Speaker Verification and Identification

Jin, Minho, and Yoo, Chang D.
Korea Advanced Institute of Science and Technology, Republic of Korea

ABSTRACT

Speaker recognition system verifies or identifies a speaker's identity based on his/her voice considered as one of the most convenient biometric characteristic for human machine communication. This chapter introduces several speaker recognition systems and examines their performances under various conditions. Speaker recognition can be classified into either speaker verification or speaker identification. Speaker verification aims to verify whether an input speech corresponds to a claimed identity, and speaker identification aims to identify an input speech by selecting one model from a set of enrolled speaker models. Both the speaker verification and identification system consist of three essential elements: feature extraction, speaker modeling, and matching. The feature extraction pertains to extracting essential features from an input speech for speaker recognition. The speaker modeling pertains to probabilistically modeling the feature of the enrolled speakers. The matching pertains to matching the input feature to various speaker models. Speaker modeling techniques including Gaussian mixture model (GMM), hidden Markov model (HMM), and phone n-grams are presented, and in this chapter, their performances are compared under various tasks. Several verification and identification experimental results presented in this chapter indicate that speaker recognition performances are highly dependent on the acoustical environment. A comparative study between human listeners and an automatic speaker verification system is presented, and it indicates that an automatic speaker verification system can outperform human listeners. The applications of speaker recognition are summarized, and finally various obstacles that must be overcome are discussed.

KEYWORDS

Speaker verification, Speaker identification, Speech recognition, Voice recognition devices

INTRODUCTION

Speaker recognition can be classified into either 1) speaker verification or 2) speaker identification (Furui, 1997; J. Campbell, 1997; Bimbot et al., 2004). Speaker verification aims to verify whether an input speech corresponds to the claimed identity. Speaker identification aims to identify an input speech by selecting one model from a set of enrolled speaker models: in some cases, speaker verification will follow speaker identification in order to validate the identification result (Park & Hazen, 2002). Speaker verification is one case of biometric authentication, where users provide their biometric characteristics as passwords. Biometric characteristics can be obtained from deoxyribonucleic acid (DNA), face shape, ear shape, fingerprint, gait pattern, hand-vein pattern, hand-and-finger geometry, iris scan, retinal scan, signature, voice, etc. These are often compared under the following criteria (Jain, Ross, & Prabhakar, 2004):

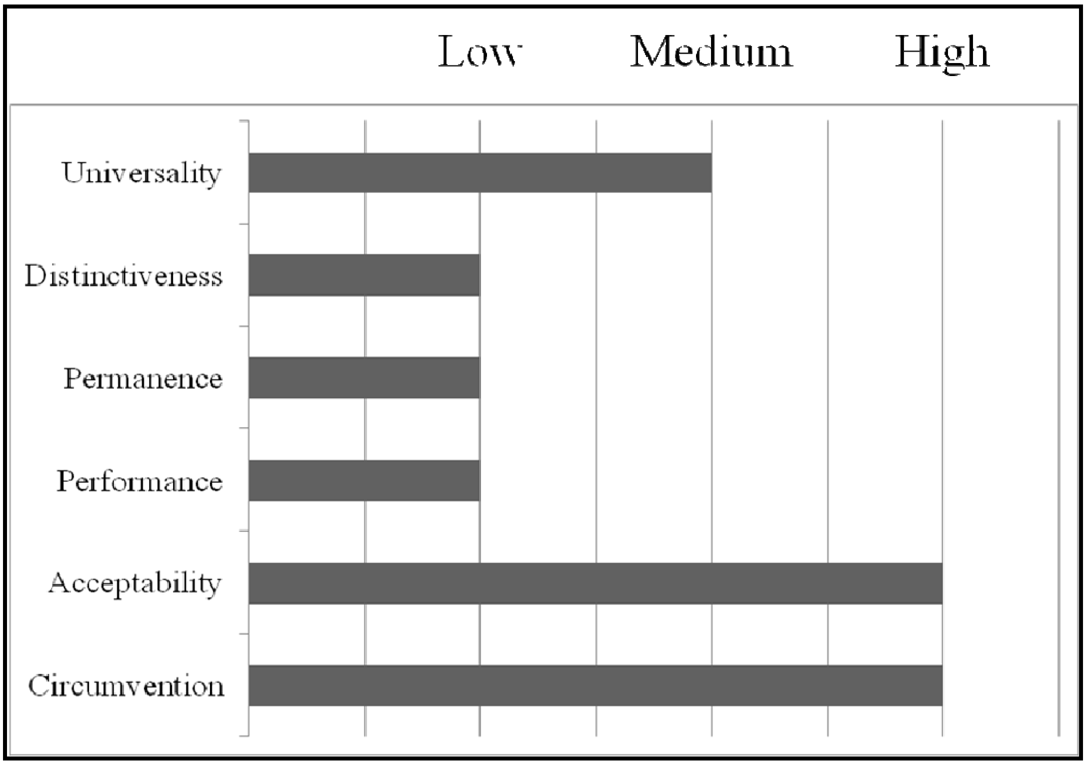

Figure 1 Properties of voice: voice can be universally available for every person, and its authentication result is acceptable. However, its performance in terms of accuracy is known to be slightly inferior to that of other biometric characteristics

- Universality: the biometric characteristic should be universally available to everyone.
- Distinctiveness: the biometric characteristics of different people should be distinctive.
- Permanence: the biometric characteristic should be invariant over a period of time that depends on the applications
- Performance: the biometric authentication system based on the biometric characteristic should be accurate, and its computational cost should be small.
- Acceptability: the result of a biometric authentication system based on certain biometric characteristic should be accepted to all users.

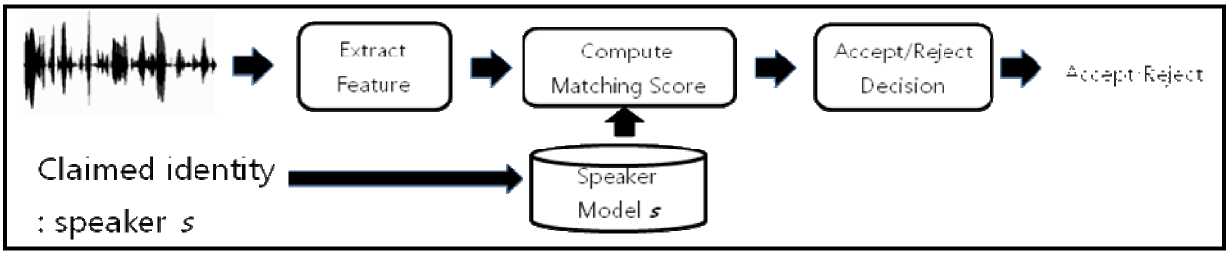

Figure 2 Conventional speaker verification system: the system extracts features from recorded voice, and it computes its matching score given the claimed speaker's model. Finally, an accept/reject decision is made based on the matching score.

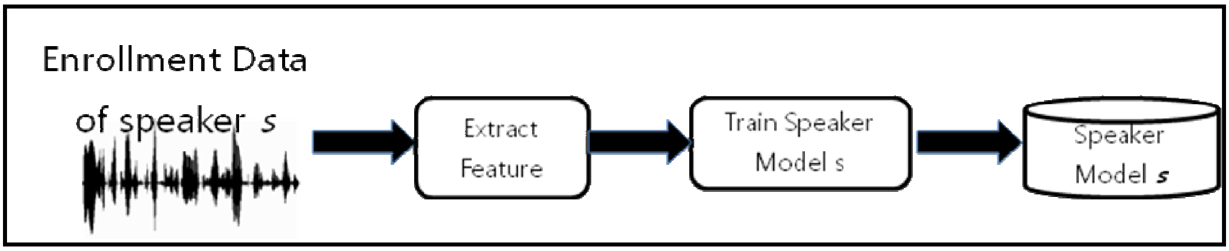

Figure 3 Enrollment of a target speaker model; each speaker's model is enrolled by training his/her model from features extracted from his/her speech data.

One additional criterion that should be included is circumvention which is given by

- Circumvention: biometric characteristics that are vulnerable to malicious attacks are leading to low circumvention.

High biometric characteristic scores on all above criteria except circumvention are preferable in real applications. As shown in Figure 1, voice is reported to have medium universality. However, in many cases, voice is the only biometric characteristic available: for example, when a person is talking over the phone. The distinctiveness of voice is considered low, and very often a speaker verification system can be fooled by an impostor mimicking the voice of an enrolled. For this, many features such as prosodic and idiosyncratic features have been incorporated to improve the speaker recognition system. The permanence of voice is low since a speaker's voice can vary under various situations, physical conditions, etc. By incorporating on-line speaker adaptation techniques that adapt a speaker's voice change on-line, the permanence of voice can be improved. We discuss the performance of a speaker recognition system in the latter part of this chapter.

BACKGROUND

Common Tasks of Speaker Recognition

Speaker recognition system verifies or identifies a speaker's identity based on his/her speech. The length of speech data varies according to the application. For example, if an application is designed to identify a speaker when he/she tells his/her name, the input speech data will be 1-2 second long. The speech data is recorded using microphones in various environments. The types of microphone, environments (e.g., telephone communication, clean sound booth) can affect the speaker recognition performance. For this reason, several benchmark databases are collected to evaluate the performance of speaker recognition algorithms in various conditions: some examples of benchmark databases will be described in the latter of this chapter.

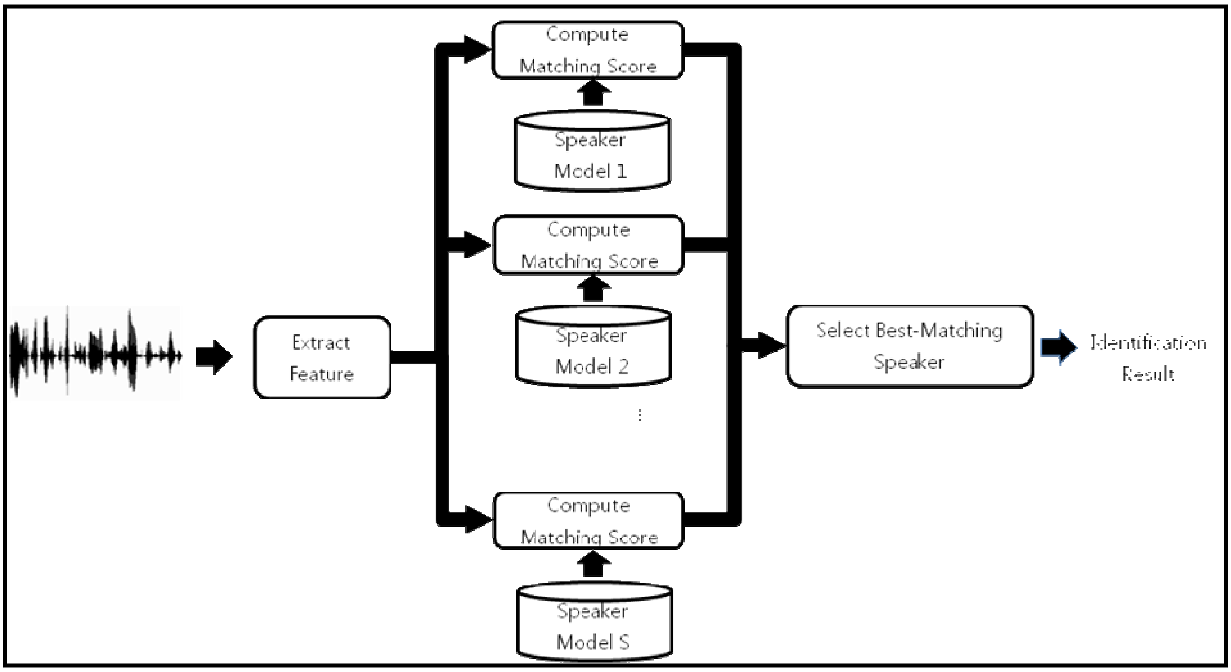

Figure 4 Conventional speaker identification system: the speaker identification system selects the best-matched speaker's model among enrolled speaker models.

Speaker Verification

Figure 2 illustrates a conventional speaker verification system. A speaker verification system takes the speech of an unknown speaker with his/her claimed identity, and it determines whether the claimed identity matches the speech. The claimed identity can be fed into the system using various channels such as keyboard, identity card, etc. To verify whether the input speech matches the claimed identity, the claimed speaker's model must be enrolled beforehand as shown in Figure 3. As the amount of enrollment data increases, the performance of speaker verification system usually improves, but the speaker may feel uncomfortable with long enrollment process. Thus, in many applications, the enrollment is performed by adapting a speaker-independent model into an enrolled speaker's model using speaker adaptation techniques.

Speaker Identification

Figure 4 illustrates a conventional speaker identification system. A speaker identification system only takes the speech of an unknown speaker, and it determines which enrolled speaker best matches the speech. The speaker identification system finds the best matching speaker among the enrolled speakers, and it may be that the unknown speaker is not enrolled. For this reason, in many systems, speaker identification is followed by speaker verification. For example, the MIT-LCS ASR-based speaker recognition system first performs speaker identification and then performs speaker verification where the identification result is used as a claimed speaker identity (Hazen, Jones, Park, Kukolich, & Reynolds, 2003).

Operation Modes of Speaker Recognition

The speaker recognition system can operate in either a text-dependent (TD) mode or a text-independent (TI) mode. In the TD mode, the user speaks a pre-defined or prompted text transcription. In the TI mode, the user is allowed to speak freely. Since the TD mode provides the speaker recognition system with extra information, thus it generally performs better than in the TI mode. Various studies have been performed in order to reduce the performance gap between the two operation modes (Park & Hazen, 2002; Che, Lin, & Yuk, 1996; Newman, Gillick, Ito, Mcallaster, & Peskin, 1996; Weber, Peskin, Newman, Emmanuel, & Gillick, 2000).

Text-Dependent Speaker Recognition

In a TD mode, a speaker recognition system can be fooled by recording and play-back of pre-defined speech of an enrolled speaker. To defend a speaker recognition system from such malicious attack, the system can request the user to utter a randomly prompted text. In most cases, a TD speaker recognition system outperforms a TI speaker recognition system since additional information (text transcription) is provided (Reynolds, 2002a). However, a TD speaker recognition system cannot be used when the true-underlying transcription is not provided, as in the situation when some is talking freely over the phone.

Text-Independent Speaker Recognition

A TI speaker recognition system does not require true-underlying transcription of an input speech. This may be useful for a forensic speaker recognition system such as identifying a speaker of a wiretapped conversation or in human-robot interface.

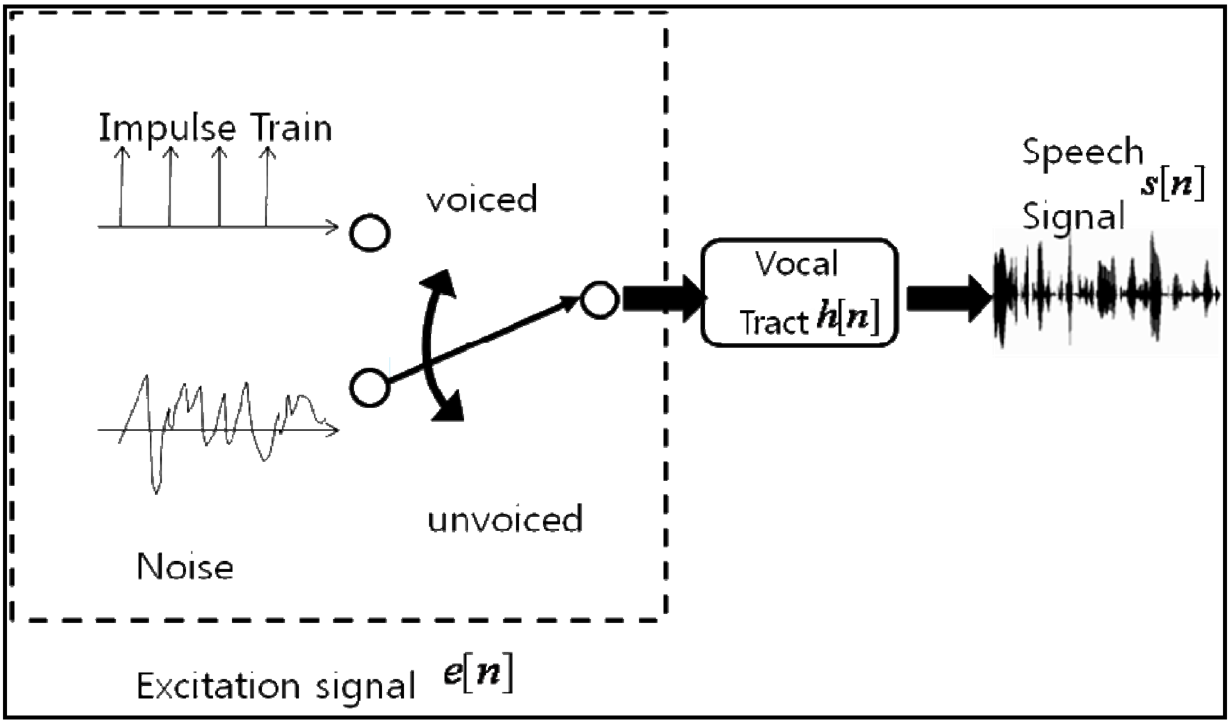

Figure 5 Vocal tract modeling in linear predictive coefficients: the speech signal is modeled as a filtered output of a vocal chords excitation signal, where the filter is determined by the shape of his/her vocal tract.

SPEAKER RECOGNITION SYSTEM

Both the speaker verification and identification system extract features from an input speech then computes matching score between enrolled speaker and the input speech. This section outlines the features and speaker modeling techniques that are used for speaker recognition.

Feature Extraction

Linear Predictive Coefficients (LPC)

The speech signal can be modeled as a filtered output of an excitation signal as shown in Figure 5. When computing P-order linear predictive coefficients (LPC), the human vocal tract is modeled as an all-pole filter as follows:

$$H(z)=\frac{S(z)}{E(z)}=\frac{1}{1-\sum_{p=1}^{P}a_k z^{-k}}, \qquad (0.1)$$

where S(z) and E(z) are the z-transforms of the speech signal $s[n]$ and its excitation signal $e[n]$, respectively. For each frame of an input speech, the LPC parameters $\{a_k\}$ are computed and used for speaker recognition. Using the all-pole model, a speech signal $s[n]$ can be written as follows:

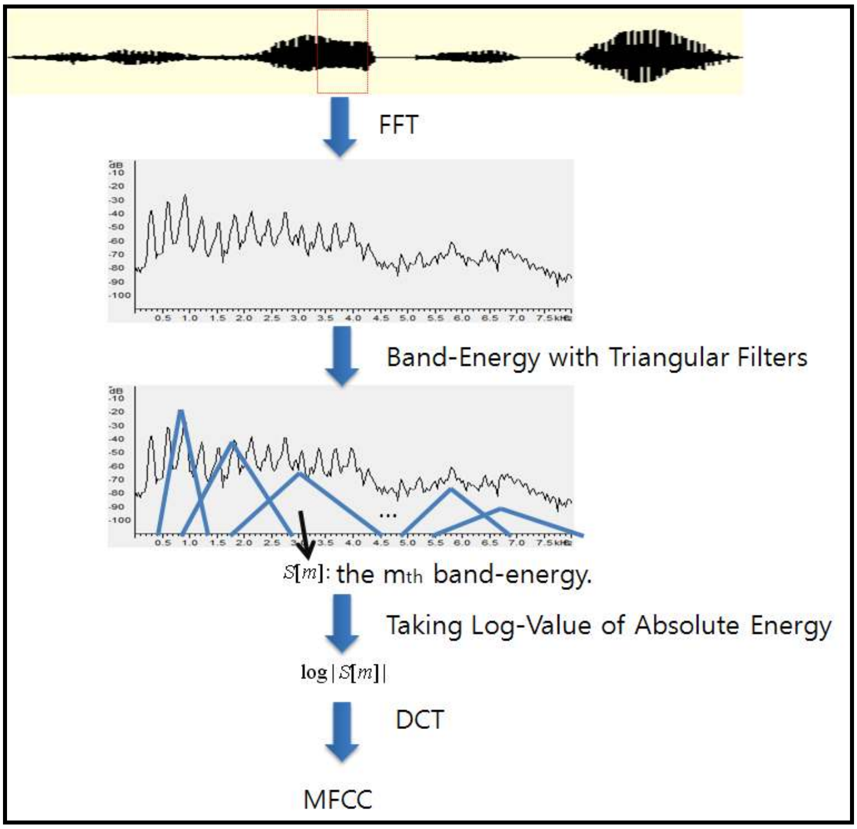
FFT
Band-Energy with Triangular Filters
$S[m]$: the mth band-energy.
Taking Log-Value of Absolute Energy
$\log|S[m]|$
DCT
MFCC

Figure 6 Extraction of MFCCs: MFCCs are extracted based on the band-energy with triangular filters.

$$s[n]=\sum_{k=1}^{P}a_k s[n-k]+e[n]. \qquad (0.2)$$

Let $\hat{s}[n]=\sum_{k=1}^{P}a_k s[n-k]$ be the estimate of $s[n]$. Then, the mean squared error (MSE) of the estimate is given as follows:

$$\begin{aligned} E &= \sum_n (s[n]-\hat{s}[n])^2 \\ &= \sum_n (s[n]-\sum_{k=1}^{P}a_k s[n-k])^2. \end{aligned} \qquad (0.3)$$

The MSE is convex function of a_k , and the minimum of (0.3) is achieved with a_k satisfying the following condition:

$$\sum_{k=1}^{P}a_k\sum_n s_{n-k}s_{n-i}=\sum_n s_n s_{n-i} \qquad (0.4)$$

for $i=1,2,...,P$. Equation (0.4) can be computed using covariance and autocorrelation methods (Huang, Acero, & Hon, 2001; Rabiner & Schafer, 1978)

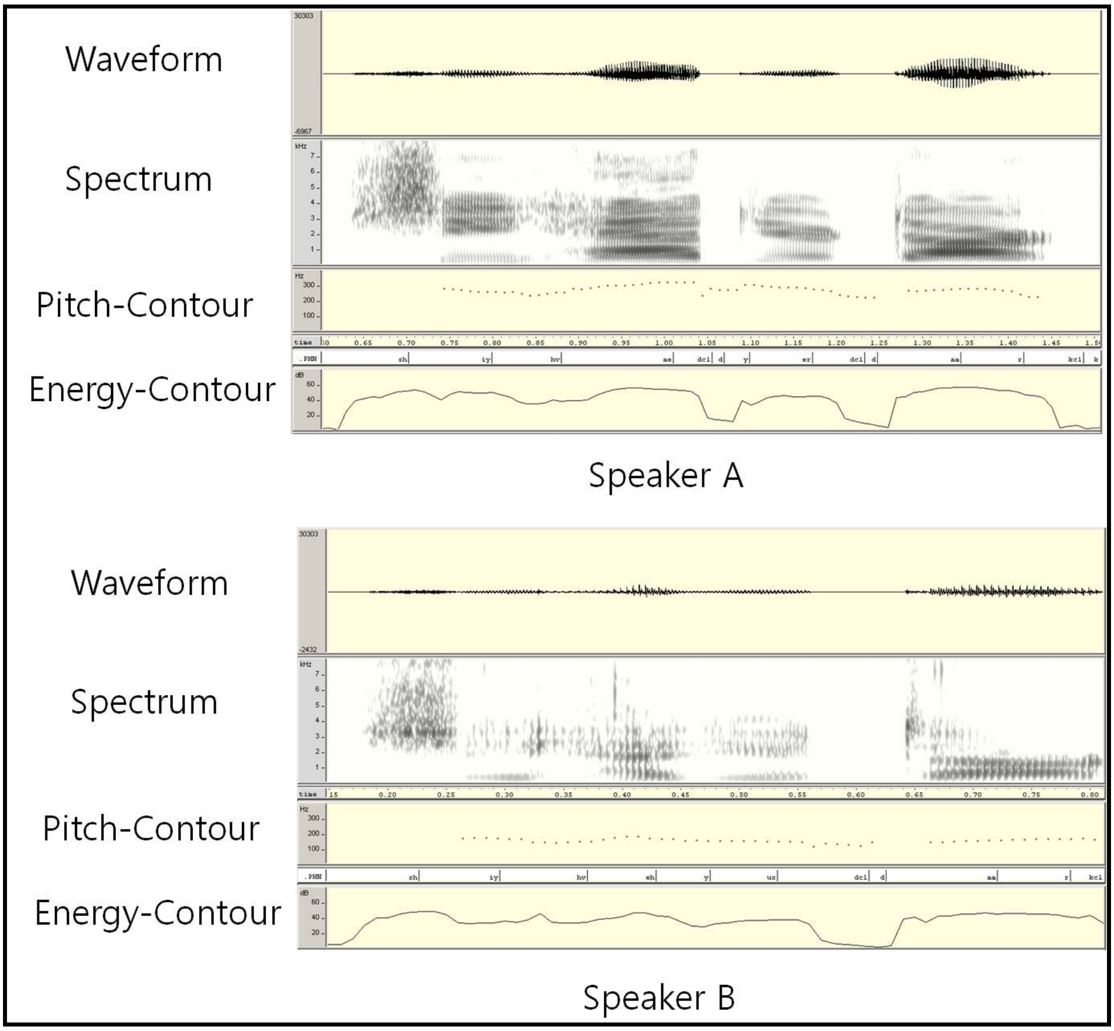

Figure 7 Waveform, spectrum, pitch-contour and energy-contour of the same sentence from different speakers: patterns from pitch or energy contour changes slowly compared to patterns from spectrum.

Mel-frequency Cepstrum Coefficients (MFCC)

Empirical studies have shown that the human auditory system resolves frequencies non-linearly, and the non-linear resolution can be approximated using the Mel-scale which is given by

$$M(f) = 1127.01048 \bullet \log_e f \qquad (0.5)$$

where f is a frequency (Volkman, Stevens, & Newman, 1937). This indicates that the human auditory system is more sensitive to frequency difference in lower frequency band than in higher frequency band. Figure 6 illustrates the process of extracting Mel-frequency cepstrum coefficients (MFCCs) with triangular filters that are equally-spaced in Mel-scale. An input speech is transformed using discrete Fourier transform, and the filter-bank energies are computed using triangular filters. The log-values of the filter-bank energies are transformed using discrete cosine transform (DCT). Finally, the M-dimensional MFCCs are extracted by taking M-DCT coefficients. Previous research has reported that using MFCCs is beneficial for both speaker and speech recognition (Davis & Mermelstein, 1980), and MFCCs are used in many speech and speaker recognition systems (Choi & Lee, 2002; Davis & Mermelstein, 1980; Hirsch &

Pearce, 2000; Li, Chang, & Dai, 2002; Milner, 2002; Molla & Hirose, 2004; Pan & Waibel, 2000; Reynolds, Quatieri, & Dunn, 2000; Xiang & Berger, 2003; Zilca, 2002)

Prosodic Features

Prosodic features include pitch and its dynamic variations, inter-pause statistics, phone duration, etc. (Shriberg, 2007) Very often, prosodic features are extracted with larger frame size than acoustical features since prosodic features exist over a long speech segment such as syllables. Figure 7 illustrates the waveform, spectrum, pitch-contour and energy-contour from speaker A and B. Two speakers' prosodic features (pitch and energy-contours) are different even though two speakers are uttering the same sentence "she had your dark …". The pitch and energy-contours change slowly compared to the spectrum, which implies that the variation can be captured over a long speech segment. Many literatures reported that prosodic features usually do not outperform acoustical features but incorporating prosodic features in addition to acoustical features can improve speaker recognition performance (Shriberg, 2007; Sönmez, Shriberg, Heck, & Weintraub, 1998; Peskin et al., 2003; Campbell, Reynolds, & Dunn, 2003; Reynolds et al., 2003).

Idiolectal Features

The idiolectal feature is motivated by the fact that people usually use idiolectal information to recognize speakers. In telephone conversation corpus, Doddington (2001) reported enrolled speakers can be verified not using acoustical features that are extracted from a speech signal but using idiolectal feature that are observed in true-underlying transcription of speech. The phonetic speaker verification, motivated by Doddington (2001)'s work, creates a speaker using his/her phone n-gram probabilities that are obtained using multiple-language speech recognizers (Andrews, Kohler, & Campbell, 2001).

Speaker Model and Matching Score

Gaussian Mixture Model (GMM)

The speaker model can be trained as a Gaussian mixture model (GMM) with the expectation-maximization (EM) algorithm. When recognizing an input speech, the system extracts from a T-length input speech, and the matching score is computed as follows:

$$S_{GMM}(\mathbf{x}_1\mathbf{x}_2 \ldots \mathbf{x}_T, \Theta_s) = \prod_{t=1}^{T} p(\mathbf{x}_t \mid \Theta_s), \qquad (0.6)$$

where Θ_s and $\mathbf{x}_t$ are a set of GMM parameters of speaker s and the tth feature vector of D dimension, respectively. The GMM parameter Θ_s includes the mean $\boldsymbol{\mu}_{k,s}$ and variance vectors $\boldsymbol{\Sigma}_{k,s}$ of the k th Gaussian kernel, and its weight $w_{k,s}$. Using this, the likelihood of a GMM can be computed as follows:

$$p(\mathbf{x}_t \mid \Theta_s) = \sum_{k=1}^{K} w_{k,s} \frac{1}{(2\pi)^{D/2} \mid \boldsymbol{\Sigma}_{k,s} \mid^{1/2}} e^{-\frac{(\mathbf{x}_t - \boldsymbol{\mu}_{k,s})^{\mathrm{T}} \boldsymbol{\Sigma}_{k,s}^{-1} (\mathbf{x}_t - \boldsymbol{\mu}_{k,s})}{2}} \qquad (0.7)$$

In the GMM-UBM (Gaussian Mixture Model – Universal Background Model) algorithm, the speaker model is adapted from the UBM which is a GMM trained using many speakers' speech (Reynolds, Quatieri, & Dunn, 2000). Mariethoz and Bengio (2002) compared the performances of GMM-UBM systems with various speaker adaptation techniques, where the maximum a posterior (MAP) adaptation performed the best: for detailed explanation on the MAP adaptation algorithm, please refer to (Gauvain & Lee, 1994). The GMM-UBM is a state of the art algorithm for TI speaker identification and verification. The GMM-UBM performs well with the small amount of enrollment data, say 2 minutes, and its computational cost is relatively small compared to other algorithms using a hidden Markov models (HMMs). The GMM-UBM is intrinsically well-fitted for a TI mode since the GMM is trained without transcription of speech data (Reynolds, et al, 2000). Nevertheless, it does not imply that the GMM cannot be used for a TD mode. For example, text-constrained GMM-UBM system incorporated GMMs trained on different word groups in order to improve the speaker verification performance (Sturim, Reynolds, Dunn, & Quatieri, 2002).

Hidden Markov Model (HMM)

The HMM is usually used to model each speaker's phonetic units or syllabic units. In a TD mode, the score is computed as follows:

$$S_{HMM,TD}(\mathbf{x}_1\mathbf{x}_2...\mathbf{x}_T,\Lambda_s) = p(\mathbf{x}_1\mathbf{x}_2...\mathbf{x}_T,Q \mid \Lambda_s), \qquad (0.8)$$

where Λ_s and Q are the set of HMM parameters that are trained on the enrollment data of speaker s and a true-underlying transcription of $\mathbf{x}_1\mathbf{x}_2...\mathbf{x}_T$, respectively . The HMM can be trained using enrollment data with Baum-Welch algorithm, and the likelihood of $\mathbf{x}_1\mathbf{x}_2...\mathbf{x}_T$ and Q given Λ_s can be computed using forward, backward and Viterbi algorithms: for mathematical details, please refer to (Huang, et al., 2001; Rabiner & Schafer, 1978) for Baum-Welch algorithm and (Huang, et al., 2001; Rabiner & Schafer., 1978; Woodland, Leggetter, Odell, Valtchev, & Young, 1995) for likelihood computation. For a TI mode operation, researchers have used a large-vocabulary continuous speaker recognition (LVCSR) system to generate 1-best transcription Q_{1b}, and the matching score is computed as follows (Hazen, et al., 2003; Newman, Gillick, Ito, Mcallaster, & Peskin, 1996; Weber, et al., 2000):

$$S_{HMM,TI-1b}(\mathbf{x}_1\mathbf{x}_2...\mathbf{x}_T,\Lambda_s) = p(\mathbf{x}_1\mathbf{x}_2...\mathbf{x}_T,Q_{1b} \mid \Lambda_s). \qquad (0.9)$$

Recently, it was reported that incorporating syllable lattice-decoding can be effective for speaker verification when Q_{1b} is highly erroneous (Jin, Soong, & Yoo, 2007). The syllable lattice from the LVCSR was incorporated in the matching score as follows:

$$S_{HMM,TI-L}(\mathbf{x}_1\mathbf{x}_2...\mathbf{x}_T,\Lambda_s) = p(\mathbf{x}_1\mathbf{x}_2...\mathbf{x}_T,L_\mathbf{x} \mid \Lambda_s), \qquad (0.10)$$

where $L_\mathbf{x}$ is the syllable lattice decoded from $\mathbf{x}_1\mathbf{x}_2...\mathbf{x}_T$, where the likelihood is computed using the lattice forward-backward algorithm.

Phone n-grams

Unlike previous modeling techniques that are based on acoustical feature vector sequence $\mathbf{x}_1\mathbf{x}_2...\mathbf{x}_T$, speaker-dependent phone n-gram probabilities can be used for speaker verification (Andrews, Kohler, Campbell, & Godfrey, 2001). Very often, phone n-gram probabilities are computed using phone sequences created by phone recognizers that are trained on several languages. For example, Andrews et al. (2001) used six phone recognizers of English, German, Hindi, Japanese, Mandarin and Spanish. The enrolled speaker model is a set of phone n-gram probabilities which are given by

$$\Xi_s = \{\xi_{l,s}\}, \qquad (0.11)$$

where s and l are the indexes of speaker and language, respectively. Let Γ_l be the set of phone n-gram types in language l. Then, $\xi_{l,s}$ is defined as follows:

$$\xi_{l,s} = \{H_{l,s}(\mathbf{w}) | \mathbf{w} \in \Gamma_l\}, \qquad (0.12)$$

where $H_{l,s}(\mathbf{w})$ is given by

$$H_{l,s}(\mathbf{w}) = \frac{N_{l,s}(\mathbf{w})}{\sum_{\mathbf{w}'\in\Gamma_l} N_{l,s}(\mathbf{w}')}, \qquad (0.13)$$

and where $N_{l,s}(\mathbf{w}')$ is the number of phone n-gram $\mathbf{w}'$ in the recognition results of speaker s 's enrollment data using the phone recognizer for language l. The matching score of speaker s can be computed as follows:

$$S_{PHN}(\mathbf{x}_1\mathbf{x}_2...\mathbf{x}_T, \Xi_s) = \sum_l \alpha_l \frac{\sum_{\mathbf{w}\in\Gamma_l} c_l(\mathbf{x}_1\mathbf{x}_2...\mathbf{x}_T, \mathbf{w})^{1-d} H_{l,s}(\mathbf{w})}{\sum_{\mathbf{w}\in\Gamma_l} c_l(\mathbf{x}_1\mathbf{x}_2...\mathbf{x}_T, \mathbf{w})^{1-d}}, \qquad (0.14)$$

where α_l and d are a weight for language l and a discounting factor between 0 and 1, respectively. In (0.13), $c_l(\mathbf{x}_1\mathbf{x}_2...\mathbf{x}_T, \mathbf{w})$ is a frequency count of $\mathbf{w} \in \Gamma_l$ in the decoded result of $\mathbf{x}_1\mathbf{x}_2...\mathbf{x}_T$. Recently, it was reported that incorporating the phone-lattice decoding can improve the phonetic speaker recognition, where the phone n-grams are extracted not from the 1-best recognition result but from the phone lattice-decoding (Hatch, Barbara Peskin, & Stolcke, 2005).

Support-Vector Machine (SVM)

In a two-class SVM, an input feature vector is classified into either class 0 or class 1 using a discriminant function below

$$f(\mathbf{x}_i) = \sum_{l=0}^{L-1} \alpha_l t_l K(\mathbf{x}_i, \mathbf{u}_l) + d \qquad (0.15)$$

where $\mathbf{u}_l$ and t_l are the l th support vector trained using training data and its label. Here, t_l is -1 if the support vector $\mathbf{u}_l$ is in class 0, and t_l is 1 if the support vector $\mathbf{u}_l$ is in class 1. The weight α_l is trained with constraints of $a_l > 0$ and $\sum_{l=0}^{L-1} \alpha_l t_l = 0$: for detailed explanation on the SVM, please refer to (Scholkopf & Smola, 2002). The kernel $K(\cdot,\cdot)$ is a pre-defined kernel function. Schmidt and Gish (1996) proposed a speaker recognition system with speaker-dependent SVMs that are trained on enrolled speakers' acoustical features. Motivated by the extreme success of GMM-UBM in speaker recognition, Wan & Renals (2005) proposed to use the Fisher-kernel for a GMM score $S_{GMM}(\mathbf{x}_1\mathbf{x}_2...\mathbf{x}_T, \Theta_s)$ and its derivatives with respect to Θ_s. Given the set of all input feature vectors $\mathbf{X} = [\mathbf{x}_0, \ \mathbf{x}_1, \ ..., \ \mathbf{x}_{T-1}]$, the output of the SVM can be used as the likelihood of a target speaker s as follows:

$$\begin{aligned} S_{SVM,WAN}(\mathbf{X} \mid \{\alpha_l, t_l, \mathbf{u}_l\}) &= f(\mathbf{X}) \\ &= \sum_{l=0}^{L-1} \alpha_l t_l K(\mathbf{X}, \mathbf{u}_l) + d, \end{aligned} \qquad (0.16)$$

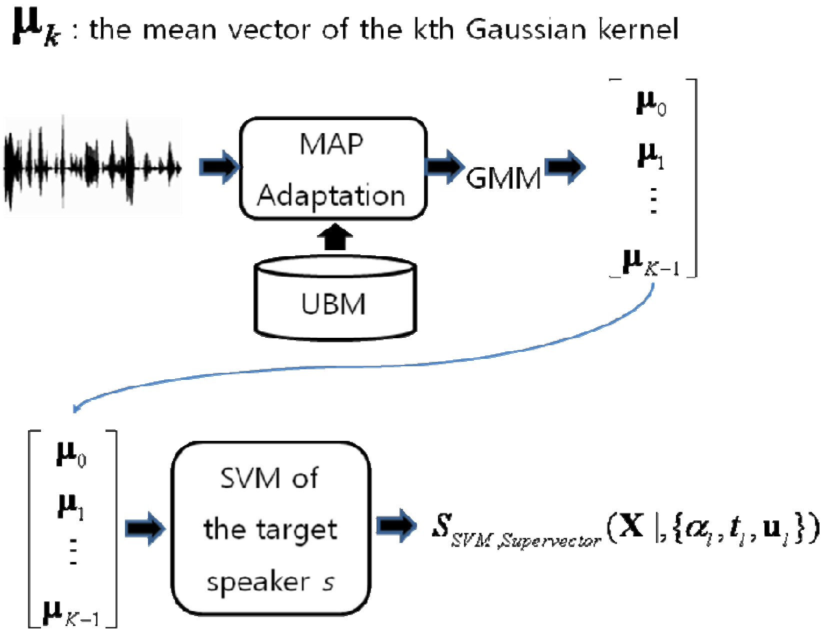
$\boldsymbol{\mu}_k$: the mean vector of the kth Gaussian kernel
MAP Adaptation
UBM
GMM
$\boldsymbol{\mu}_0$
$\boldsymbol{\mu}_1$
$\vdots$
$\boldsymbol{\mu}_{K-1}$
$\boldsymbol{\mu}_0$
$\boldsymbol{\mu}_1$
$\vdots$
$\boldsymbol{\mu}_{K-1}$
SVM of the target speaker s
$S_{SVM,Supervector}(\mathbf{X} \mid, \{\alpha_l, t_l, \mathbf{u}_l\})$

Figure 8 SVM based on GMM supervector for speaker recognition

where the kernel can be a Fisher kernel defined on the score space as follows:

$$K(\mathbf{X}, \mathbf{u}_l) = \psi_s(\mathbf{X}) \bullet \psi_s(\mathbf{u}_l), \qquad (0.17)$$

and where

$$\psi_s(\mathbf{X}) = \begin{bmatrix} \nabla_{\Theta_s} \log p(\mathbf{X} \mid \Theta_s) \\ \log p(\mathbf{X} \mid \Theta_s) \end{bmatrix} \qquad (0.18)$$

By using the Fisher kernel, the input feature vector is mapped into a hyper plane where the mapped feature vector is more sparsely distributed than that in the original feature vector space (Jaakkola & Haussler, 1998).

Figure 8 illustrates the system using the SVM for GMM parameters proposed by Campbell et al (2006). Given input feature vectors, the system first adapts the UBM; the UBM is trained as a GMM. The supervector of the adapted GMM, which is a concatenation of mean parameters of all Gaussian kernels in the adapted GMM, is used an input to the SVM. The experimental results have shown that incorporating the GMM parameters into the SVM can improve the speaker verification performance.

In the phonetic speaker recognition, where the speaker model is trained as phone n-grams, (W. Campbell, J. Campbell, Reynolds, Jones, & Leek, 2004) demonstrated that the kernels constructed using vector of $H_{l,s}(\mathrm{w})$ in (0.13) can improve the speaker verification performance.

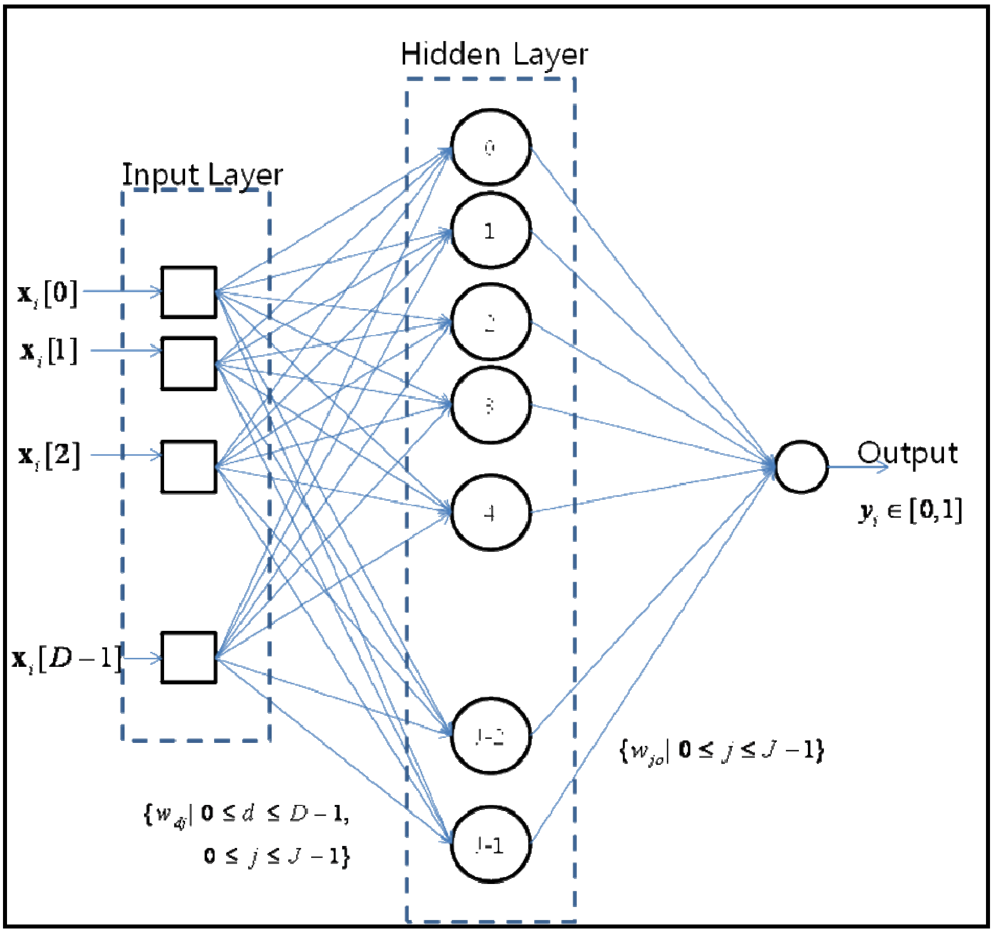

Figure 9 A Multi-Layer Perceptron Neural Network: this figure illustrates an example of MLP which consists 1 input, 1 hidden layer, and an output layer.

Neural Network

Speaker recognition can benefit from the neural network since the exact mapping between an input (speech features) and an output (speaker identity) is not known. Farrell et al (1994) have considered two kinds of neural netoworks for speaker recognition, multi-layer perceptron (MLP) (Haykin, 1999) and neural tree network (NTN) (Sankar & Mammone, 1991). The MLP can be trained for speech vectors such as pitch, LPC, etc. When training the MLP of a target speaker s, feature vectors from that speaker are labelled as 1, and feature vectors from non-target speakers are labelled as 0. Very often, the number of feature vectors from a target speaker is much smaller than those from non-target speakers. With such unbalanced label data, an MLP can be trained to alwaly output 0 (impostor). For this reason, Kevein et al (1994) performed VQ on training data from non-target speakers, and only the codewords are used as training data for label 0. Figure 9 illustrates an MLP with 1 hidden layer. Let $\mathbf{x}_i = \left[x_i[0], \quad x_i[1], \quad \ldots, \quad x_i[D-1]\right]$ be the input to the MLP. A weight w_{ij} corresponds to the weight of the arrowed line from a node i in the input layer to a node j in the hidden layer. The input to the hidden node j is a weighted sum of input nodes given by

$$v_j = \sum_{d=0}^{D-1} w_{dj} x_i[d] \qquad (0.19)$$

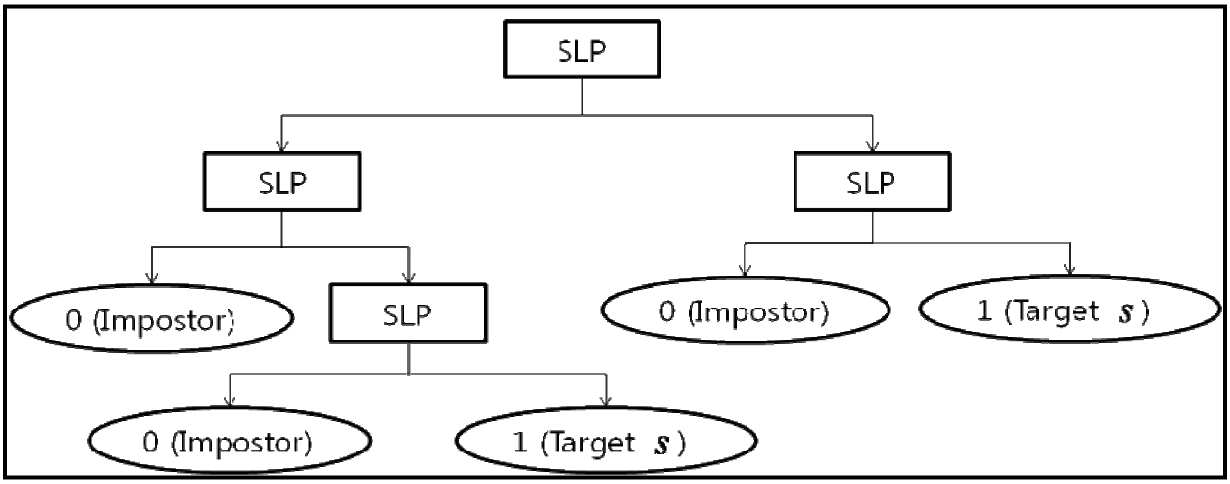

Figure 10 An NTN trained for a target speaker s : the decision tree consists of several SLPs.

Then, the output y_j of node j is computed as follows:

$$y_j = \varphi_j(v_j) \qquad (0.20)$$

where $\varphi_j(\cdot)$ is the activation function of node j . Very often, the activation function is modeled as either logistic function

$$\varphi_j(v_j) = \frac{1}{1+\exp(-av_j)} \qquad (0.21)$$

or hyperbolic tangent function

$$\varphi_j(v_j) = b\tanh(cv_j) \qquad (0.22)$$

where a, b and c are preset parameters. Finally, the output of the MLP is obtained as follows:

$$y_i = \sum_{j=0}^{J-1} w_{jo} y_j \qquad (0.23)$$

For binary classification, the output y_i is often compared to a preset threshold, and is mapped into either 0 or 1. The weights $\{w_{dj},\ \ w_{jo}\}$ can be trained using the back-propagation algorithm : for details, please refer to (Haykin, 1999). The weights in the MLP can be trained by the back-propagation algorithm, but the number of layers and nodes must be pre-defined. For speaker recognition, Farrell et al. (1994) trained the MLP of each target speaker with following configuration:

- Input feature: 12th order LPC
- MLP structure: 64 hidden nodes within one hidden layer
- Activation function: Logistics function

Given a sequence of input feature $\mathrm{x}_1\mathrm{x}_2 \ldots \mathrm{x}_T$, the likelihood of a target speaker s can be computed as follows:

$$S_{MLP}(\mathrm{x}_1\mathrm{x}_2 \ldots \mathrm{x}_T, \Pi_s) = \frac{N_1}{N_1 + N_0} \qquad (0.24)$$

where N_0 and N_1 are the number of feature vectors whose output is 0 (impostor) and 1 (target), respectively. The above function measures how many feature vectors are classified into a target speaker with its dynamic range normalized from 0 to 1. The normalized dynamic range can be beneficial to the speaker verification where it is important to set a proper threshold of an accept/reject decision.

In addition to the MLP, the NTN can be used for speaker recogntion. The NTN is a hybrid of a decision tree and a neural network, where the decision is made by the class of a leaf node that the input feature enters. Figure 10 llustrates an NTN using single-layer perceptron (SLP) neural network trained for a target speaker s . The hierarchical structure of NTN can be self-orgainzed as follows:

1. Create a root node, and assign entire training data to the root node
2. If there exists a node whose training data are not uniformly labeled (i. e., if there exists a node where both 0 and 1 training labels are observed in its training data)
    A. Train an SLP with its training data
    B. Let $\mathbf{p}$ and $\mathbf{q}$ be the set of training data whose SLP output is 0 and 1, respectively.
    C. Create two children nodes, and assign $\mathbf{p}$ and $\mathbf{q}$ to these children nodes.
    E. Go to 2.
3. Terminate.

The above procedure ends when all leaf nodes are uniformly labelled, i. e., all training data in each leaf node are either 0 or 1, respectively. In Figure 10, leaf nodes are assigned to either 0 for impostor or 1 for the target speaker. When classifying an input feature vector, at every non-leaf node (rectangular shaped), an input feature is classified into one of two child nodes. When the input feature enters a leaf node (oval shaped), the NTN outputs the class of that leaf node: in the speaker recognition considered in Figure 10, the output class is either a target speaker s or impostor for s. Similarly to the MLP, the likelihood of an input feature vector can be obtained as follows:

$$S_{NTN}(\mathbf{x}_1\mathbf{x}_2 ... \mathbf{x}_T, \Psi_s) = \frac{N_1}{N_1 + N_0} \quad (0.25)$$

Additionally, Kevin et al (1994) have used the following likelihood

$$S_{MNTN}(\mathbf{x}_1\mathbf{x}_2 ... \mathbf{x}_T, \Pi_s) = \frac{\sum_{k=0}^{N_1-1} c_k^1}{\sum_{k=0}^{N_1-1} c_k^1 + \sum_{l=0}^{N_0-1} c_l^0} (0.26)$$

where c_k^1 is the confidence score for output 1 obtained from the SLP for the kth input feature vector that is classified as the target speaker. Similarly, c_l^0 is the confidence score for output 0 obtained from the SLP for the lth input feature vector that is classified as an impostor. Farrell et al. (1994) reported that using the confidence score in (0.26) can significantly improve the performance compared to using (0.25).

Speaker Verification

The speaker verification system first extracts feature from input speech and computes the matching score of the claimed speaker model. The speaker verification system differs from the speaker identification system in that 1) it requests additional information of claimed identity and 2) the decision is an “Accept/Reject” decision. The accept/reject decision can be formulated as follows:

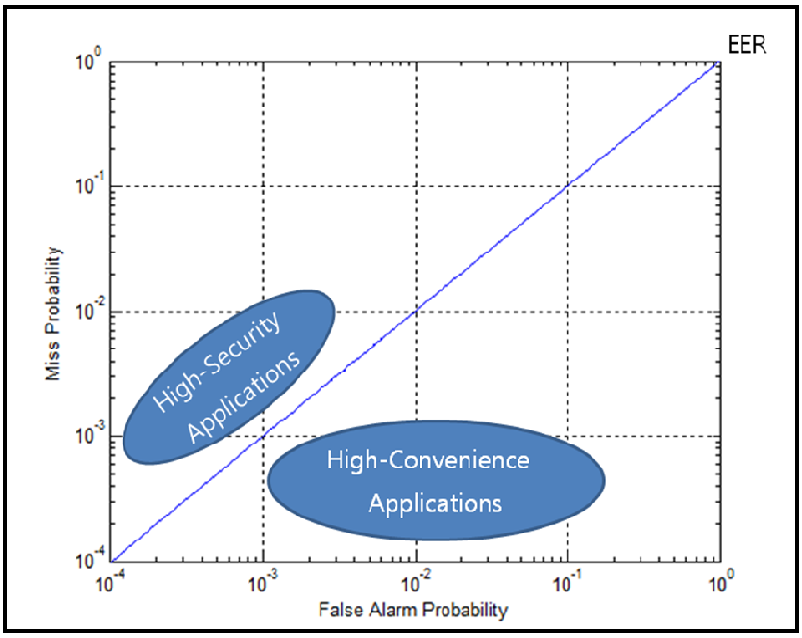

Figure 11. Detection Error Trade-off Curve: the horizontal and the vertical axis represent the false alarm and miss probabilities, respectively. The blue straight line intersects a DET curve at the EER operating point.

- H_0: The input speech is accepted as the claimed speaker s .
- H_1: The input speech is rejected.

$$\frac{S(\mathbf{x}_1\mathbf{x}_2...\mathbf{x}_T \mid H_0)}{S(\mathbf{x}_1\mathbf{x}_2...\mathbf{x}_T \mid H_1)} \underset{H_1}{\overset{H_0}{\gtrless}} \tau, \qquad (0.27)$$

where τ, $S(\mathbf{x}_1\mathbf{x}_2...\mathbf{x}_T \mid H_0)$ and $S(\mathbf{x}_1\mathbf{x}_2...\mathbf{x}_T \mid H_1)$ are a preset threshold, the scores of H_0 and H_1, respectively. Let $S(\mathbf{x}_1\mathbf{x}_2...\mathbf{x}_T \mid s)$ be the matching score of the claimed speaker s : for example, (0.6), (0.8) and (0.14) can be used as $S(\mathbf{x}_1\mathbf{x}_2...\mathbf{x}_T \mid s)$. Very often, $S(\mathbf{x}_1\mathbf{x}_2...\mathbf{x}_T \mid s)$ is used as $S(\mathbf{x}_1\mathbf{x}_2...\mathbf{x}_T \mid H_0)$. In order to compute the score of H_1, the UBM-based and cohort-based approaches can be used. In the UBM-based approach, the score $S(\mathbf{x}_1\mathbf{x}_2...\mathbf{x}_T \mid H_1)$ is approximated using the score of UBM. For example, in the GMM-UBM system, (0.27) is performed as follows:

$$\frac{S(\mathbf{x}_1\mathbf{x}_2...\mathbf{x}_T \mid H_0)}{S(\mathbf{x}_1\mathbf{x}_2...\mathbf{x}_T \mid H_1)} = \frac{S_{GMM}(\mathbf{x}_1\mathbf{x}_2...\mathbf{x}_T, \Theta_s)}{S_{GMM}(\mathbf{x}_1\mathbf{x}_2...\mathbf{x}_T, \Theta_{UBM})} \underset{H_1}{\overset{H_0}{\gtrless}} \tau, \quad (0.28)$$

where Θ_{UBM} is the set of GMM parameters that are trained on many speakers' development data: the development data is separate from the test and enrollment data. In the cohort model-based approach, cohort speakers are randomly selected speakers or most competitive speakers (Rosenberg, Lee, & Soong, 1990; Che et al., 1996; Higgins, Bahler, & Porter, 1991). Let $\breve{s}_k$ be the k th cohort speaker of speaker s . Then, the score of H_1 is computed as follows:

$$S(\mathbf{x}_1\mathbf{x}_2...\mathbf{x}_T \mid H_1) = f\left(S(\mathbf{x}_1\mathbf{x}_2...\mathbf{x}_T \mid \breve{s}_1), S(\mathbf{x}_1\mathbf{x}_2...\mathbf{x}_T \mid \breve{s}_2), ..., S(\mathbf{x}_1\mathbf{x}_2...\mathbf{x}_T \mid \breve{s}_K)\right) \qquad (0.29)$$

where f is a function that computes statistics of arguments $S(\mathbf{x}_1\mathbf{x}_2...\mathbf{x}_T \mid \breve{s}_k)$ for $k = 1, 2, ..., K$: for example, f can be arithmetic, harmonic and geometric mean of $S(\mathbf{x}_1\mathbf{x}_2...\mathbf{x}_T \mid \breve{s}_k)$ for $k = 1, 2, ..., K$.

The performance of the speaker verification system can be measured with various metrics such as equal-error rate (EER) and a detection error trade-off (DET) curve. The EER is the false alarm probability or the miss probability at the threshold τ where the false alarm probability equals the false rejection probability. Figure 11 illustrates the DET curve. Applications that require high security should have low false alarm probability, and applications that require high-convenience should have low miss probability. In this chapter, the EER is used to compare performances of speaker recognition systems.

Speaker Identification

The speaker identification system first extracts feature from an input speech, and it computes the matching scores of target speaker models. Then, the speaker s^* whose matching score is the largest is identified as a result as follows:

$$s^* = \arg\max_{s} S(\mathbf{x}_1\mathbf{x}_2...\mathbf{x}_T \mid s). \qquad (0.30)$$

The performance of a speaker identification system can be measured in terms of the accuracy.

PERFORMANCE OF SPEAKER RECOGNITION SYSTEMS

The performance of speaker recognition system can be affected by many factors (Bimbot et al., 2004; Furui, 1997):

- Distortions in speech
  - Noise: speech can be collected in noisy environment
  - Channel distortions: speech can be collected via telephone and other channels, which can distort the speech signal.
- Speech variability
  - Physical condition of the user: the speech is affected by the physical condition of the user, especially by the condition of laryngitis.
  - Intersession variability: the difference between the enrollment environment and test environment can affect the speaker recognition performance.
  - Speaking style: the performance of speaker recognition system varies with speaking styles.
- Amount of data
  - Amount of enrollment data: the amount of enrollment data affects the quality of enrolled speaker model
  - Amount of test data: the accuracy of speaker recognition system can be improved with long test speech.

In order to reduce the effect of distortions in speech, normalization in both the speech feature and scores have been proposed (Auckenthaler, Carey, & Lloyd-Thomas, 2000; Bimbot et al., 2004; D. A. Reynolds, 1995)

- Z-norm
  - Each speaker model has different dynamic range of matching scores. In order to normalize the matching scores of different models, the model-specific mean and the variance of the matching score are computed, and these parameters are used to normalize the matching score to follow a standard normal distribution.
- H-norm
  - The telephone speech is affected by the types of handset and microphone used. The effects of handset can be reduced by first recognizing which type of handsets is used for the input speech, and the matching score is normalized to follow a standard normal distribution using the handset-specific parameters (mean and variance).

Researchers have reported their experimental results of speaker recognition with different evaluation data, thus it is difficult to compare various experimental results. In the rest of this chapter, speaker recognition performance is evaluated with following criteria:

- Speaker Verification: EER
  - The amount of enrollment data
  - Operation mode: TD (fixed and variable text), TI
  - Acoustical environment: recorded in sound booth, telephone (land line and cellular)

- Speaker Identification: accuracy

	TIMIT	NTIMIT	YOHO	Switchboard I	Mercury
Mode	TI	TI	TD (variable)	TI	TI
# of speakers	630	630	138	500	38
Enrollment data / speaker	24sec	24sec	100sec	6min	30-90sec, variable
Length of Test Utterance	3sec	3sec	2.5sec	1min	2sec
Speaking Style	Read Sentence	Read Sentence	Combination Lock	Conversation	Conversation
Acoustical Environment	Clean, Recorded in Sound Booth	PSTN channel, Recorded in Sound Booth	Clean, Recorded in Sound Booth	Telephone	Telephone
EER	GMM : 0.24% (Reynolds, 1995) NTN : 1.9% (Farrell et al., 1994)	GMM : 7.19% (Reynolds, 1995)	HMM : 0.51% (Che et al., 1996)	GMM : 5.15% (Reynolds, 1995)	GMM&HMM : 4.83% (Hazen et al., 2003)

Table 1 EERs of speaker verification system in various tasks

- The number of speakers to be identified
- The amount of enrollment data
- Operation mode: TD (fixed and variable text), TI

EER of Speaker Verification Systems

Table 1 summarizes the EERs for speaker verification systems that are reported for YOHO (Campbell, 1995), TIMIT (Garofolo, 1993), NTIMIT (Fisher et al., 1993), Switchboard (Godfrey, Holliman, &

Modeling	Feature	EER
GMM-UBM (2,048 Kernels)	MFCC	0.70%
GMM-UBM (512 Kernels)	Pitch + Energy Slopes + Duration + Phoneme Context	5.20%
Phone n-grams (5 multiple languages)	Phone n-grams from speech recognizers with 5 language phone models	4.80%
Word n-grams	Word n-grams from English speech recognizer	11.00%
Fusion	Matching scores of 8 different algorithms	0.20%

Table 2 EERs of various speaker verification systems in NIST SRE 2001 extended task (Reynolds et al., 2003)

McDaniel, 1992) and Mercury dataset (Hazen et al., 2003). In the TIMIT and the YOHO database, where both enrollment and test data are recorded in sound-booth, the EER is less than 0.6%. The NTIMIT database is identical to the TIMIT database except that the speech is distorted with PSTN network, and it results in the verification performance degradation as shown in Table 1. According to Reynolds (2002), the accuracies of speaker verification systems are usually as follows:

- TD with combination lock: 0.1%-1%
  - Clean data
  - 3 min. enrollment data
  - 2 sec. test utterance
- TD with 10 digit strings: 1%-5%
  - Telephone data from multiple handset and multiple sessions
- TI with conversational speech: 7%-15%
  - Telephone data from multiple handset and multiple sessions
  - 2 min. enrollment data
  - 30 sec. test utterance
- TI with read sentences: 20%-35%
  - Noisy radio data from military radios
  - 30sec. enrollment data
  - 15sec. test utterance

The verification performances of various TI systems were reported in the SuperSID workshop (Reynolds et al., 2003). Various speaker modeling techniques are evaluated by NIST 2001 SRE Extended Task

	TIMIT	NTIMIT	YOHO	Switchboard	Mercury
Mode	TI	TI	TD (variable)	TI	TI
# of speakers	630	630	138	500	38
Enrollment data / speaker	24sec	24sec	100sec	6min	30-90sec, variable
Length of Test Utterance	3sec	3sec	2.5sec	1min	2sec
Speaking Style	Read Sentence	Read Sentence	Combination Lock	Conversation	Conversation
Acoustical Environment	Clean, Recorded in Sound Booth	PSTN channel, Recorded in Sound Booth	Clean, Recorded in Sound Booth	Telephone	Telephone
Accuracy	GMM : 99.50% (Reynolds, 1995) MLP : 90% (Farrell et al., 1994) NTN : 96 % (Farrell et al., 1994)	GMM : 60.70% (Reynolds, 1995)	GMM&HMM : 99.75% (Park & Hazen, 2002)	GMM : 82.80% (Reynolds, 1995)	GMM&HMM : 81.70% (Hazen et al., 2003)

Table 3 Accuracy of speaker identification system in various tasks

	16-individuals (Average EER)	**Median of Scores from 16 individuals**	**Mean of Scores from 16 individuals**	**Machine (Hierarchical GMM)**
EER	23%	12.50%	12.00%	12.00%
	(Schmidt-Nielsen & Crystal, 2000)	(Schmidt-Nielsen & Crystal, 2000)	(Schmidt-Nielsen & Crystal, 2000)	(Liu et al., 2002)

Table 4 Comparison of human listeners and automatic speaker verification system

(Adami, Mihaescu, Reynolds, & Godfrey, 2003; Jin et al., 2003; Klusacek, Navratil, D. A. Reynolds, & J. P. Campbell, 2003; Navratil et al., 2003; Reynolds et al., 2003) .

- NIST 2001 SRE Extended Task
  - ○ Data source: Switchboard I (Conversational speech on telephone line)
  - ○ Enrollment data : Nominally 20 min. per each target speaker
  - ○ Test data : 14-45 sec

Table 2 summarizes verification performances of a GMM-UBM with MFCCs, phone n-grams from speech recognizers of 5 different languages, word n-grams from English recognizers: for details, please refer to (Reynolds, et al., 2003). The GMM-UBM with MFCCs showed outstanding performance. Using other features such as prosodic features and phone n-gram features do not perform as well as the GMM-UBM with MFCCs, but it is shown that fusing the scores from different systems can reduce the EER. This implies that acoustical features, prosodic features, and phone and word n-gram features are complementary features for speaker recognition.

Accuracy of Speaker Identification Systems

Table 3 summarizes the speaker identification accuracies. As in speaker verification, the acoustical environment affects the accuracy. In speaker verification experiments, many studies have reported score and feature normalization techniques (Auckenthaler et al., 2000; Barras & Gauvain, 2003; Pelecanos & Sridharan, 2001). Applying these techniques to the NTIMIT database could improve the accuracy of NTIMIT DB. The number of speakers to be identified in the Switchboard database is larger than that in the Mercury database, and the length of test utterance and the amount of enrollment data in the Switchboard database is also larger than those in the Mercury database. For this reason, the identification accuracies in the Switchboard and the Mercury database are similar.

Human and Machine Performance Comparison on Speaker Verification

Schmidt-Nielson and Crystal (2000) compared the speaker verification performances between human listeners and an automatic speaker verification (ASV) system, where the following restrictions are forced to human listeners and the ASV system:

- The text-transcription of speech is unavailable for the ASV system.

- Human listeners are allowed to hear 2 min. enrollment data 3 times.
- The gender of speakers are provided for both human listeners and the ASV system
- The verification experiments of human listeners are performed by following procedures:
    - ○ Listening to 2 min. enrollment data of target speakers
    - ○ For each test sample

        Listen to test sample twice.
        Make 10-level scores on accept/reject decision

The experiments were performed using a database from NIST speaker recognition evaluation (SRE) 1998 (M. Przybocki & A. F. Martin, 2004). The number of human listeners participated in this experiment is 16. The average EER of these 16 human listeners is 23%, which is worse than the machine performance reported using the hierarchical GMM (Liu, Chang, & Dai, 2002). When matching scores from 16-individuals are available, the EERs of using the mean and the median are around 12%, which is similar to the EER of machine. The EERs of human listeners are reported to range from 10% to 25%, where the speaker verification performance is dependent on the human listener involved in the task. Human listeners did not outperform the ASV system in average, but they tend to outperform the ASV system for degraded speech as stated in Schmidt-Nielsen and Crystal (2000) as follows:

> Both human and algorithm performance also went down when the signal was degraded by background noise, crosstalk, poor channel conditions, etc., but again the humans were more robust for the worst conditions.

APPLICATIONS

The major application of speaker recognition system is in an on-site or in a network-based access control. Since the speaker recognition system is not reliable enough for sensitive applications, very often, the speaker recognition system must be incorporated into a service that allows a certain level of verification/identification error. Applications to on-site access control for secured rooms can be considered. In such applications, a user claims his/her identity using ID card, badge, or personal identification number (PIN), and the claimed identity is verified using voice input. The voice signal is acquired in controlled environments, thus very often the speaker recognition system is reliable. The network-based access control differs from the on-site access control in that the voice is acquired through a remote terminal such as a telephone or a computer. Very often, the network-based access control applications are designed for customer relationship management (CRM), and various applications related to the CRM can be found in Markowitz (2002)'s survey. For example, the Union Pacific Railroad employed the speaker verification system. When a customer's goods are arrived, the customer calls the company to release empty railcars. Since the railcar number is complex, customers often mispronounce the railcar number. The speaker verification system determines whether the voice comes from the customer who is supposed to release the pronounced railcar number, which reduces the number of incorrectly released rail cars. Another CRM application is speaker verification based password reset system, which have been used by Volkswagen Financial Services. When a user forgets his/her password, the system resets the password only when the user's voice matches the enrolled speaker model. This can reduce many steps that are required to validate the user's identity.

The automatic speaker recognition techniques are also applicable to forensic speaker recognition which can be termed as follows:

> Expert opinion is increasingly being sought in the legal process as to whether two or more recordings of speech are from the same speaker. This is usually termed forensic speaker identification, or forensic speaker recognition. (Rose 2002, p. 19)

In the forensic speaker recognition, the speaker model is often enrolled using limited data obtained from the questioning, and the voice evidence to be matched with the speaker model is usually distorted by acoustical noise and channel distortions. In addition, speakers are not cooperative and attempt to disguise their speaking style (Bimbot et al., 2004; Campbell, Reynolds, Campbell, & Brady, 2005; Kunzel, 1994; Meuwly & Drygajlo, 2001). Automatic speaker recognition systems have been evaluated on forensic speaker recognition tasks, but still, automatic speaker recognition systems do not achieve sufficient performance to be used in the forensic speaker recognition task (Nakasone & Beck, 2001).

CONCLUSION

In this chapter, we have considered basic concepts of speaker recognition and common speaker recognition techniques. The performances of speaker recognition systems are shown using experimental results from various tasks of YOHO, TIMIT, NTIMIT, Switchboard, Mercury and NIST database. The performances of both speaker verification and identification are highly dependent on the acoustical environment and speaking styles. In order to compare performances of various speaker modeling techniques, the experimental results of super SID evaluation were presented. The GMM-UBM with MFCCs performed the best, and fusing scores from different systems improved the performance. Schmidt-Nielson and Crystal (2000)'s work was presented to compare the human listeners with the ASV system in speaker verification. The ASV system slightly outperformed human listeners, but human listeners were reported to be robust against distortions. This implies that more research on human listener's speaker recognition mechanism can improve the ASV system. The applications of speaker recognition were described. Many systems were designed for CRM applications using telephone. These applications have shown that the performance of speaker recognition system is not sufficient to be used for user authentication, but it can help the CRM system to reduce costs incurred by mistakes.

Currently, it is not sufficient for the speaker recognition system to be used as a stand-alone user authentication system. One alternative is to use the verbal information verification which incorporates speech recognition into speaker recognition (Li, Juang, Zhou, & Lee, 2000). With a properly defined confidence measure, the speaker recognition system can measure the reliability of the recognition results. If the reliability is smaller than a preset threshold, then the system may ask the speaker some private information which is enrolled beforehand, e.g., mother's maiden name, telephone number, etc. Then, the user's answer is recognized and determined to be correct or not by automatic speech recognition. By iterating this procedure until the system can make a reliable decision, the speaker recognition performance can be significantly improved (Li, Juang, Zhou, & Lee, 2000).

In addition, the multi-modal speaker recognition can be an alternative. Very often, speech, face and lip shapes are incorporated into the multimodal speaker recognition systems (Brunelli & Falavigna, 1995; Jourlin, Genound, & Wassner, 1997; Ben-Yacoub, Abdeljaoued, & Mayoraz, 1999; Chatzis, Bors, & Pitas, 1999; Wark & Sridharan, 2001; Sanderson & Paliwal, 2003). The face and lip recognition performance may also deteriorate under distortions in vision such as illumination change or head rotation. The probability that both the voice and the vision are degraded may be smaller than the probability that either the voice or the vision is degraded. In addition, recent development of information technology lowers the price of devices that can capture and process the face and lip shapes, which indicates that the speaker, face and lip recognition can be incorporated with a small increment in cost. Multi-modal systems extract features from considered biometric characteristics in the system. Some systems concatenate these features, and verify or identify a speaker by using a classifier trained with a concatenated feature (Bengio, Marcel & Mariethoz 2002). Other systems compute matching scores using these features, and verify or identify a speaker by using a classifier trained with these scores (Jourlin, Genound, & Wassner, 1997; Hazen et al., 2003b,;Erzin, Yemez, & Tekalp 2005). The multi-modal speaker recognition system performs well especially when the voice is contaminated under noisy conditions. In future, the advancement of speech recognition may provide accurate idiosyncratic features that can improve the speaker recognition performance. Relevant challenges to speaker recognition include the estimation of speaker's physical characteristics. Recently, some researchers have considered the estimation of speaker's physical characteristics (height, weight, vocal tract length) via voice. (Smith, & Nelson, 2004; Dusan 2005). The estimated speaker's physical characteristics can be used not only to profile unknown speaker but also to improve the speaker recognition performance.

REFERNCES

Adami, A., Mihaescu, R., Reynolds, D. A., & Godfrey, J. J. (2003). Modeling prosodic dynamics for speaker recognition In *Proceedings of the International conference on Acoustics, Speech, and Signal Processing* .

Alex Park, & Timothy J. Hazen. (2002). ASR dependent techniques for speaker recognition. In *International conference on Spoken Language Processing* (pp. 1337-1340).

Andrews, W. D., Kohler, M. A., Campbell, J. P., & Godfrey, J. J. (2001). Phonetic, idiolectal and acoustic speaker recognition. In A *Speaker Odyssey-The Speaker Recognition Workshop*. ISCA.

Andrews, W. D., Kohler, M. A., & Campbell, J. P. (2001). Phonetic speaker recognition. In *Proceedings of the Eurospeech* (pp. 2517-2520).

Auckenthaler, R., Carey, M., & Lloyd-Thomas, H. (2000). Score normalization for text-independent speaker verification systems. *Digital Signal Processing*, *10*, 42-54.

Barras, C., & Gauvain, J. L. (2003). Feature and score normalization for speaker verification of cellular data. In *Proceedings of the International conference on Acoustics, Speech, and Signal Processing* (Vol. 2).

Ben-Yacoub, S. , Abdeljaoued, Y. , & Mayoraz, E. (1999). Fusion of face and speech data for person identity verification, *IEEE Transactions on Neural Network 10(5)*, 1065-1074

Bengio, S., Marcel, C., Marcel, S., Mariethoz, J. (2002). Confidence measures for multimodal identity verification, *Information Fusion, 3*, 267-276

Bimbot, F., Bonastre, J. F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., et al. (2004). A tutorial on text-independent speaker verification. *EURASIP Journal on Applied Signal Processing*, 430-451.

Brunelli R., & Falavigna, D. (1995). Person identification using multiple clues, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *17(10),* 955-966.

Campbell Jr, J. P. (1995). Testing with the YOHO CD-ROM voice verification corpus. *in: Proceedings of the International conference on Acoustics, Speech, and Signal Processing* , *1.*

Campbell, J. (1997). Speaker recognition: a tutorial. *Proceedings of the IEEE*, *85*, 1437—1462.

Campbell, J. P., Reynolds, D. A., & Dunn, R. B. (2003). Fusing high-and low-level features for speaker recognition. In *Proceedings of the Eighth European Conference on Speech Communication and Technology.*

Campbell, W. M., Campbell, J. P., Reynolds, D. A., Jones, D. A., & Leek, T. R. (2004). Phonetic speaker recognition with support vector machines. *Advances in Neural Information Processing Systems*, *16*, 57.

Campbell, W. M., Reynolds, D. A., Campbell, J. P., & Brady, K. J. (2005). Estimating and evaluating confidence for forensic speaker recognition. In *Proceedings of the International conference on Acoustics, Speech, and Signal Processing*

Campbell, W. M., Sturim, D. E., & Reynolds, D. A. (2006). Support vector machines using GMM supervectors for Speaker Verification. *IEEE Signal Processing Letters*, *13*(5), 308.

Chatzis, V., Bors, A. G., & Pitas, I. (1999). Multimodal decision-level fusion for person authentication *IEEE Transactions on System, Man, Cybernetics A, 29(6),* 674-680.

Che, C., Lin, Q., & Yuk, D. (1996). An HMM approach to text-prompted speaker verification. In *Proceedings of the International conference on Acoustics, Speech, and Signal Processing*

Choi, E., Hyun, D., & Lee, C. (2002). Optimizing feature extraction for English word recognition. *in: Proceedings of the International conference on Acoustics, Speech, and Signal Processing*

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing, 28*(4), 357-366.

Doddington, G. R., Przybocki, M. A., Martin, A. F., & Reynolds, D. A. (2000). The NIST speaker recognition evaluation - overview, methodology, systems, results, perspective-. *Speech Communication*, *31*, 225-254.

Dusan, S. (2005) Estimation of speaker's height and vocal tract length from speech signal, In *Proceedings of INTERSPEECH.*

Erzin, E., Yemez, Y., Tekalp, A. M. (2005) Multimodal speaker identification using an adaptive classifier cascade based on modality reliability. *IEEE Transactions on Multimedia, 7(5),* pp. 840-852

Farrell, K. R., Mammone, R. J.,& Assaleh, T. (1994). Speaker recognition using neural networks and conventional classifiers. *IEEE Trans on speech and audio processing*, 194-205.

Fisher, W. M., Doddington, G. R., Goudie-Mashall, K. M., Jankowski, C., Kalyanswamy, A., Basson, S., et al. (1993). NTIMIT. *Linguistic Data Consortium.*

Furui, S. (1997). Recent advances in speaker recognition. *Pattern Recognition Letters*, *18*(9), 859-872.

Garofolo, J. S., (US, N. I. O. S. A. T., Consortium, L. D., Office, I. S. A. T., States, U., & Agency, D. A. R. P. (1993). TIMIT acoustic-phonetic continuous speech corpus. *Linguistic Data Consortium.*

Gauvain, J., & Lee, C. (1994). Maximum *a* posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Transactions on Audio, Speech, and Language Processing*, *2*(2), 291-298.

Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992). SWITCHBOARD: telephone speech corpus for research and development. In *Proceedings of the International conference on Acoustics, Speech, and Signal Processing*

Hatch, A., Peskin, B., & Stolcke, A. (2005). Improved phonetic speaker recognition using lattice decoding. In *Proceedings of the International conference on Acoustics, Speech, and Signal Processing.*

Haykin, S. (1999). *Neural networks: a comprehensive foundation*. Upper Saddle River, New Jersey: Prentice-Hall, Inc.

Hazen, T. J., Jones, D. A., Park, A., Kukolich, L. C., & Reynolds, D. A. (2003a). Integration of speaker recognition into conversational spoken dialogue systems. In *Proceedings of the Eurospeech*

Hazen, T. J., Weinstein, E., Kabir, R., Park, A., Heisele, B. (2003b) Multi-modal face and speaker identificaiton on a handheld device In. *Proceedings of Workshop on Multimodal User Authentication,* 113-120.

Higgins, A., Bahler, L., & Porter, J. (1991). Speaker verification using randomized phrase prompting. *Digital Signal Processing*, *1*(2), 89-106.

Hirsch, H. G., & Pearce, D. (2000). The AURORA experimental framework for the Performance evaluation of speech recognition systems under noisy conditions. In *ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW).* ISCA.

Huang, X., Acero, A., & Hon, H. (2001). *Spoken language processing*. Prentice Hall PTR, Upper Saddle River, New Jersey.

Jaakkola, T. S. & Haussler, D. (1998). Exploiting generative models in discriminative classifiers. *Advances in Neural Information Processing Systems, 11*

Jain, A. K., Ross, A., & Prabhakar, S. (2004). An introduction to biometric recognition. *Circuits and Systems for Video Technology, IEEE Transactions on*, *14*(1), 4-20.

Jin, M., Soong, F., & Yoo, C. (2007). A syllable lattice approach to speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing, 15*(8), 2476-2484.

Jin, Q., Navratil, J., Reynolds, D. A., Campbell, J. P., Andrews, W. D., & Abramson, J. S. (2003). Combining cross-stream and time dimensions in phonetic speaker recognition. In *Proceedings of the International conference on Acoustics, Speech, and Signal Processing.*

Jourlin, P., Luettin, J., Genoud, D., & Wassner, H. (1997). Acoustic-labial speaker veriifcaiton. *Pattern recognition, 18,* 853-856

Kunzel, H. J. (1994). Current approaches to forensic speaker recognition. In *Automatic Speaker Recognition, Identification and Verification*. ISCA.

Klusacek, D., Navratil, J., Reynolds, D. A., & Campbell, J. P. (2003). Conditional pronunciation modeling in speaker detection. In *Proceedings of the International conference on Acoustics, Speech, and Signal Processing*.

Leeuwen, D. A. V., Martin, A. F., Przybocki, M. A., & Bouten, J. S. (2006). NIST and NFI-TNO evaluations of automatic speaker recognition. *Computer Speech and Language*, *20*, 128-158.

Li, Q., Juang, B.-H., Zhou, Q., & Lee, C. –H. (2000) Automatic verbal information verification for user authenticaion*, IEEE Transactions on Audio Processing, 8(5)*, 585-596.

Li, X., Chang, E., & Dai, B. (2002). Improving speaker verification with figure of merit training. In *Proceedings of the International conference on Acoustics, Speech, and Signal Processing* .

Liu, M., Chang, E., & Dai, B. (2002). Hierarchical Gaussian mixture model for speaker verification. In *Seventh International Conference on Spoken Language Processing*. ISCA.

Markowitz, Judith (2002). Speaker recognition. *Biometric Technology Today*, 10(6), 9-11.

Meuwly, D., & Drygajlo, A. (2001). Forensic speaker recognition based on a Bayesian framework and Gaussian mixture modeling (GMM). In *A Speaker Odyssey-The Speaker Recognition Workshop*. ISCA.

Milner, B. (2002). A comparison of front-end configurations for robust speech recognition. *Proceedings of the International conference on Acoustics, Speech, and Signal Processing.*

Molla, K., & Hirose, K. (2004). On the effectiveness of MFCCs and their statistical distribution properties in speaker identification. In *IEEE Symposium on Virtual Environments, Human-Computer Interfaces and Measurement Systems (VECIMS).*

Nakasone, H., & Beck, S. D. (2001). Forensic automatic speaker recognition. In *A Speaker Odyssey-The Speaker Recognition Workshop*. ISCA.

Navratil, J., Jin, Q., Andrews, W. D., & Campbell, J. P.,(2003). Phonetic speaker recognition using maximum-likelihood binary-decision tree models. In *Proceedings of the International conference on Acoustics, Speech, and Signal Processing*.

Newman, M., Gillick, L., Ito, Y., Mcallaster, D., & Peskin, B. (1996). Speaker verification through large vocabulary continuous speech recognition. In *Proceedings International conference on Spoken Language Processing*

Pan, Y., & Waibel, A. (2000). The effects of room acoustics on MFCC speech parameter. In *International conference on Spoken Language Processing*.

Pelecanos, J., & Sridharan, S. (2001). Feature warping for robust speaker verification. In *A Speaker Odyssey-The Speaker Recognition Workshop*. ISCA.

Peskin, B., Navratil, J., Abramson, J., Jones, D., Klusacek, D., Reynolds, D., et al. (2003). Using prosodic and conversational features for high-performance speaker recognition: report from JHU

WS'02. In *Proceedings of the International conference on Acoustics, Speech, and Signal Processing*

Przybocki, M., & Martin, A. F. (2004). NIST speaker recognition evaluation chronicles. In *A Speaker Odyssey-The Speaker Recognition Workshop*. ISCA.

Rabiner, L. R., & Schafer, R. W., (1978). *Digital processing of speech signals*. Prentice-Hall Englewood Cliffs, NJ.

Reynolds, D. (1995). Speaker identification and verification using Gaussian mixture speaker models. *Speech Communication*, *17*(1-2), 91-108.

Reynolds, D., Quatieri, T., & Dunn, R. (2000). Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, *10*, 19-41.

Reynolds, D. (2002). An overview of automatic speaker recognition technology. In *Proceedings of the International conference on Acoustics, Speech, and Signal Processing.*

Reynolds, D., Andrews, W., Campbell, J., Navratil, J., Peskin, B., Adami, A., et al. (2003). The SuperSID project: Exploiting high-level information for high-accuracy speaker recognition. In *Proceedings of the International conference on Acoustics, Speech, and Signal Processing*.

Rosenberg, A. E., Lee, C., & Soong, F. K. (1990). Sub-word unit talker verification using hidden Markov models. In *Proceedings of the International conference on Acoustics, Speech, and Signal Processing*.

Sanderson C.,& Paliwal, K. K. (2003). Noise compensation in a person verification system using face and multiple speech features*, Pattern Recognit., 36(2),* 293-302.

Sankar, A. & Mammone, R. J. (1991). *Neural tree networks in neural networks: theory and applications*. San Diego, CA: Academic.

Schmidt-Nielsen, A., & Crystal, T. H. (2000). Speaker verification by human listeners: experiments comparing human and machine performance using the NIST 1998 speaker evaluation data. *Digital Signal Processing*, *10*(1-3), 249-266.

Scholkopf, B. & Smola, A. J. (2002). *Learning with kernels*. Cambridge, London: The MIT Press

Shriberg, E. (2007). Higher-level features in speaker recognition. In *Speaker Classification I.*

Smith, L. H., Nelson, D. J. (2004) An estimate of physical scale from speech, In *Proceedings of the International conference on Acoustics, Speech, and Signal Processing.*

Sönmez, K., Shriberg, E., Heck, L., & Weintraub, M. (1998). Modeling dynamic prosodic variation for speaker verification. In *the Proceedings of International conference on Spoken Language Processing*.

Sturim, D. E., Reynolds, D. A., Dunn, R. B., & Quatieri, T. F. (2002). Speaker verification using text-constrained Gaussian mixture models. *Proceedings of the International conference on Acoustics, Speech, and Signal Processing*.

Volkmann, J., Stevens, S. S., & Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *Jounal of the Acoustical Society of America*, 8(3), 1937.

Wan, V., & Renals, S. (2005). Speaker verification using sequence discriminant support vector machines. *IEEE Transactions on Speech and Audio Processing*, *13*(2), 203-210.

Wark, T., & Sridharan, S. (2001) Adaptive fusion of speech and lip information for robust speaker identification, *Digital Signal Processing, 11(3)*, 169-186.

Weber, F., Peskin, B., Newman, M., Emmanuel, A. C., & Gillick, L. (2000). Speaker recognition on single- and multispeaker data. *Digital Signal Processing*, *10*, 75-92.

Woodland, P. C., Leggetter, C. J., Odell, J. J., Valtchev, V., & Young, S. J. (1995). The 1994 HTK large vocabulary speech recognition system. In *Proceedings of the International conference on Acoustics, Speech, and Signal Processing*

Xiang, B., & Berger, T. (2003). Efficient text-independent speaker verification with structural Gaussian mixture models and neural network. *IEEE Transactions on Speech and Audio Processing*, *11*(5), 447-456.

Zilca, R. D. (2002). Text-independent speaker verification using utterance level scoring and covariance modeling. *IEEE Transactions on Speech and Audio Processing*, *10*(6), 363-370.