IDIAP COMMUNICATION

Speech Processing & Text-Independent Automatic Person Verification

Conrad Sanderson (*)

IDIAP–Com 02-08

December 2002
(Minor revision: January 2004)

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 − 27 − 721 77 11
fax +41 − 27 − 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

(*) conradsand @ ieee.org

Contents

List of Figures

List of Tables

Mathematical Notation

$\vec{x}$	a column vector
$\vec{x}^T$	vector transpose of $\vec{x}$
x_i	i-th element of vector $\vec{x}$, e.g., $\vec{x}^T = [\ x_1\ x_2\ ...\ x_D\ ]$, or, $\vec{x}^T = [\ x_i\ ]_{i=1}^{D}$
$\vec{x}_i$	i-th vector in a set
$\{\vec{x}_i\}_{i=1}^{N_V}$	set of N_V vectors
A^T	matrix transpose of A
A^{-1}	inverse of matrix A
$\|A\|$	determinant of matrix A
$\mathbf{\Sigma}$	covariance matrix
λ	parameter set (e.g., parameters of a GMM)

Acronyms

ATM	Automatic Teller Machine
BMS	Background Model Set
CMS	Cepstral Mean Subtraction
DCT	Discrete Cosine Transform
EER	Equal Error Rate
EM	Expectation Maximization
FA	False Acceptance
FA%	False Acceptance rate
FR	False Rejection
FR%	False Rejection rate
GMM	Gaussian Mixture Model
MACVs	Maximum Auto-Correlation Values
MFCCs	Mel-Frequency Cepstral Coefficients
SNR	Signal-to-Noise Ratio
SSC	Spectral Subband Centroid
UBM	Universal Background Model
VAD	Voice Activity Detector
ZCPA	Zero-Crossing with Peak Amplitude

1 Overview

In this communication we first review the human speech production process and feature extraction approaches commonly used in a speaker verification system. Mel Frequency Cepstral Coefficients (MFCCs), delta (regression) features and Cepstral Mean Subtraction (CMS) are covered. A recently proposed feature set, termed Maximum Auto-Correlation Values (MACVs), which utilizes information from the source part of the speech signal, is also covered. A parametric Voice Activity Detector (VAD), used for disregarding silence and noise segments of the speech signal, is briefly described.

Experiments on the telephone speech NTIMIT database suggest that the performance degradation of a speaker verification system used in noisy conditions can be reduced by extending traditional feature vectors with MACV features.

This communication is an extended and significantly reworked version of part of [51].

2 Introduction

Identity verification (or authentication) systems pervade our every day life. For example, Automatic Teller Machines (ATMs) employ simple identity verification where the user is asked to enter their password after inserting their ATM card. If the password matches the one prescribed to the card, the user is allowed access to their bank account. Similar verification systems can be used to restrict access to rooms and buildings.

While the above verification technique is quite effective, it suffers from a major drawback: only the validity of the combination of a certain possession (in this case, the ATM card) and certain knowledge (the password) is verified. The ATM card can be lost or stolen, and the password can be compromised (e.g., somebody looks over your shoulder while you're entering it). Hence new verification methods have emerged, where the password has either been replaced by, or used in addition to, biometrics such as the person's speech, face image or fingerprints; such physical attributes cannot be lost and vary significantly from person to person.

Apart from the applications listed above, biometrics can be applied to other areas, such as telephone & Internet based banking, airline reservations & check-in and access to computer networks [8, 29, 30], as well as forensic work, where the task is to determine whether a given biometric sample belongs to a given suspect [9, 11], and law enforcement applications [4, 56].

It must be stressed that a verification system is different from an identification system: an identification system attempts to find the identity of a given person out of a pool of N people, while verification is inherently a two class task (from a security point of view this translates to: either the claimant is who he/she claims to be or he/she is an impostor). It must also be noted that while the identification task has received considerable scientific interest, the verification task has the greatest application potential [11, 13]. Both verification and identification systems fall under the general umbrella of *recognition* systems.

As mentioned above, one biometric is the speech signal. Speech based verification systems fall into two categories: *text-dependent* and *text-independent*. In a text-dependent system, the claimant must recite a phrase specified by the system; this is in contrast to a text-independent system, where the claimant can say whatever he or she wishes. The main advantage of a text-independent system is the general absence of idiosyncrasies in the task definition, which allows the system to be applied to many tasks [11]; for this reason, this communication concentrates on the latter category.

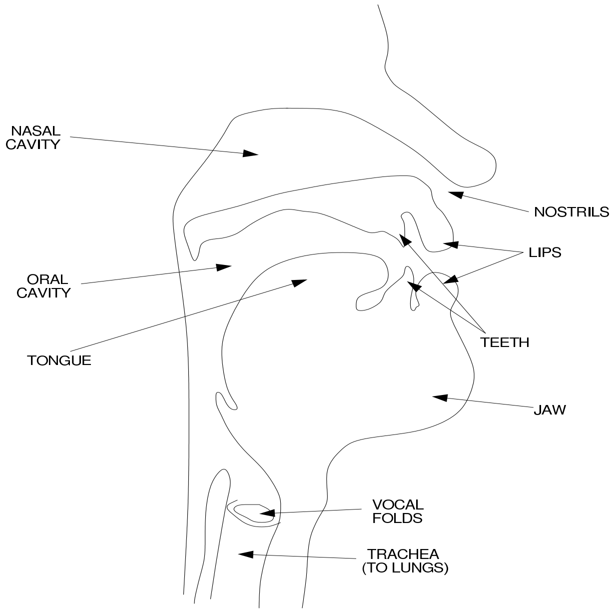

Figure 1: Major vocal tract components (after [52])

2.1 Speech Production Process

Speech can be categorized into two main sound types: *voiced* and *unvoiced*. Voiced sounds are produced as follows. Quasi-periodic opening and closing of the vocal folds, measured in terms of *fundamental* or *pitch* frequency (often abbreviated as $F0$), generates a *glottal wave* composed of energy at $F0$ and at harmonics of $F0$ (i.e. integral multiples of $F0$). The glottal wave is then passed through the vocal tract (see Figure 1). The vocal tract can be modeled as an acoustic tube (starting at the vocal folds and terminating at the lips) with resonances and anti-resonances. The resonances are referred to as *formants*, and are abbreviated to Fi, where $F1$ is the formant with the lowest frequency. The vocal tract, in effect, amplifies energy around formant frequencies and attenuates energy at anti-resonant frequencies. Formant frequencies are changed by modifying the configuration of the articulators (such as the tongue, jaw, lips and teeth), allowing the production of different sounds (e.g., 'aaa' vs 'eee'). In normal speech the articulators are almost constantly moving, indicating that voiced sounds are at best quasi-stationary over short periods of time (tens of milliseconds) [52].

The opening and closing of the vocal folds is accomplished by the following mechanism. At the start of the cycle the vocal folds are closed. Air pressure beneath the vocal folds is increased (due to the constriction of the lungs) and once it overcomes the resistance of the vocal fold closure, it forces the vocal folds apart. Shortly afterward the air pressure is temporarily equalized, and the vocal folds close again, completing the cycle. The cycle occurs at a typical frequency of 60-160 Hz for males and 160-400 Hz for females [38, 23] (average values are 132 Hz and 223 Hz for males and females, respectively [53]). Changes in $F0$ by the speaker are used to denote prosodic information, such as whether a spoken sentence is a statement or a question. While most speakers are capable of changing their $F0$ by two octaves, variation of $F0$ is limited in normal speech since extremes of $F0$ require increased labour.

During the production of unvoiced sounds, the vocal folds do not vibrate. Instead, some of the articulators constrict a point in the vocal tract, causing high speed air flow, which in turn produces an *aperiodic* noise-like signal. The signal is then shaped by the section of the vocal tract in front of the constriction.

As a simplification, the speech signal production process can be thought of as being composed of two parts:

1. The *source part*. Here the source signal may be either periodic, resulting in voiced sounds, or noisy and aperiodic, resulting in unvoiced sounds.

2. The *filter part*, where the source signal is filtered to produce a particular sound.

Thus for voiced sounds the source part generates a signal with spectral energy concentrated at $F0$ (the fundamental frequency) and all its harmonics. The signal is then filtered by the filter part, where the required formants are emphasized, while other parts of the signal are attenuated.

Apart from linguistic information, speech carries person dependent information due to the largely unique configuration of the vocal tract and vocal folds for each person; this causes the time course of $F0$ and the formant frequencies to be person dependent [52].

2.2 Text-Independent Automatic Speaker Verification

Popular speech based verification systems use information from the filter part in the form of a short-time Fourier spectrum represented by Mel Frequency Cepstral Coefficients (MFCCs) [5, 11, 44, 47]. While MFCC features are quite effective for discriminating speakers, they are affected by channel distortion and/or ambient noise. This causes a degradation in the performance of a verification system due to a mismatch between training and testing conditions. There are two popular techniques to reduce the effects of channel distortion and ambient noise: the use of delta (regression) features [14, 54] and Cepstral Mean Subtraction (CMS) [14].

Recently Wildermoth and Paliwal [55] proposed a new feature set, termed Maximum Auto-Correlation Values (MACVs), which utilizes information from the source part of the speech signal; as will be shown, the use of MACV features reduces the performance degradation present due to mismatched conditions.

2.3 Organization

The rest of this communication is organized as follows. In Section 3 we describe the MFCC, CMS, delta and MACV speech feature extraction techniques. In Section 4 we describe a Gaussian Mixture Model (GMM) based classifier which will be used in the experiments we report. In Section 5 we define the (normally used) error measures for finding the performance of a verification system. In Section 6 we describe a parametric Voice Activity Detector (VAD), which is used for disregarding silence and noise segments of the speech signal. Section 7 is devoted to evaluating the use of MACV features to reduce the effects of mismatched conditions. The communication is summarized in Section 8 and some future research is suggested in Section 9.

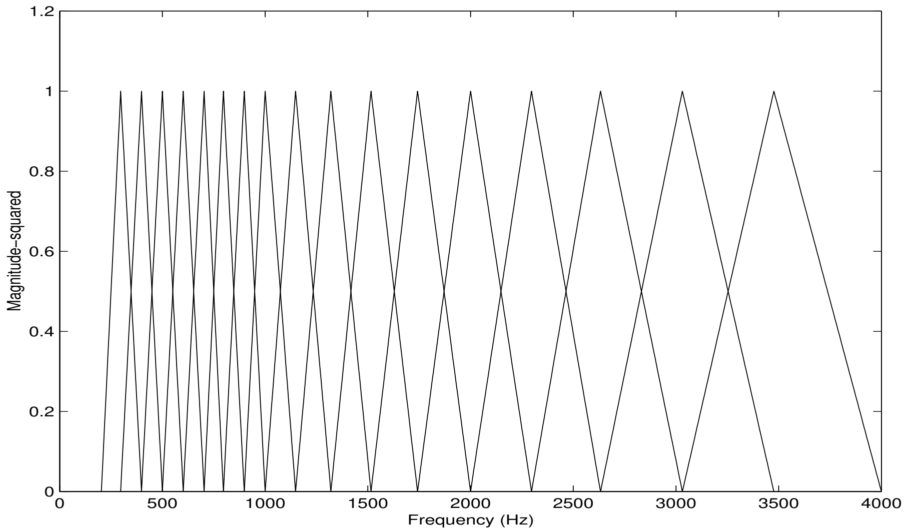

Figure 2: Mel-scale filter bank

3 Feature Extraction Methods

3.1 MFCC Features

In MFCC feature extraction, the speech signal is analyzed on a frame by frame basis, with a typical frame length of 20 ms and a frame advance of 10 ms. For a frame length of 20 ms it can be assumed that the speech signal is stationary, allowing the computation of the short-time Fourier spectrum [36].

Let us denote the speech frame as $\vec{s}^T = [\ s_i\ ]_{i=1}^{N_S}$, where N_S is the number of samples (for a speech signal sampled at 8 kHz, $N_S = 160$ when using 20 ms frames). Each frame is multiplied by a Hamming window to reduce the effects of spectral leakage [41]:

$$\hat{s}_i = s_i h_i, \quad i = 1, 2, ..., N_S \tag{1}$$

where

$$h_i = 0.54 - 0.46 \cos\left(\frac{2\pi(i-1)}{N_S - 1}\right), \quad i = 1, 2, ..., N_S \tag{2}$$

The complex spectrum of $\vec{\hat{s}}^T = [\ \hat{s}_i\ ]_{i=1}^{N_S}$ is then obtained using the Fast Fourier Transform (FFT) algorithm [40, 41]. The square of the magnitude of the complex spectrum is represented as $\vec{S}$ (in our experiments we use a 2048 point representation).

A set of triangular-shaped filters is spaced according to the Mel-scale [39], simulating the processing done by the human ear [23, 33, 34]. For filters chosen to cover the telephone bandwidth, the center frequencies are (in Hz): 300, 400, 500, 600, 700, 800, 900, 1000, 1149, 1320, 1516, 1741, 2000, 2297, 2639, 3031 and 3482. Moreover, to simulate critical bandwidths [39], the upper and lower passband frequencies of each filter are the center frequencies of adjacent filters. For the filter centered at 300 Hz, the lower passband frequency is 200 Hz, while the upper passband frequency for the filter centered at 3482 Hz is 4000 Hz. The responses of $N_F = 17$ filters are shown in Figure 2.

Let $\vec{f}_i$ be the magnitude-squared response of the i-th filter in the frequency domain. The energy output of each filter is obtained using:

$$e_i = \vec{f}_i^T \vec{S}, \quad i = 1, 2, ..., N_F \tag{3}$$

The above equation can be rewritten to obtain an N_F-dimensional energy vector $\vec{e}$:

$$\vec{e} = F^T \vec{S} \tag{4}$$

where $F = [\ \vec{f}_1\ \vec{f}_2\ ...\ \vec{f}_{N_F}\ ]$. It must be noted that Eqn. (4) can be interpreted as a form of dimensionality reduction. In effect, the energy vector $\vec{e}$ represents the smoothed (Mel-warped) spectrum of $\vec{s}$, which is a good representation of the filter part of speech [52].

In order to obtain amplitude normalization, as well as to take into account the diagonal covariance matrix constraint in the GMM classifier (see Section 4), a form of 1D Discrete Cosine Transform (1D DCT) [18] is applied to the log version of $\vec{e}$:

$$g_i = \frac{1}{N_F} \sum_{j=1}^{N_F} \log(e_j) \cos\left(\frac{\pi(i-1)(2j-1)}{2N_F} \right) \quad i = 1, 2, ..., N_F \tag{5}$$

One reason for using the log version of $\vec{e}$ is explained in Section 3.2. Eqn. (5) can be rewritten in matrix notation:

$$\vec{g} = C^T \vec{e}_{\text{log}} \tag{6}$$

where

$$\vec{e}_{\text{log}}^{\,T} = [\ \log(e_i)\ ]_{i=1}^{N_F} \tag{7}$$

and $C = [\ \vec{c}_1\ \vec{c}_2\ ...\ \vec{c}_{N_F}\ ]$, where

$$\vec{c}_i = \left[\frac{1}{N_F} \cos\left(\frac{\pi(i-1)(2j-1)}{2N_F} \right) \right]_{j=1}^{N_F} \qquad i = 1, 2, ..., N_F \tag{8}$$

are the 1D DCT basis vectors.

In Eqn. (5), it can be seen that g_1 represents the average log energy of the spectrum. Since we prefer to use a feature set which is not susceptible to varying background noise and loudness of speech, g_1 is omitted, resulting in a $(N_F - 1)$-dimensional MFCC feature vector:

$$\vec{x} \ = \ [\ g_2\ g_3\ ...\ g_{N_F}\ ]^T \tag{9}$$

Disregarding g_1 can be interpreted as a form of amplitude normalization.

Another popular speech feature extraction method is based on Linear Prediction Cepstral Coefficients (LPCC) [36], which originated from speech compression applications [2, 23, 28]. However, MFCC features are used for experiments in this communication since it has been shown that they are generally more robust than LPCC features for speaker recognition applications [43].

3.2 CMS Features

Let us assume that a signal z is comprised of an original speech signal a that is being filtered by a channel[1] b:

$$z = a * b \tag{10}$$

where $*$ denotes the convolution operation. Thus in the frequency domain the above translates to:

$$Z = AB \tag{11}$$

where Z, A and B are the spectra of z, a and b, respectively. Taking the logarithm of Eqn. (11) yields:

$$\log(Z) = \log(A) + \log(B) \tag{12}$$

Hence in the log domain, the speech signal and the channel are superimposed. Because the energy vector $\vec{e}$ from Eqn. (4) represents the smoothed (Mel-warped) spectrum, Eqn. (11) is analogous to:

$$\vec{e}^{\,T} = [\ e_i\ ]_{i=1}^{N_F} = [\ e_i^a e_i^b\ ]_{i=1}^{N_F} \tag{13}$$

where $\vec{e}^{\,a}$ and $\vec{e}^{\,b}$ represent the smoothed spectrum of a and b, respectively. Taking the log of (13) yields:

$$\vec{e}^{\,T}_{\text{log}} = [\ \log(e_i)\ ]_{i=1}^{N_F} = \left[\ \log(e_i^a) + \log(e_i^b)\ \right]_{i=1}^{N_F} \tag{14}$$

Applying 1D DCT decorrelation to $\vec{e}_{\text{log}}$ yields:

$$\begin{aligned} \vec{g} &= C^T \left(\vec{e}^{\,a}_{\text{log}} + \vec{e}^{\,b}_{\text{log}} \right) && (15) \\ &= C^T \vec{e}^{\,a}_{\text{log}} + C^T \vec{e}^{\,b}_{\text{log}} && (16) \\ &= \vec{g}^{\,a} + \vec{g}^{\,b} && (17) \end{aligned}$$

Thus the effect of the channel is an additive component on the MFCC feature vector:

$$\vec{x} = \vec{x}^{\,a} + \vec{x}^{\,b} \tag{18}$$

Let us define the mean MFCC feature vector for an entire utterance, $\{\vec{x}_i\}_{i=1}^{N_V}$, as:

$$\begin{aligned} \vec{x}^{\,\mu} &= \frac{1}{N_V} \sum_{i=1}^{N_V} \vec{x}_i && (19) \\ &= \frac{1}{N_V} \sum_{i=1}^{N_V} \left(\vec{x}_i^{\,a} + \vec{x}_i^{\,b} \right) && (20) \\ &= \frac{1}{N_V} \sum_{i=1}^{N_V} \vec{x}_i^{\,a} + \frac{1}{N_V} \sum_{i=1}^{N_V} \vec{x}_i^{\,b} && (21) \end{aligned}$$

Assuming that channel characteristics are time invariant leads to:

$$\vec{x}^{\,\mu} = \frac{1}{N_V} \sum_{i=1}^{N_V} \vec{x}^{\,a} + \vec{x}^{\,b} \tag{22}$$

Moreover, if we assume that speech energy is uniformly distributed over the entire spectrum for the duration of the utterance (i.e., the average speech spectrum is flat), then the term $\frac{1}{N_V} \sum_{i=1}^{N_V} \vec{x}^{\,a}$ tends toward zero [6]. Thus $\vec{x}^{\,b}$ can be found using Eqn. (19) and we can obtain channel compensated vectors using:

$$\{\vec{x}_i^{\,\text{comp}}\}_{i=1}^{N_V} = \{\vec{x}_i - \vec{x}^{\,\mu}\}_{i=1}^{N_V} \tag{23}$$

[1]For example, a telephone channel.

The above procedure is known as Cepstral Mean Subtraction (CMS) and Cepstral Mean Normalization (CMN) [3, 6, 14, 43, 45].

As shown in Eqn. (22), the mean feature vector also represents the average speech spectrum; in most practical applications the length of the utterance is not long enough for the second assumption to be valid [6, 17], thus removal of the mean from MFCC features is a double-edged sword: on one hand it makes the verification system more robust to channel mismatches, while on the other it reduces the accuracy of the system in matched conditions (since the average speech spectrum contains speaker information).

In Eqn. (22) it is assumed that the channel characteristics are not changing over time. However, if the characteristics are time-variant, an adaptive bias removal method, such as RASTA processing [21, 22], can be used.

For the sake of convenience, we shall refer to MFCC features with CMS applied simply as CMS features.

3.3 Delta Features

It has been shown that transitional spectrum information contains information which is relatively complimentary to instantaneous spectral information, as well as being less affected by channel effects [54]. Given a sequence of instantaneous spectrum feature vectors, $\{\vec{x}_i\}_{i=1}^{N_V}$, the corresponding transitional spectrum feature vectors are calculated using a modified 1st order orthogonal polynomial fit [14, 25, 54]:

$$\Delta\vec{x}_i = \frac{\sum_{k=-K}^{K} h_k k \, \vec{x}_{i+k}}{\sum_{k=-K}^{K} h_k k^2} \quad \text{for} \quad i = (K+1) \ \text{to} \ (N_V - K) \tag{24}$$

where $\vec{h}$ is a $2K+1$ dimensional symmetric window vector. Typically, $K = 2$ and a rectangular window is used [5, 45, 47] (thus $\Delta\vec{x}_i$ is the slope of the least squares linear fit over the duration of the window).

Transitional spectrum features are better known as delta features. Consequently, instantaneous spectrum features are often referred to as static features [45].

While being more robust to channel effects, delta features do not perform as well as static features in matched conditions [54]. Thus it is general practice to combine the two feature sets by concatenating the delta feature vector with the static feature vector:

$$\vec{y} = \left[\, \vec{x}^T \ \Delta\vec{x}^T \, \right]^T \tag{25}$$

It must be noted that the above concatenation operation can be interpreted as a form of information fusion (see [50] for more information).

Since it is convenient to have the same number of delta and static feature vectors, the "missing" delta feature vectors are generated using:

$$\Delta\vec{x}_i = \Delta\vec{x}_K \quad \text{for} \quad i = 1 \ \text{to} \ K \tag{26}$$

$$\Delta\vec{x}_i = \Delta\vec{x}_{N_V-K} \quad \text{for} \quad i = (N_V - K + 1) \ \text{to} \ N_V \tag{27}$$

Delta-delta (or acceleration) feature vectors ($\Delta\Delta\vec{x}$) can be obtained by applying Eqn. (24) to delta feature vectors. However, use of delta-delta features has shown no measurable improvement in speaker verification performance [11].

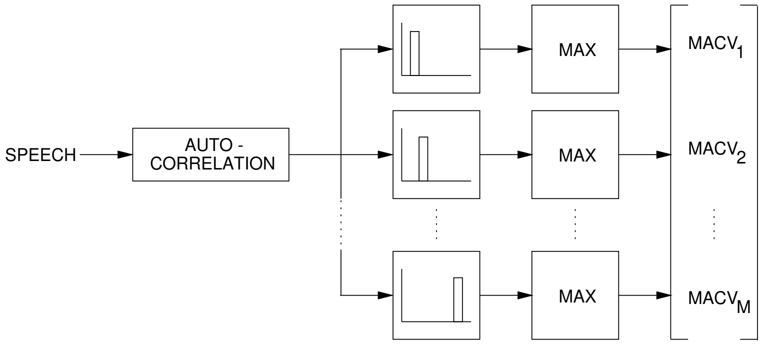

Figure 3: MACV feature extractor (after [55])

3.4 MACV Features

In MFCC features (and hence CMS and delta features) only the system part of the speech signal is effectively utilized. There can be two ways of utilizing pitch (or pitch-related) information:

1. Using a dedicated pitch-based verification sub-system and fusing its output with that of a traditional speaker verification system before reaching the final accept/reject decision. The front-end for the dedicated sub-system can be comprised, for example, of a voiced/unvoiced frame detector, followed by a pitch frequency extractor.

2. Incorporating pitch or pitch-related information directly into the feature vector.

In this communication we will pursue the second approach. The simplest method for detecting the pitch period is by using the autocorrelation function, which for a speech frame $\vec{s}^T = [\ s_i\ ]_{i=1}^{N_S}$ is defined as [41]:

$$R(k) = \frac{1}{N_S} \sum_{i=1}^{N_S-k} s_i\ s_{i+k} \quad k = 0, 1, ..., N_S - 1 \tag{28}$$

If $\vec{s}$ is periodic with a period equal to P samples, then $\{R(k)\}_{k=0}^{N_S-1}$ will show a peak at a lag equal to P. The pitch frequency is typically between 60-160 Hz for males and 160-400 Hz for females [23, 38], indicating that valid pitch lags are approximately between 2.5ms and 16ms. Thus the period of $\vec{s}$ can be found by searching for the maximum of $\{R(k)\}_{k=0}^{N_S-1}$ in the 2.5ms to 16ms range. Due to the harmonic nature of the formants, this approach also allows the recovery of the pitch period when using a telephone channel (which limits the bandwidth of speech signals to between 300 and 3400 Hz).

Unfortunately the auto-correlation method (and other time-domain techniques, such as the Normalized Cross-Correlation Method [1] and the Average Magnitude Difference Method [35, 49]), suffer from pitch doubling and halving as well as other errors [23].

If the signal is periodic with period P, it is also periodic with period $2P$, $3P$, etc. Hence, $\{R(k)\}_{k=0}^{N_S-1}$ will also have maxima at lags equal to $2P$, $3P$, etc. Due to the presence of interfering signals (e.g., noise) and since the speech signal is only quasi-stationary (e.g., the pitch can drift during the duration of the frame), one of the "extra" maxima may be the global maximum; thus the pitch period can be identified as $2P$, which is referred to as pitch halving. When the M-th formant dominates the signal's energy (which can easily occur when using a telephone channel), there will be a maximum at a lag equal to P/M; thus the pitch period can be identified as $P/2$, which is referred to as pitch doubling.

When the speech frame is unvoiced, the above mentioned pitch extraction techniques essentially provide random values [23], indicating that their output cannot be incorporated into the feature vector for each frame.

The recently proposed Maximum Auto-Correlation Value (MACV) feature set [55] overcomes the above problems by deriving pitch related information from the auto-correlation function rather than trying to find the pitch period directly. This is accomplished by dividing the auto-correlation function into several pitch-candidate regions and then finding the maximum value in each region. Formally, the MACV features are obtained as follows:

1. Compute the auto-correlation function $\{R(k)\}_{k=0}^{N_S-1}$.
2. Normalize $\{R(k)\}_{k=0}^{N_S-1}$ by its maximum, i.e., $\left\{\hat{R}(k)\right\}_{k=0}^{N_S-1} = \left\{\frac{R(k)}{R(0)}\right\}_{k=0}^{N_S-1}$.
3. Divide the higher portion (from 2.5 ms to 16 ms) of $\{\hat{R}(k)\}_{k=0}^{N_S-1}$ into N_M equal parts (typically, $N_M = 5$ [55]).
4. Find the maximum value of each of the N_M parts.
5. The N_M Maximum Auto-Correlation Values (MACVs) form an N_M-dimensional feature vector.

A conceptual block diagram of this process is shown in Figure 3. It should be noted that the MACV feature set can also be considered as a non-linear approximation of the mid-section of the autocorrelation function.

Since the MACVs for an unvoiced frame will be relatively low when compared to MACVs for a voiced frame, the MACV feature set also contains voicing information. Moreover, since the MACV feature set does not attempt to extract salient features of the spectrum for each frame (as in MFCC features) it may be less affected by background noise; this conjecture is experimentally tested in Section 7.

4 Gaussian Mixture Model Based Classifier

The distribution of feature vectors for each person is modeled here by a Gaussian Mixture Model (GMM). Given a set of training vectors, an N_G-Gaussian GMM is trained using 10 iterations of the Expectation Maximization (EM) algorithm [7, 10, 12, 32] (which was initialized using a *k*-means clustering algorithm [12, 27]).

Given a claim for person C's identity and a set of feature vectors $X = \{\vec{x}_i\}_{i=1}^{N_V}$ supporting the claim, the average log likelihood of the claimant being the true claimant is calculated using:

$$\mathcal{L}(X|\lambda_C) = \frac{1}{N_V}\sum_{i=1}^{N_V}\log p(\vec{x}_i|\lambda_C) \tag{29}$$

$$\text{where} \quad p(\vec{x}|\lambda) = \sum_{j=1}^{N_G} m_j\, \mathcal{N}(\vec{x}; \vec{\mu}_j, \mathbf{\Sigma}_j) \tag{30}$$

$$\lambda = \{m_j, \vec{\mu}_j, \mathbf{\Sigma}_j\}_{j=1}^{N_G} \tag{31}$$

Here λ_C is the model[2] for person C. N_G is the number of Gaussians, m_j is the weight for Gaussian j (with constraints $\sum_{j=1}^{N_G} m_j = 1$ and $\forall j : m_j \geq 0$), and $\mathcal{N}(\vec{x}; \vec{\mu}, \mathbf{\Sigma})$ is a multi-variate Gaussian function with mean $\vec{\mu}$ and diagonal covariance matrix $\mathbf{\Sigma}$:

$$\mathcal{N}(\vec{x}; \vec{\mu}, \mathbf{\Sigma}) = \frac{1}{(2\pi)^{\frac{D}{2}}|\mathbf{\Sigma}|^{\frac{1}{2}}} \exp\left[\frac{-1}{2}(\vec{x}-\vec{\mu})^T\mathbf{\Sigma}^{-1}(\vec{x}-\vec{\mu})\right] \tag{32}$$

[2]Strictly speaking, model = structure + parameters. In this communication we shall assume that λ represents both the structure (i.e. mixture of Gaussians) and the associated parameters.

where D is the dimensionality of $\vec{x}$. The average log likelihood that the claim for person C's identity is from an impostor is calculated[3] using a Background Model Set (BMS), $B = \{\lambda_b\}_{b=1}^{N_B}$:

$$\mathcal{L}(X|\lambda_{\overline{C}}) = \log\left[\frac{1}{N_B}\sum_{b=1}^{N_B}\exp\mathcal{L}(X|\lambda_b)\right] \tag{33}$$

where $\exp\mathcal{L}(X|\lambda_b)$ can be interpreted as $p(X|\lambda_b)$ which has been normalized to take into account the length of the observation. The BMS for each client is found using the method described in Section 4.1. An opinion on the claim is then found using:

$$\Lambda(X) = \mathcal{L}(X|\lambda_C) - \mathcal{L}(X|\lambda_{\overline{C}}) \tag{34}$$

The opinion reflects the "probability" that a given claimant is the true claimant (i.e., a low opinion suggests that the claimant is an impostor, while a high opinion suggests that the claimant is the true claimant). Given a threshold t, the verification decision is reached as follows:

$$\text{decision} = \begin{cases} \text{true claimant} & \text{if } \Lambda(X) \geq t \\ \text{impostor} & \text{otherwise} \end{cases} \tag{35}$$

4.1 Background Model Set

The set of background models[4] for each client is selected here from the pool of all client models, using the method described by Reynolds [44]; the method is summarized as follows. Using training data, pair-wise distances between each client model are found. For models λ_D and λ_E with corresponding training feature vector sets X_D and X_E (which were used during the construction of the models), the distance is defined as:

$$d(\lambda_D, \lambda_E) = [\mathcal{L}(X_D|\lambda_D) - \mathcal{L}(X_D|\lambda_E)] + [\mathcal{L}(X_E|\lambda_E) - \mathcal{L}(X_E|\lambda_D)] \tag{36}$$

The above symmetric distance defines how similar (or close) the models λ_D and λ_E are. The background model set for each client contains models which are the closest to as well as the farthest from the client model. While it may intuitively seem that only the close models are required (which represent the expected impostors), this would leave the system vulnerable to impostors which are very different from the client. This is demonstrated by inspecting Eqn. (34) where both terms would contain similar likelihoods, leading to an unreliable opinion on the claim.

For a given client model λ_K, N_Φ closest models ($N_\Phi \geq N_B$) are placed in set Φ. Similarly, N_Ψ farthest models ($N_\Psi \geq N_B$) are placed in set Ψ. *Maximally spread* models from the Φ set are moved to set B_{close} using the following procedure:

1. Move the closest model from Φ to B_{close}.
2. Move λ_i from Φ to B_{close}, where λ_i is found using:

$$\lambda_i = \arg\max_{\lambda_j\in\Phi}\left[\frac{1}{N_{B_{close}}}\sum_{\lambda_b\in B_{close}}\frac{d(\lambda_b,\lambda_j)}{d(\lambda_C,\lambda_j)}\right] \tag{37}$$

   where $N_{B_{close}}$ is the cardinality of B_{close}.

3. Repeat step (2) until $N_{B_{close}} = \frac{N_B}{2}$.

[3] It must be noted that the Universal Background Model (UBM) can also be used to calculate $\mathcal{L}(X|\lambda_{\overline{C}})$ [46, 47].

[4] Also known as cohort models [15, 48].

Next, *maximally spread* models from the Ψ set are moved to set B_{far} using the following procedure:

1. Move the farthest model from Ψ to B_{far}.
2. Move λ_i from Ψ to B_{far}, where λ_i is found using:

$$\lambda_i = \arg\max_{\lambda_j \in \Psi} \left[\frac{1}{N_{B_{far}}} \sum_{\lambda_b \in B_{far}} d(\lambda_b, \lambda_j)\, d(\lambda_C, \lambda_j) \right] \quad (38)$$

   where $N_{B_{far}}$ is the cardinality of B_{far}.
3. Repeat step (2) until $N_{B_{far}} = \frac{N_B}{2}$.

Finally, $B = B_{close} \cup B_{far}$. The above procedures for selecting *maximally spread* models are required to reduce redundancy in the B set [44].

5 Error Measures

Since the verification system is inherently a two-class decision task, it follows that the system can make two types of errors. The first type of error is a False Acceptance (FA), where an impostor is accepted. The second error is a False Rejection (FR), where a true claimant is rejected. Thus the performance is measured in terms of False Acceptance rate (FA%) and False Rejection rate (FR%), defined as:

$$\text{FA\%} = \frac{I_A}{I_T} \times 100\% \quad (39)$$

$$\text{FR\%} = \frac{C_R}{C_T} \times 100\% \quad (40)$$

where I_A is the number of impostors classified as true claimants, I_T is the total number of impostor classification tests, C_R is the number of true claimants classified as impostors, and C_T is the total number of true claimant classification tests.

Since the errors are related, minimizing the FA% increases the FR% (and vice versa). The trade-off between FA% and FR% is adjusted using the threshold t in Eqn. (35). Depending on the application, more emphasis may be placed on one error over the other. For example, in a high security environment, it may be desired to have the FA% as low as possible, even at the expense of a high FR%.

The trade-off between FA% and FR% can be graphically represented by a Receiver Operating Characteristics (ROC) plot or a Detection Error Trade-off (DET) plot [11]. The ROC plot is on a linear scale, while the DET plot is on a quasi-log scale (which can improve the visual appearance of the curves). In both cases the FR% is plotted as a function of FA%.

To quantify the performance into a single number, Equal Error Rate (EER), is often used [15]. Here the system is configured to operate with FA% = FR%.

It must be noted that the threshold is adjusted to obtain desired performance on *test* data (data unseen by the system up to this point). Such a threshold is known as the *a posteriori* threshold. However, if the threshold is fixed before finding the performance, the threshold is known as the *a priori* threshold [14]. The *a priori* threshold can be found via experimental means using training data or *evaluation* data (data which has also not been seen by the system up to this point, but is separate from *test* data).

Logically, the *a priori* threshold is more realistic. However, it is often difficult to find a reliable *a priori* threshold [11, 14]. The *test* section of a database is often divided into two sets: *evaluation* data and true *test* data. If the *evaluation* data is not representative of the *test* data, then the *a priori* threshold will achieve significantly different results on *evaluation* and *test data*. Moreover, such a database division reduces the number of verification tests, thus decreasing the accuracy of the results. For these reasons, many researchers prefer to use the *a posteriori* threshold and interpret the performance obtained as the *expected* performance.

In keeping with tradition, the *a posteriori* threshold (set to obtain EER performance) is used in verification experiments in this communication.

6 Voice Activity Detector

In addition to pauses between words, the start and the end of speech signals in many databases often contains only background noise. Since these segments do not contain speaker dependent information, it would be advantageous to disregard them during modeling and testing. Decomposing a signal into speech and non-speech segments can be approximately accomplished via a Voice Activity Detector (VAD). Rather than using the heuristic energy based detector presented by Reynolds in [42] (seemingly used in his following work, i.e., [44, 45, 46, 47]) we have developed a parametric VAD based on the work by Haigh [19, 20].

The parametric VAD is implemented as follows. Each utterance is completely parameterized using a given feature extraction technique, resulting in a set of feature vectors, $X = \{\vec{x}_i\}_{i=1}^{N_V}$. A single Gaussian GMM (representing the background noise) is constructed using the first N_{noise} vectors[5]. Using the background noise GMM (λ_{noise}), the log-likelihood for each vector is found. If the log-likelihood for a given feature vector is below a predefined threshold (T_{VAD}), the vector is classified as containing speech. The following threshold has been experimentally found to provide good discrimination ability across various parameterization methods:

$$T_{\text{VAD}} = \frac{1}{3} l_{\text{noise}} \tag{41}$$

where

$$l_{\text{noise}} = \frac{1}{N_{\text{noise}}} \sum_{i=1}^{N_{\text{noise}}} \log p(\vec{x}_i | \lambda_{\text{noise}}) \tag{42}$$

The result of typical speech selection is shown in Figure 4.

A few words of caution: The VAD described here assumes that there is an initial part of the signal does not contain speech; moreover, for this VAD to work well, the background noise conditions have to be stationary during the duration of the speech utterance.

7 Evaluation of MACVs in Noisy Conditions

Speech signals were taken from the *test* section of the telephone speech NTIMIT database [24], which contains 10 utterances each from 168 persons (56 female and 112 male). The utterances have an average duration of approximately 4 seconds and have been degraded by the effects of a carbon button microphone and telephone line conditions (local and long-distance).

20 fixed persons (first 10 females and last 10 males, alpha-numerically sorted by subject ID) were selected to be the impostors; the remaining 148 persons were used as clients. As in [44], the BMS for each client was comprised of 10 models (N_Φ and N_Ψ were set to 20 (see Section 4.1)); the BMS was constructed by considering the other 147 client models. The first six sentences for each client were used for model training purposes, leaving the last four sentences for simulating true claimant tests. Impostor accesses were simulated using the last four utterances from each impostor. In total there were 592 (148 $\times$ 4) true claimant tests and 11840 (20 $\times$ 4 $\times$ 148) impostor tests. The decision threshold was set to obtain performance as close as possible to EER.

[5]For the NTIMIT database [24], $N_{\text{noise}} = 10$.

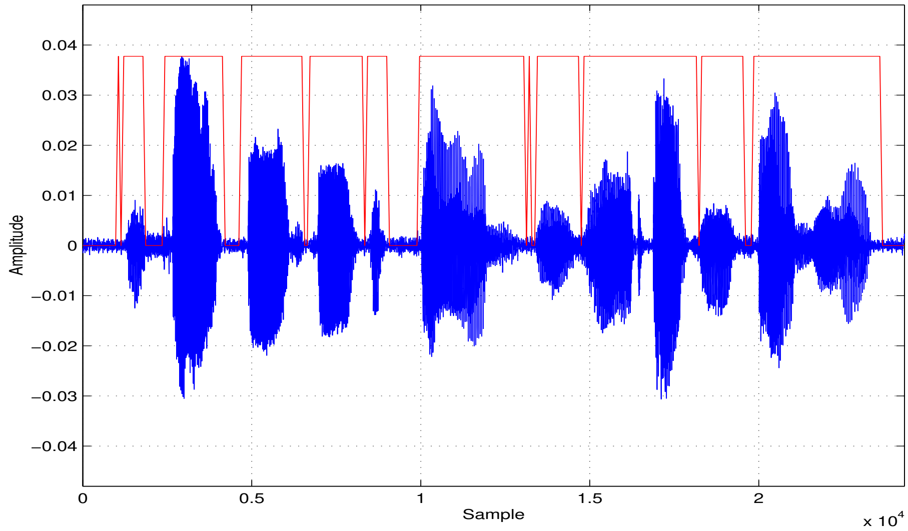

Figure 4: Typical result of speech selection using the parametric VAD. High level of the red line indicates the segments that have been selected as speech.

Number of Gaussians	1	2	4	8	16	32	64
EER (%)	14.28	12.73	11.73	9.96	9.58	9.99	11.16

Table 1: EER for varying number of Gaussians (MFCC parameterization)

In the first experiment we studied the effect of the number of Gaussians on verification performance while utilizing MFCC features. From the results shown in Table 1 it can be observed that the performance starts to level off at eight Gaussians. Increasing the number of Gaussians to 16 causes only minor performance gains. Further increases in the number of Gaussians reduces the performance, indicating that *overfitting* is occurring [12, 31]. Overfitting is said to occur when the classifier is “too tuned” to the training data, resulting in poor generalization on test data. Taking into account Occam’s Razor principle [12, 31], which in effect pleads for the simplest solution that provides adequate performance, the number of Gaussians in the second experiment was fixed at eight.

In the second experiment, the performance of each of the following feature sets was found: MFCC, CMS, MACV, MFCC+Δ, MFCC+Δ+MACV, CMS+Δ and CMS+Δ+MACV. A feature vector of type MFCC+Δ indicates that the MFCC feature vector ($\vec{x}$) has been concatenated with the feature vector containing delta versions of the MFCC features ($\Delta\vec{x}$). Similarly, MFCC+Δ+MACV indicates that the MACV feature set has also been appended.

Results were obtained for non-corrupted (clean) test utterances as well as for noisy test utterances where the SNR was varied from 28 dB to -8 dB. The utterances were corrupted by adding stationary white Gaussian noise, simulating background noise. The results are presented in Figures 5 through 7.

In Figure 5 it can be seen that the CMS features are the least affected by changes in the SNR, at the expense of slightly worse performance than MFCC features on clean speech (as expected; see Section 3.2). MFCC features are the most affected by noise, with rapid degradation in performance as the SNR is lowered. Performance of MACV features in clean and low noise conditions (SNR $>$ 16 dB) is not as good as for MFCC and CMS features, indicating that pitch and voicing information is not sufficient by itself to distinguish speakers. However, as the SNR drops to 16 dB and lower, MACVs perform better than MFCCs, suggesting that MACV features are more immune to the effects of noise.

In Figure 6 it can be observed that extending the MFCC feature vector with delta features reduces the performance degradation as the SNR is lowered. Extending the MFCC+Δ feature vector with MACV features obtains slightly better performance on clean speech and further reduces the performance degradation. However, by comparing Figures 6 and 7 it can be seen that CMS features obtain better performance than the MFCC+Δ+MACV feature set for SNRs of 16 dB and lower.

Figure 7 shows that extending the CMS feature vector with corresponding delta features causes only minor differences. Extending the CMS+Δ feature vector with MACV features alleviates some of the performance loss experienced by CMS features in clean conditions, and causes the performance in noisy conditions to be visibly improved up to a SNR of 4 dB.

These results thus support the conjecture described in Section 3.4, and suggest that use of the MACV feature set has beneficial effects on the performance of a verification system in noisy conditions.

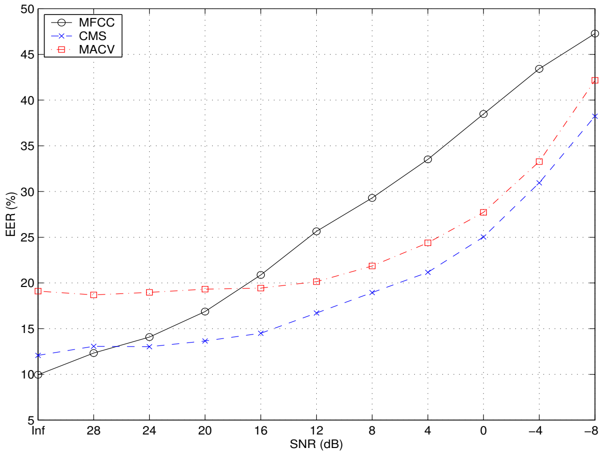

Figure 5: Performance of baseline features

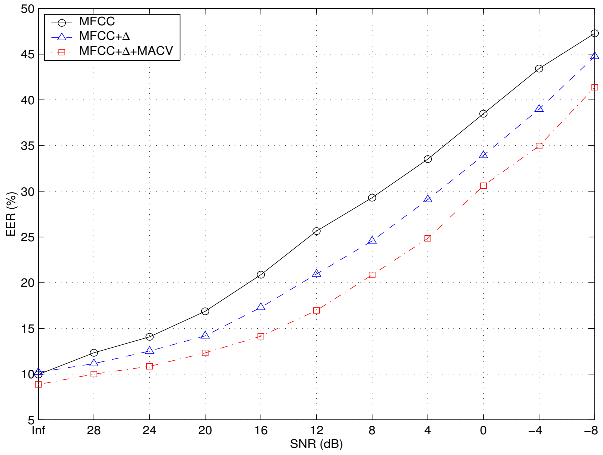

Figure 6: Performance of MFCC based features

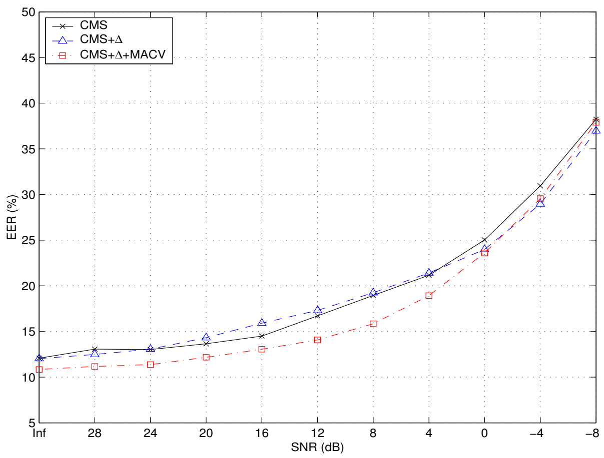

Figure 7: Performance of CMS based features

8 Summary

This communication first reviewed the human speech production process and feature extraction approaches used in a speaker verification system. Mel Frequency Cepstral Coefficients (MFCCs), delta (regression) features and Cepstral Mean Subtraction (CMS) were covered. A recently proposed feature set, termed Maximum Auto-Correlation Values (MACVs), which utilizes information from the source part of the speech signal, was also covered. A parametric Voice Activity Detector (VAD), used for disregarding silence and noise segments of the speech signal, was briefly described.

Experiments on the telephone speech NTIMIT database suggest that the performance degradation of a verification system used in noisy conditions can be reduced by extending MFCC or CMS feature vectors with MACV features.

9 Future Research

The experiments reported in this communication have used only stationary additive white Gaussian noise; it would be interesting to see if using MACV features would also help when using other noise types (for example, babble noise, car noise, etc.).

In speech recognition systems, it has been recently been shown that Spectral Subband Centroid (SSC) features [37] and biologically inspired Zero-Crossing with Peak Amplitude (ZCPA) features [26] are quite robust to the effects of additive noise. While the speaker verification task is significantly different from the speech recognition task, the SSC and ZCPA features may still contain person-dependent information; thus it would be interesting to evaluate their usefulness for robust person verification purposes.

10 Acknowledgments

The author thanks Professor Kuldip K. Paliwal (Griffith University, Australia) and Dr Andrew Morris (IDIAP, Switzerland) for their valuable suggestions. This work was partially supported by the Swiss National Science Foundation through the National Center of Competence in Research (NCCR) on Interactive Multimodal Information Management (IM2).

References

[1] B. S. Atal, *Automatic Speaker Recognition Based on Pitch Contours*, PhD Thesis, Polytechnic Institute of Brooklyn, 1968.

[2] B. S. Atal and M. R. Schroeder, "Predictive Coding of Speech Signals", *Report of the 6th International Congress on Acoustics*, Tokyo, 1968.

[3] B. S. Atal, "Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification", *Journal of the Acoustical Society of America*, Vol. 55, No. 6, 1974, pp. 1304-1312.

[4] W. Atkins, "A testing time for face recognition technology", *Biometric Technology Today*, Vol. 9, No. 3, 2001, pp. 8-11.

[5] R. Auckenthaler, M. Carey and H. Lloyd-Thomas, "Score Normalization for Text-Independent Speaker Verification Systems", *Digital Signal Processing*, Vol. 10, 2000, pp. 42-54.

[6] R. Balchandran, V. Ramanujam and R. Mammone, “Channel Estimation and Normalization by Coherent Spectral Averaging For Robust Speaker Verification”, *Proc. 6th European Conf. Speech Communication and Technology*, Budapest, 1999, pp. 755-758.

[7] J. A. Bilmes, “A Gentle Tutorial of the EM Algorithm and its Applications to Parameter Estimation for Gaussian Mixture and Hidden Markov Models”, Technical Report TR-97-021, International Computer Science Institute, Berkeley, California, 1998.

[8] M. M. Buechner, “Eye of the Beholder”, *Time Australia*, 27 November 2000 (No. 47), pp. 89-92.

[9] C. Champod and D. Meuwly, “The inference of identity in forensic speaker recognition”, *Speech Communication*, Vol. 31, No. 2-3, 2000, pp. 193-203.

[10] A.P. Dempster, N.M. Laird and D.B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm”, *J. Royal Statistical Soc., Ser. B*, Vol. 39, No. 1, 1977, pp. 1-38.

[11] G. R. Doddington, M. A. Przybycki, A. F. Martin and D. A. Reynolds, “The NIST speaker recognition evaluation - Overview, methodology, systems, results, perspective”, *Speech Communication*, Vol. 31, No. 2-3, 2000, pp. 225-254.

[12] R. O. Duda, P. E. Hart and D. G. Stork, *Pattern Classification*, John Wiley & Sons, USA, 2001.

[13] K. R. Farrell, “Text-Dependent Speaker Verification Using Data Fusion”, *Proc. IEEE International Conf. Acoustics, Speech and Signal Processing*, Detroit, Michigan, 1995, Vol. 1, pp. 349-352.

[14] S. Furui, “Cepstral Analysis Technique for Automatic Speaker Verification”, *IEEE Trans. Acoustics, Speech and Signal Processing*, Vol. 29, No. 2, 1981, pp. 254-272.

[15] S. Furui, “Recent Advances in Speaker Recognition”, *Pattern Recognition Letters*, Vol. 18, No. 9, 1997, pp. 859-872.

[16] J-L. Gauvain and C-H. Lee, “Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains”, *Proc. IEEE Trans. Speech and Audio Processing*, Vol. 2, No. 2, 1994. pp. 291-298.

[17] H. Gish and M. Schmidt, “Text-independent speaker identification”, *IEEE Signal Processing Magazine*, Vol. 11, No. 4, 1994, pp. 18-32.

[18] R. C. Gonzales and R. E. Woods, *Digital Image Processing*, Addison-Wesley, Reading, Massachusetts, 1993.

[19] J. A. Haigh and J. S. Mason, “A voice activity detector based on cepstral analysis”, *Proc. European Conf. Speech Communication and Technology*, 1993, Vol. 2, pp. 1103-1106.

[20] J. A. Haigh, *Voice Activity Detection for Conversational Analysis*, Masters Thesis, University of Wales, 1994.

[21] H. Hermansky, N. Morgan, A. Bayya and P. Kohn, “RASTA-PLP Speech Analysis Technique”, *Proc. IEEE International Conf. Acoustics, Speech and Signal Processing*, San Francisco, 1992, Vol. 1, pp. 121-124.

[22] H. Hermansky and N. Morgan, “RASTA Processing of Speech”, *IEEE Trans. Speech and Audio Processing*, Vol. 2, No. 4, 1994, pp. 578-589.

[23] X. Huang, A. Acero and H-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*, Prentice Hall PTR, New Jersey, 2001.

[24] C. Jankowski, A. Kalyanswamy, S. Basson and J. Spitz, "NTIMIT: A Phonetically Balanced, Continuous Speech Telephone Bandwidth Speech Database", *Proc. International Conf. Acoustics, Speech and Signal Processing*, Albuquerque, 1990, Vol. 1, pp. 109-112.

[25] N. L. Johnson and F. C. Leone, *Statistics and Experimental Design in Engineering and the Physical Sciences* (Vol. 1), John Wiley & Sons, USA, 1977.

[26] D-S. Kim, S-Y. Lee and R. M. Kil, "Auditory Processing of Speech Signals for Robust Speech Recognition in Real-World Noisy Environments", *IEEE Trans. Speech and Audio Processing*, Vol. 7, No. 1, 1999, pp. 55-69.

[27] Y. Linde, A. Buzo and R. M. Gray, "An Algorithm for Vector Quantization", *IEEE Trans. Communications*, Vol. 28, No. 1, 1980, pp. 84-95.

[28] J. Makhoul, "Spectral Analysis of Speech by Linear Prediction", *IEEE Trans. Acoustics Speech and Signal Processing*, Vol. 21, No. 3, 1973, pp. 140-148.

[29] J. Markowitz, "Speech systems work together in harmony", *Biometric Technology Today*, Vol. 9, No. 4, 2001, pp. 7-8.

[30] E. Messmer, "Pentagon lab may give biometrics needed boost", *CNN.com* web site (http://www.cnn.com/2001/TECH/science/03/20/pentagon.biometrics.idg/index.html), 20 March 2001.

[31] Tom M. Mitchell, *Machine Learning*, WCB/McGraw-Hill, USA, 1997.

[32] T. K. Moon, "Expectation-maximization Algorithm", *IEEE Signal Processing Magazine*, Vol. 13, No. 6, 1996, pp. 47-60.

[33] B. C. J. Moore, "Frequency Analysis and Masking", in: *Hearing* (editors: D. A. Eddins and D. M. Green), Academic Press, USA, 1995.

[34] B. C. J. Moore, "Information Extraction and Perceptual Grouping in the Auditory System", in: *Human and Machine Perception: Information Fusion* (editors: V. Cantoni, V. D. Gesù, A. Setti and D. Tegolo), Plenum Press, New York, 1997.

[35] J. A. Moorer, "The Optimum Comb Method of Pitch Period Analysis of Continuous Digitized Speech", *IEEE Trans. Acoustics, Speech and Signal Processing*, Vol. 22, No. 3, 1974, pp. 330-338.

[36] K. K. Paliwal, "Speech Processing Techniques" in: *Advances in Speech, Hearing and Language Processing* (editor: W. A. Ainsworth), Vol. 1, 1990, pp. 1-78.

[37] K. K. Paliwal, "Spectral Subband Centroid Features for Speech Recognition", *Proc. International Conf. on Acoustics, Speech and Signal Processing*, Seattle, Washington, 1998, Vol. 2, pp. 617-620.

[38] T. W. Parsons, *Voice and Speech Processing*, McGraw-Hill, USA, 1987.

[39] J. Picone, "Signal Modeling Techniques in Speech Recognition", *Proceedings of the IEEE*, Vol. 81, No. 9, 1993, pp. 1215-1247.

[40] W. H. Press, S. A. Teukolsky, W. T. Vetterling and B. P. Flannery, *Numerical Recipes in C: the Art of Scientific Computing*, Cambridge University Press, Cambridge, 1992.

[41] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing: Principles, Algorithms, and Applications* (3rd ed.), Prentice Hall, USA, 1996.

[42] D. A. Reynolds, “A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification”, *Technical Report 967*, Lincoln Laboratory, Massachusetts Institute of Technology, 1993.

[43] D. A. Reynolds, “Experimental Evaluation of Features for Robust Speaker Identification”, *IEEE Trans. Speech and Audio Processing*, Vol. 2, No. 4, 1994, pp. 639-643.

[44] D. A. Reynolds, “Speaker Identification and Verification Using Gaussian Mixture Speaker Models”, *Speech Communication*, Vol. 17, No. 1-2, 1995, pp. 91-108.

[45] D. A. Reynolds, “Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models”, *IEEE Trans. Speech and Audio Processing*, Vol. 3, No. 1, 1995, pp. 72-83.

[46] D. A. Reynolds, “Comparison of Background Normalization Methods for Text-Independent Speaker Verification”, *Proc. 5th European Conf. Speech Communication and Technology*, Rhodes, Greece, 1997, Vol. 2, pp. 963-966.

[47] D. Reynolds, T. Quatieri and R. Dunn, “Speaker Verification Using Adapted Gaussian Mixture Models”, *Digital Signal Processing*, Vol. 10, No. 1-3, 2000, pp. 19-41.

[48] A. E. Rosenberg, J. DeLong, C-H. Lee, B-H. Juang and F. K. Soong, “The Use of Cohort Normalized Scores for Speaker Verification”, *Proc. International Conf. Spoken Language Processing*, Alberta, 1992, Vol. 1, pp. 599-602.

[49] M. J. Ross, “Average Magnitude Difference Function Pitch Extractor”, *IEEE Trans. Acoustics, Speech and Signal Processing*, Vol. 22, No. 5, 1974, pp. 353-362.

[50] C. Sanderson, “Information Fusion and Person Verification Using Speech & Face Information”, IDIAP Research Report, IDIAP-RR 02-33, Martigny, Switzerland, September 2002.

[51] C. Sanderson and K. K. Paliwal, “Noise compensation in a person verification system using face and multiple speech features”, *Pattern Recognition*, Vol. 36, No. 2, 2003, pp. 293-302.

[52] D. O’Shaughnessy, *Speech communications: human and machine* (2nd ed.), IEEE Press, New York, 2000.

[53] B. Sonesson, “The functional anatomy of the speech organs” in: *Manual of Phonetics* (editor: B. Malmberg), North-Holland, Amsterdam, 1968, pp. 45-75.

[54] F. K. Soong and A. E. Rosenberg, “On the Use of Instantaneous and Transitional Spectral Information in Speaker Recognition”, *IEEE Trans. Acoustics, Speech and Signal Processing*, Vol. 36, No. 6, 1988, pp. 871-879.

[55] B. Wildermoth and K. K. Paliwal, “Use of Voicing and Pitch Information for Speaker Recognition”, *Proc. 8th Australian International Conf. Speech Science and Technology*, Canberra, 2000, pp. 324-328.

[56] J. D. Woodward, “Biometrics: Privacy’s Foe or Privacy’s Friend?”, *Proceedings of the IEEE*, Vol. 85, No. 9, 1997, pp. 1480-1492.