H. Lee, “Maximum a Posteriori Estimation for Multivariate Gaussian Overlappture Observations of Markov Chains,” *IEEE Transactions on Speech and Audio Processing,* vol.2, no.2, pp.291-298, Apr.1994.

List of related papers by the author

Journal Papers

[1] Haoze Lu, Masafumi Nishida, Yasuo Horiuchi, Shingo Kuroiwa, “Text-Independent speaker identification in phoneme-independent subspace using PCA transformation,” *International Journal of Biometrics, Vol.2, No.4,* pp.379-390, Sep. 2010.

[2] Haoze Lu, Wenbin Zhang, Yasuo Horiuchi, Takahiro Shinozaki and Shingo Kuroiwa, “PCA Transformation Based Inter-session Variability Suppression for Text-Independent Speaker Identification,” *International Journal of Advanced Intelligence Vol.5, No.1,* pp.120-129, July. 2013.

[3] Haoze Lu, Wenbin Zhang, Yasuo Horiuchi, Shingo Kuroiwa, “Phoneme dependent inter-session variability reduction for speaker verification,” *International Journal of Biometrics, Vol.7, No.2,* pp.83-96, 2015.

[4] Haoze Lu, Yuma Akaiwa, Yasuo Horiuchi, Shingo Kuroiwa, “Robust Extraction of Desired Speaker’s Utterance in Overlapped Speech, ” *IEEJ Transactions on Electronics, Information and Systems, Vol.135, No.8,* pp.1009-1016

[5] Wenbin Zhang, Haoze Lu, Yasuo Horiuchi, Satoru Tsuge, Kenji Kita, Shingo Kuroiwa, “Text-Independent Speaker Identification Based on Reducing Inter-Session Variability of Speech Feature Using PCA Transformation,” *Journal of Signal Processing, Vol.15, No.4,* pp.275-278, July 2011.

International Conferences

[1] Haoze Lu, Haruka Okamoto, Masafumi Nishida, Yasuo Horiuchi, and Shingo Kuroiwa, “Text-Independent Speaker Identification Based on Feature Transformation to Phoneme-Independent Subspace,” *Proceedings of the 11th IEEE International Conference on Communication Technologies (ICCT)*, pp. 692-695, Nov. 2008.

[2] Haoze Lu,Wenbin Zhang,Yasuo Horiuchi, Takahiro Shinozaki and Shingo Kuroiwa, “PCA Transformation Based Inter-session Variability Suppression for Text-Independent Speaker Identification,” *Proceedings of the 8th Conference on Natural Language Processing and Knowledge Engineering (NLPKE2012),* Hefei, China, pp.463-473, Sep. 2012.

[3] Wenbin Zhang, Haoze Lu, Yasuo Horiuchi, Satoru Tsuge, Kenji Kita, Shingo Kuroiwa, “Text-Independent Speaker Identification Based on Reducing Inter-Session Variability of Speech Feature Using PCA Transformation,” *2011 International Workshop on Nonlinear Circuits, Communication and Signal Processing*, Tianjin, China, pp.421-424, Mar. 2011.

Japanese Domestic Contests

[1] Haoze Lu, Masafumi Nishida, Yasuo Horiuchi, Shingo Kuroiwa, “Speaker Recognition Based on Phoneme Suppressed Feature Transformation,” Acoustical Society of Japan, Spring Meeting, 3-Q-1, pp.199-200, Mar 2009.

[2] Wenbin Zhang, Haoze Lu, Takahiro Shinozaki, Yasuo Horiuchi and Shingo Kuroiwa, “Robust Speaker Identification Method Based on Reducing Inter-session Variability,” The First Symposium on Biometrics, Recognition and Authentication, pp.88-91, Nov. 2011.