PERGAMON

Pattern Recognition 36 (2003) 293–302

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

Noise compensation in a person verification system using face and multiple speech features

Conrad Sanderson*, Kuldip K. Paliwal

School of Microelectronic Engineering, Griffith University, Brisbane, QLD 4111, Australia

Received 21 December 2001

Abstract

In this paper, we demonstrate that use of a recently proposed feature set, termed Maximum Auto-Correlation Values, which utilizes information from the source part of the speech signal, significantly improves the robustness of a text independent identity verification system. We also propose an adaptive fusion technique for integration of audio and visual information in a multi-modal verification system. The proposed technique explicitly measures the quality of the speech signal, adjusting the amount of contribution of the speech modality to the final verification decision. Results on the VidTIMIT database indicate that the proposed approach outperforms existing adaptive and non-adaptive fusion techniques. For a wide range of audio SNRs, the performance of the multi-modal system utilizing the proposed technique is always found to be better than the performance of the face modality.

Keywords: Multi-modal; Audio-visual; Identity verification; Adaptive fusion; Source features

1. Introduction

Access control systems are becoming an increasingly important part of our life. As an example, Automatic Teller Machines (ATMs) employ a simple identity verification where the user is asked to enter their Personal Identification Number (PIN), known only to the user, after inserting their ATM card. If the PIN matches the one prescribed to the card, the user is allowed access to their bank account. Similar verification systems are widely employed to restrict access to rooms and buildings.

The verification system such as the one used in the ATM only verifies the validity of the combination of a certain possession (in this case, the ATM card) and certain knowledge (the PIN). The ATM card can be lost or stolen, and the PIN can be compromised (e.g. somebody looks over your shoulder while you're entering the PIN). Hence new verification methods have emerged, where the PIN can be replaced by, or used in addition to, biometrics such as the person's speech, face image or fingerprints. The use of biometrics is attractive since they cannot be lost or forgotten and vary significantly between people. The basic operation of a speech-based verification system is as follows:

(1) A claim for an identity is presented along with a supporting sample of the person's speech.
(2) The system extracts person-dependent information (known as feature extraction) from the speech and compares it against a model of features from the person whose identity is being claimed. Let us refer to the result of this comparison as the client likelihood, L_C.
(3) The system also compares the information against a model of possible impostors. Let us refer to the result of this comparison as the impostor likelihood, L_I.
(4) An opinion, O, on the claim is found using $O = L_C/L_I$. A relatively high opinion indicates the person is a true claimant, while a relatively low opinion suggests the person is an impostor.

* Corresponding author. Tel.: +61-7-3875-6578; fax: +61-7-3875-5198.

E-mail addresses: c.sanderson@me.gu.edu.au (C. Sanderson), k.paliwal@me.gu.edu.au (K.K. Paliwal).

0031-3203/02/$22.00
PII: S0031-3203(02)00031-6

(5) The opinion is thresholded to achieve the final decision to either accept or reject the claim.

The performance of a verification system is measured in terms of False Acceptance rate (FA%) and False Rejection rate (FR%), defined as

$$\mathrm{FA} = \frac{I_A}{I_T}100\%, \quad \mathrm{FR} = \frac{C_R}{C_T}100\%,$$

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. To quantify the performance into a single number, Equal Error Rate (EER) is often used. Here the system is configured so the false acceptance rate is equal to the false rejection rate.

Speech-based verification systems fall into two categories: text dependent and text independent. In text-dependent systems, the claimant must recite a phrase specified by the system. In text independent systems, the claimant can say whatever he or she wishes. In this paper we will concentrate on the latter category.

The speech signal can be thought of as being composed of two parts [1]:

1. *The source part.* Here the source signal may be either periodic, resulting in voiced speech, or noisy and aperiodic, causing unvoiced speech. The periodic signal is generated in the larynx where periodic opening and closing of the vocal cord determines the pitch of the voice.
2. *The system part.* Here the source signal is filtered by the vocal tract. The vocal tract includes the oral cavity which is continuously changing due to the movement of articulators, such as the tongue, jaw, lips, etc.

Popular speech-based verification systems use information from the system part in the form of an instantaneous spectrum represented by Mel Frequency Cepstral Coefficients (MFCCs). Verification systems using MFCC features have proven to be quite effective [2]. However, their performance easily degrades in the presence of a mismatch between training and testing conditions. Usually this is in the form of a channel distortion and/or ambient noise. There are two popular techniques to reduce the effects of mismatch: use of delta (regression) features [3] and Cepstral Mean Subtraction (CMS) [4].

Recently Wildermoth and Paliwal [5] proposed a new feature set, termed Maximum Auto-Correlation Values (MACV), which utilizes information from the source part. As we will show, use of MACV features significantly increases the robustness of a speech-based verification system.

An alternative method to achieve increased robustness (and higher performance) is to use features from both speech and face images. It is also possible to use biometrics such as the iris, fingerprints and hand geometry [6,7]. A system employing more than one biometric is known as multi-modal verification system [8].

The crucial part of a multi-modal system is the fusion technique that combines the separate sources of information. There are two classes of fusion: non-adaptive and adaptive. In non-adaptive fusion, the degree of contribution of information from speech and face modalities to the overall decision process is fixed. Conversely, in adaptive fusion, the degree of contribution is varied according to the current reliability of each modality. For example, the contribution of speech information is lowered when the Signal-to-Noise Ratio (SNR) is poor.

The performance of a multi-modal verification system should at all times, be better than, or at worst, be equal to the best corresponding single-modality system. *Catastrophic fusion* is said to occur when the performance is worse than the best corresponding single-modality system [9].

In this paper we propose an adaptive fusion technique where the resulting performance is never worse than that of the underlying modalities. Its performance is compared against existing adaptive and non-adaptive fusion techniques.

The rest of the paper is organized as follows. In Section 2 we describe the MFCC, CMS, delta and MACV speech feature extraction techniques; we also describe the eigenfaces approach, where facial features are derived from Principal Component Analysis (PCA). In Section 3 we describe a Gaussian Mixture Model (GMM) classifier which shall be used as the basis for experiments. In Section 4 we describe several popular fusion techniques as well as the proposed technique. Section 5 is devoted to experiments evaluating the use of MACV features to reduce the effects of mismatched conditions in a speech-based verification system. In Section 6 we evaluate the performance of the presented fusion techniques in a multi-modal verification system.

To keep consistency with traditional matrix notation, image sizes are described using the number of rows first, followed by the number of columns, e.g. an image of size $Y \times X$ has Y rows and X columns.

2. Feature extraction methods

2.1. MFCC features

The human ear processes the speech signal using a bank of non-uniformly spaced filters [10]. Features extracted using such a filter-bank have been shown to be effective for speaker verification [2].

The speech signal is analyzed on a frame by frame basis, with a frame length of 20 ms and a frame advance of 10 ms. Hence for each second of speech we extract features from 100 frames. Each frame is multiplied by a Hamming window and the spectrum is obtained using the Fast Fourier Transform (FFT) algorithm. The square of the magnitude of the spectrum is taken. Seventeen Mel-scale

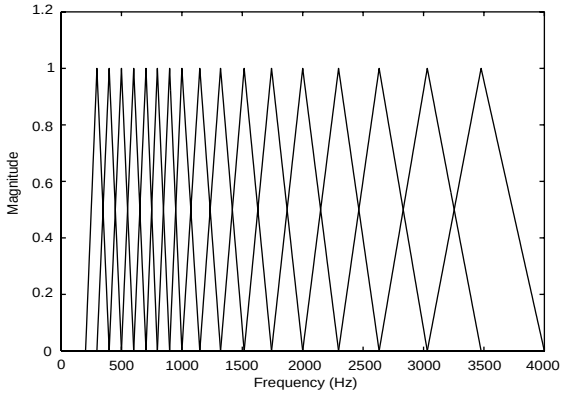

Fig. 1. Mel-scale filterbank.

triangular filter-bank energies [11] are then calculated and are expressed on a logarithmic scale [12]. The frequency range of the filters was chosen to cover the telephone bandwidth—the central frequencies are (in Hz): 300, 400, 500, 600, 700, 800, 900, 1000, 1149, 1320, 1516, 1741, 2000, 2297, 2639, 3031 and 3482. The upper and lower passband frequencies of each filter are the center frequencies of the adjacent filters. A graphical representation of the filters is shown in Fig. 1.

Since the filter bank coefficients are highly correlated, a discrete cosine transform is used to de-correlate them:

$$c_i = \frac{1}{N_F}\sum_{j=1}^{N_F} f_j \cos\left[\frac{\pi i}{N_F}(j-0.5)\right],$$

$$i = 0, 1, \ldots, N_F - 1, \tag{1}$$

where N_F is the number of filters and f_j are the log filter-bank energies. c_0 is not used since it represents the average value of the spectrum and hence is susceptible to varying background noise. Therefore, the MFCC feature vector is constructed using

$$\vec{c} = [c_1 \quad c_2 \quad \ldots \quad c_{N_F-1}]^{\mathrm{T}}, \tag{2}$$

where T denotes the transpose operation.

2.2. CMS features

It has been shown that by applying CMS to MFCC features results in new features which are significantly more immune to the effects of channel distortion [4]. For the sake of convenience, we shall refer to MFCC-CMS features simply as CMS features.

Given a sequence of MFCC feature vectors from a speech utterance, $\{\vec{c}_i,\ i = 0, 1, \ldots, N_V - 1\}$, we define their mean as $\vec{c}_\mu$. The mean is assumed to represent the cepstrum of the channel [13]. Thus the sequence of CMS feature vectors is obtained using

$$\vec{d}_i = \vec{c}_i - \vec{c}_\mu, \quad i = 0, 1, \ldots, N_V - 1. \tag{3}$$

It must be noted that the cepstral mean also contains the average speech cepstrum, which contains speaker information [13,14]. Thus removal of $\vec{c}_\mu$ from MFCC features is a double-edged sword: on one hand it makes the verification system more robust against channel mismatches, while on the other it reduces the accuracy of the system in clean conditions.

2.3. Delta features

It has been shown that use of transitional spectral information in addition to the instantaneous spectrum increases robustness in speaker recognition systems [3]. Given a sequence of feature vectors from a speech utterance, $\{\vec{c}_i,\ i = 0, 1, \ldots, N_V - 1\}$, their corresponding delta representations are calculated using a first-order orthogonal polynomial fit

$$\Delta\vec{c}_i = \frac{\sum_{k=-K}^{K} k\vec{c}_{i+k}}{\sum_{k=-K}^{K} k^2} \quad \text{for } i = K \text{ to } N_V - 1 - K \tag{4}$$

and

$$\Delta\vec{c}_i = \Delta\vec{c}_K \quad \text{for } i = 0 \text{ to } K - 1, \tag{5}$$

$$\Delta\vec{c}_i = \Delta\vec{c}_{N_V-1-K} \quad \text{for } i = N_V - K \text{ to } N_V - 1. \tag{6}$$

2.4. MACV features

Given a speech frame $\{s(n),\ n = 0, 1, \ldots, N_S - 1\}$ the auto-correlation function is defined as [15,16]

$$R(k) = \frac{1}{N_S}\sum_{n=0}^{N_S-1-k} s(n)s(n+k), \quad k = 0, 1, \ldots, N_S - 1. \tag{7}$$

If $\{s(n)\}$ is periodic with a period equal to P samples, then $\{R(k)\}$ will show a peak at a lag equal to P. Valid pitch lags are approximately between 2 and 16 ms. Assuming $\{s(n)\}$ contains voiced speech, the period of $\{s(n)\}$ can be found by searching for the maximum in $\{R(k)\}$ in the 2–16 ms range [17]. However, current methods of detecting whether a given speech frame is voiced or unvoiced are unreliable—leading to pitch estimation errors.

The MACV feature set [5] overcomes this problem by finding, for each speech frame, an M-point approximation of the auto-correlation function. This is done as follows:

(1) Compute the auto-correlation function $\{R(k)\}$.
(2) Normalize $\{R(k)\}$ by its maximum, i.e., $\hat{R}(k) = \frac{R(k)}{R(0)},\ k = 0, 1, \ldots, N_S - 1$.
(3) Divide the higher portion (from 2 to 16 ms) of $\{\hat{R}(k)\}$ into M equal parts.
(4) Find the maximum value of each of the M parts.
(5) The M MACVs form an M-dimensional feature vector.

A conceptual block diagram of this process is shown in Fig. 2. It must be noted the MACV feature set also contains voicing information, i.e. whether the current frame is voiced or unvoiced.

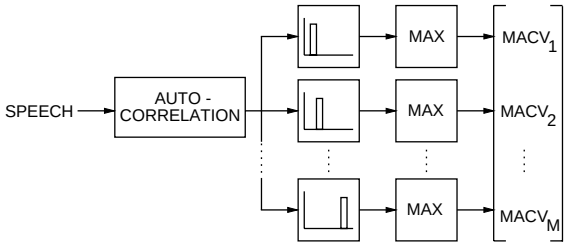

Fig. 2. MACV feature extractor (after Ref. [5]).

2.5. Principal component analysis

A face image can be represented by a matrix containing grey level pixel values. A straightforward approach of representing facial information with a feature vector is to simply concatenate all the columns of the matrix. However, even for a low resolution face image, the resulting feature vector is of prohibitive dimensionality. For example, given a 56×64 pixel face image, the resulting feature vector has 3584 dimensions. Hence dimensionality reduction methods, such as the PCA approach have been used [18].

The features are obtained as follows. Given a face image matrix F of size $Y \times X$, we construct a vector representation by concatenating all the columns of F to form a column vector $\vec{f}$ of dimensionality YX. Given a set of training vectors $\{\vec{f}_i,\ i = 1, 2, \ldots, N_P\}$ for all persons, we define the mean of the training set as $\vec{f}_\mu$. A new set of mean subtracted vectors is formed using

$$\vec{g}_i = \vec{f}_i - \vec{f}_\mu, \quad i = 1, 2, \ldots, N_P. \tag{8}$$

The mean subtracted training set is represented as matrix $\boldsymbol{G} = [\vec{g}_1\ \vec{g}_2\ \ldots\ \vec{g}_{N_P}]$. The covariance matrix is calculated using

$$\boldsymbol{C} = \boldsymbol{G}\boldsymbol{G}^{\mathrm{T}}. \tag{9}$$

Due to the size of $\boldsymbol{C}$, calculation of the eigenvectors of $\boldsymbol{C}$ can be computationally infeasible. However, if the number of training vectors (N_P) is less than their dimensionality (YX), there will be only $N_P - 1$ meaningful eigenvectors. Turk and Pentland [18] exploit this fact to determine the eigenvectors using an alternative method, summarized as follows. Let us denote the eigenvectors of matrix $\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G}$ as $\vec{v}_j$ with corresponding eigenvalues λ_j:

$$\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G}\vec{v}_j = \lambda_j \vec{v}_j. \tag{10}$$

Pre-multiplying both sides by $\boldsymbol{G}$ gives us

$$\boldsymbol{G}\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G}\vec{v}_j = \lambda_j \boldsymbol{G}\vec{v}_j. \tag{11}$$

Letting $\vec{u}_j = \boldsymbol{G}\vec{v}_j$ and substituting for $\boldsymbol{C}$ from Eq. (9)

$$\boldsymbol{C}\vec{u}_j = \lambda_j \vec{u}_j. \tag{12}$$

Hence the eigenvectors of $\boldsymbol{C}$ can be found by pre-multiplying the eigenvectors of $\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G}$ by $\boldsymbol{G}$. To achieve dimensionality reduction, let us construct matrix $\boldsymbol{U} = [\vec{u}_1\ \vec{u}_2\ \ldots\ \vec{u}_D]$, containing D eigenvectors of $\boldsymbol{C}$ with largest corresponding eigenvalues. Here, $D < N_P$. A feature vector $\vec{x}$ of dimensionality D is then derived from a face vector $\vec{f}$ using

$$\vec{x} = \boldsymbol{U}^{\mathrm{T}}(\vec{f} - \vec{f}_\mu), \tag{13}$$

i.e. face vector $\vec{f}$ decomposed in terms of D eigenvectors.

3. GMM classifier

The distribution of feature vectors for each person is modeled by a GMM. Given a set of training vectors, an N_M-mixture GMM is trained using a k-means clustering algorithm followed by 10 iterations of the Expectation Maximization (EM) algorithm [19].

Given a claim for person C's identity and a set of feature vectors $X = \{\vec{x}_i,\ i = 1, 2, \ldots, N_V\}$ supporting the claim, log likelihood of the claimant being the true claimant is calculated using

$$\log p(X|\lambda_C) = \frac{1}{N_V}\sum_{i=1}^{N_V} \log\left[p(\vec{x}_i|\lambda_C)\right], \tag{14}$$

where

$$p(\vec{x}|\lambda) = \sum_{j=1}^{N_M} m_j \mathcal{N}(\vec{x}; \vec{\mu}_j, \boldsymbol{\Sigma}_j) \tag{15}$$

and

$$\lambda = \{m_j, \vec{\mu}_j, \boldsymbol{\Sigma}_j,\ j = 1, 2, \ldots, N_M\}. \tag{16}$$

Here λ_C is the model for person C. N_M is the number of mixtures, m_j is the weight for mixture j (with constraint $\sum_{j=1}^{N_M} m_j = 1$), and $\mathcal{N}(\vec{x}; \vec{\mu}, \boldsymbol{\Sigma})$ is a multi-variate Gaussian function with mean $\vec{\mu}$ and diagonal covariance matrix $\boldsymbol{\Sigma}$:

$$\mathcal{N}(\vec{x}; \vec{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left[\frac{-1}{2}(\vec{x} - \vec{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\vec{x} - \vec{\mu})\right], \tag{17}$$

where D is the dimensionality of $\vec{x}$. Given a set $\{\lambda_b,\ b = 1, 2, \ldots, B\}$ of B background person models (also known as cohorts [20]) for person C, the log likelihood of the claimant being an impostor is found using

$$\log p(X|\lambda_{\bar{C}}) = \log\left[\frac{1}{B}\sum_{b=1}^{B} p(X|\lambda_b)\right]. \tag{18}$$

The set of background person models is found using the method described in Ref. [2]. An opinion on the claim is found using the following likelihood ratio:

$$o = \frac{p(X|\lambda_C)}{p(X|\lambda_{\bar{C}})}. \tag{19}$$

In the log domain this becomes

$$O = \log p(X|\lambda_C) - \log p(X|\lambda_{\bar{C}}). \tag{20}$$

The verification decision is reached as follows: given a threshold t, the claim is accepted when $O \geqslant t$; the claim is rejected when $O < t$. A conceptual block diagram of a verification system employing the GMM is shown in Fig. 3.

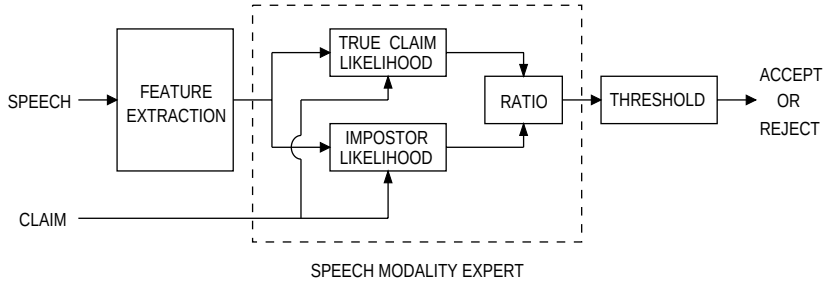

Fig. 3. Conceptual block diagram of a verification system.

4. Fusion of audio and visual information

4.1. Feature vector concatenation

In this fusion approach, the basic idea is to concatenate the speech and face feature vectors to form a new feature vector. However, before concatenation can be done, the frame rates from the speech and face feature extractors must match. Recall that the frame rate for speech features is 100 frames per second (fps) while the standard frame rate for video is 25 fps (using off the shelf commercial video cameras). A straightforward approach to match the frame rates is to artificially increase the video frame rate and generate the missing frames by copying original frames.

We shall refer to this fusion approach as *concatenation fusion*. This type of fusion is also known in the literature as *pre-categorical integration* [21] and *early integration* [9].

4.2. Opinion fusion

4.2.1. Non-adaptive

In opinion fusion, each modality is processed independently by a *modality expert* which produces an opinion on the claim. A modality expert is the GMM classifier described in Section 3 without the final thresholding stage.

The opinions from Ψ modality experts then form a Ψ-dimensional opinion vector which is used by a *decision stage*. Since there are only two possible outcomes (accept or reject), the decision stage can be a binary classifier [8]. The classifier is trained with example opinions for known impostor and true claims. It then classifies a given opinion vector as belonging to either the impostor or true claimant class. This type of fusion technique is also known in the literature as *post-categorical integration* [21] and *late integration* [9].

Many different binary classifiers can be used for the opinion fusion approach [8]—here we use a linear combination approach, described as follows. The opinion O_i for modality i is first normalized to the $[0,1]$ interval using a sigmoïd

$$P_i = \frac{1}{1+\exp[-\tau_i(O_i)]}, \tag{21}$$

where

$$\tau_i(O_i) = \frac{O_i - (\mu_i - 2\sigma_i)}{2\sigma_i}. \tag{22}$$

Here μ_i and σ_i are the mean and the standard deviation of opinions for true claims for modality i, respectively.

Assuming the opinions for true and impostor claims follow Gaussian distributions $\mathcal{N}(O_i;\mu_i,\sigma_i^2)$ and $\mathcal{N}(O_i;\mu_i - 4\sigma_i,\sigma_i^2)$, respectively, 95% of the value lie in the $[\mu_i - 2\sigma_i, \mu_i + 2\sigma_i]$ and $[\mu_i - 6\sigma_i, \mu_i - 2\sigma_i]$ intervals, respectively. Eq. (22) maps the opinions to the $[-2,2]$ interval, which corresponds to the approximately linear portion of the sigmoïd in Eq. (21). The sigmoïd is necessary to take care of outliers and situations where the assumptions do not hold entirely. The normalized opinions are then fused using

$$F = \sum_{i=1}^{\Psi} w_i P_i, \tag{23}$$

where w_i is the weight for modality i, with the constraint $\sum_{i=1}^{\Psi} w_i = 1$. The normalization to the $[0,1]$ interval is required to ensure opinions from all modalities are equally represented—this prevents any modality from dominating the fused opinion prior to weighting. The weight for each modality expert (also in the $[0,1]$ interval) can then be selected according to its discrimination ability and robustness.

Given a threshold t, the claim is accepted when $F \geqslant t$ (i.e. true claimant); the claim is rejected when $F < t$ (i.e. impostor). It must be noted that this fusion approach results in a linear decision boundary in Ψ-dimensional space [22].

4.2.2. Proposed adaptive method

As will be shown in Section 5, out of all the presented speech features, the MFCC features are most susceptible to changes in the SNR (preliminary experiments have also shown that they are also the most consistently affected features). We exploit this to detect the amount of mismatch between the training and testing conditions, and hence select the weight of the speech modality accordingly.

A global background noise GMM (λ_{noise}) is constructed using the first N_{noise} MFCC feature vectors from all training speech signals. The first N_{noise} vectors, $\{\vec{x}_i,\ i = 0,1,\ldots,N_{noise}-1\}$ are assumed to contain only the

background noise. During a claim, a quality measure for a given speech signal is found using

$$q = \frac{1}{N_{noise}} \sum_{i=0}^{N_{noise}-1} \log[p(\vec{x}_i|\lambda_{noise})]. \quad (24)$$

The larger the difference between the training and testing conditions, the lower q is going to be. We convert q to a posterior weight for the speech modality using

$$v_1 = \frac{1}{1+\exp[-a(q-b)]}, \quad (25)$$

where a and b are prior knowledge on how q changes according to the amount of mismatch. Given an a priori weight u_1, the final weight for the speech modality is found using

$$w_1 = u_1 v_1. \quad (26)$$

The corresponding weight for the video modality is then found simply using

$$w_2 = 1 - w_1. \quad (27)$$

4.2.3. Wark's adaptive method

In Ref. [9], Wark proposed an adaptive method to fuse the opinions in a multi-modal system. We summarize the method as follows. Given normalized opinions P_1 and P_2 from the speech and face modality experts respectively, the fused opinion is found using

$$F = \left[\frac{\zeta_2}{\zeta_1+\zeta_2}\right]\left[\frac{\kappa_1}{\kappa_1+\kappa_2}\right]P_1 + \left[\frac{\zeta_1}{\zeta_1+\zeta_2}\right]\left[\frac{\kappa_2}{\kappa_1+\kappa_2}\right]P_2, \quad (28)$$

where

$$\zeta_1 = \sqrt{\frac{\sigma_{i,C}^2}{N_C} + \frac{\sigma_{i,I}^2}{N_I}} \quad (29)$$

and

$$\kappa_i = \frac{|\mathscr{M}(P_i)_{i,C} - \mathscr{M}(P_i)_{i,I}|}{\mu_{i,C}}. \quad (30)$$

Here, for modality i, ζ_i is the a priori confidence (found during training), while κ_i is the a posteriori confidence (found during testing). N_C and N_I are the number of opinions for true and impostor claims, respectively. $\mathscr{M}(P_i)_{i,C} = (P_i - \mu_{i,C})/\sigma_{i,C}^2$ is the one-dimensional Mahalanobis distance, where $\mu_{i,C}$ and $\sigma_{i,C}^2$ are the mean and variance of opinions for true claims. Similarly, $\mathscr{M}(P_i)_{i,I} = (P_i - \mu_{i,I})/\sigma_{i,I}^2$ where $\mu_{i,I}$ and $\sigma_{i,I}^2$ are the mean and variance of opinions for impostor claims.

Under clean conditions, the Mahalanobis distance between a given opinion for a true claim and the model of opinions for true claims (i.e. represented by $\mu_{i,C}$ and $\sigma_{i,C}^2$) should be small. Similarly, the distance between a given opinion for a true claim and the model of opinions for impostor claims should be large. Vice versa applies for a given opinion for an impostor claim. Hence under clean conditions, κ_i should be large. Wark argued that under noisy conditions, the distances should decrease, hence κ_i should decrease.

By letting $\eta = 1/\zeta$, Eq. (28) becomes

$$F = \left[\frac{\eta_1}{\eta_1+\eta_2}\right]\left[\frac{\kappa_1}{\kappa_1+\kappa_2}\right]P_1 + \left[\frac{\eta_2}{\eta_1+\eta_2}\right]\left[\frac{\kappa_2}{\kappa_1+\kappa_2}\right]P_2. \quad (31)$$

We can now generalize the above equation for Ψ modalities

$$F = \sum_{i=1}^{\Psi}\left[\frac{\eta_i}{\sum_{j=1}^{\Psi}\eta_j}\right]\left[\frac{\kappa_i}{\sum_{j=1}^{\Psi}\kappa_j}\right]P_i. \quad (32)$$

Hence we define Wark's weight for modality i as

$$\phi_i = \left[\frac{\eta_i}{\sum_{j=1}^{\Psi}\eta_j}\right]\left[\frac{\kappa_i}{\sum_{j=1}^{\Psi}\kappa_j}\right]. \quad (33)$$

We can then use ϕ_i in place of w_i in Eq. (23) by placing a $\sum_{i=1}^{\Psi}\phi_i = 1$ constraint.

5. Performance of speech features

5.1. Experimental setup

In this section we evaluate the performance of a speech-based verification system using different feature sets while varying the SNR. The speech utterances were corrupted by adding white Gaussian noise. The experiments were done on the ubiquitous telephone speech NTIMIT database [23]. As in Ref. [2], only the *test* section of the database is used, containing 10 utterances from each of the 168 persons (56 female and 112 male). The first six utterances (sorted alpha-numerically by filename) were used for training the models, while the last four were used for testing purposes.

In order to reduce detecting and modeling the environment rather than the speaker, a parametric voice activity detector (VAD) is used [24]. Each utterance is completely parameterized using a given feature extraction technique. The VAD then builds a 1 mixture GMM of the background noise using the first 10 feature vectors. For each feature vector the log-likelihood is calculated. If the log-likelihood falls below a predefined threshold, the feature vector is classified as containing speech. Only feature vectors containing speech are used for training and testing purposes.

Sixteen mixture GMMs were used as client models. Four test utterances, each from 20 fixed persons (10 male and 10 female) were used for simulating impostor accesses against the remaining 148 persons. As in Ref. [2], 10 background person models were used for the impostor likelihood calculation. For each of the remaining 148 persons, their four test utterances were used separately as true claims. In total there were 11,840 impostor and 592 true claims. The decision threshold was then set for EER performance.

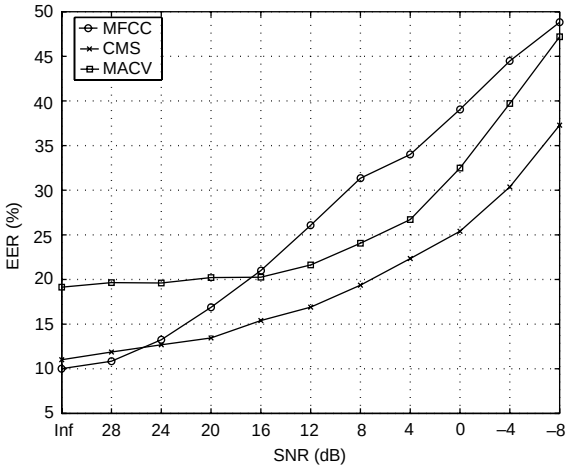

Fig. 4. Performance of baseline features.

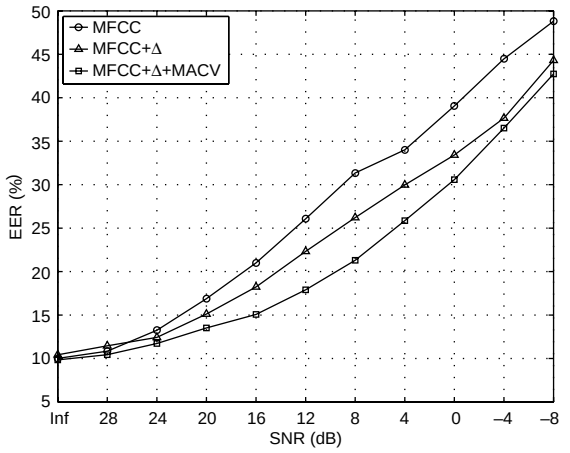

Fig. 5. Performance of MFCC-based features.

Performance of the following feature sets was found: MFCC, CMS, MACV, MFCC + Δ, MFCC + Δ + MACV, CMS + Δ and CMS + Δ + MACV.

As in Ref. [25], we have used $K = 2$ for delta features. A feature vector of type MFCC + Δ indicates that the MFCC feature vector ($\vec{c}$) has been concatenated with the feature vector containing delta versions of the MFCC features ($\Delta\vec{c}$). Thus the concatenated vector is expressed as

$$\vec{e} = [\vec{c}^{\mathrm{T}} \quad \Delta\vec{c}^{\mathrm{T}}]^{\mathrm{T}}. \tag{34}$$

Similarly, MFCC + Δ + MACV indicates that the feature vector $\vec{e}$ has been concatenated with the MACV feature vector. For MACV features, $M = 5$ was found to be optimal in preliminary experiments. Results were obtained for non-corrupted speech as well as for corrupted speech where the SNR was varied from 28 to −8 dB. The results are presented in Figs. 4–6.

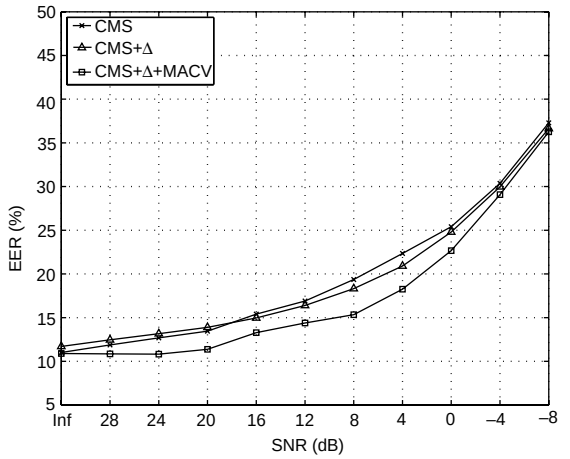

Fig. 6. Performance of CMS-based features.

5.2. Discussion

In Fig. 4 we can see that the CMS features are the most immune to changes in the SNR, at the expense of slightly worse performance than MFCC features on clean speech. MFCC features are the most affected by noise, with rapid degradation in performance as the SNR is lowered. Performance of MACV features is the worst up to SNR of 20 dB, indicating that pitch and voicing information is not sufficient by itself to distinguish speakers. However from 16 dB onwards MACVs are better than MFCCs.

In Fig. 5 we can observe that extending the MFCC feature vector with deltas reduces the performance degradation as the SNR is lowered. Extending the MFCC + Δ feature vector with MACV features reduces the performance degradation even further. However, it must be noted that CMS features obtain better performance than MFCC + Δ + MACV features from 12 dB onwards.

Fig. 6 shows that extending the CMS feature vector with deltas causes only minor differences. However, extending the CMS + Δ feature vector with MACV features significantly improves the performance.

Based on these results we conclude that use of MACV features has beneficial effects on the performance of a verification system in noisy conditions.

6. Fusion of audio and video information

6.1. VidTIMIT audio–visual database

With the help of many volunteers, we have created an audio–visual database used for experiments in multi-modal identity verification. It is comprised of video and corresponding audio recordings of 43 people (19 female and 24

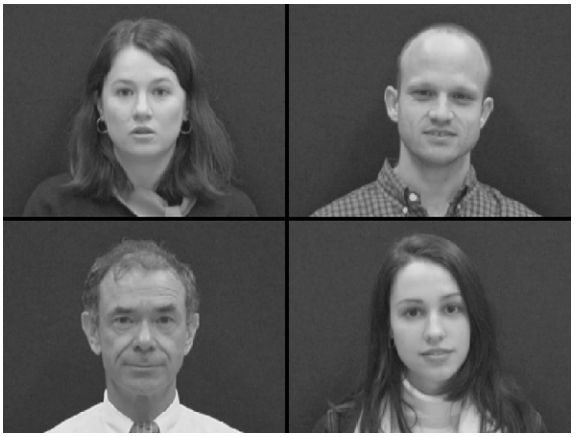

Fig. 7. Example images from the VidTIMIT database.

male), reciting short sentences. It was recorded in three sessions, with a mean delay of 7 days between Session 1 and 2, and 6 days between Session 2 and 3.

The sentences were chosen from the test section of the NTIMIT corpus [23]. There are 10 sentences per person. The first six sentences are assigned to Session 1. The next two sentences are assigned to Session 2 with the remaining two to Session 3. The first two sentences for all persons are the same, with the remaining eight generally different for each person.

The recording was done in a noisy office environment using a broadcast quality digital video camera. The video of each person is stored as a sequence of JPEG images with a resolution of 384×512 pixels. The corresponding audio is stored as a mono, 16-bit, 32 kHz WAV file. Several example images are presented in Fig. 7. For more information on the database, please visit http://spl.me.gu.edu.au/vidtimit/

6.2. Experimental setup

In this section we evaluate the performance of a multi-modal verification system, using different fusion techniques, while varying the SNR.

The experiments were done on the VidTIMIT database. Based on the results from Section 5, CMS $+ \Delta +$ MACV feature extraction was used for speech signals.

Before PCA feature extraction can occur, the face must first be located [26]. Furthermore, to account for varying distances to the camera, a geometrical normalization must be performed. To find the face, we use template matching with several prototype faces of varying dimensions. Using the distance between the eyes as a size measure, an affine transformation is used [27] to adjust the size of the image. This makes the distance between the eyes the same for each person. Finally a 56×64 pixel face window containing the eyes and the nose is extracted from the image. This is the most invariant face area to changes in the expression and hair style. Fig. 8 shows an example face window.

Fig. 8. Example face window.

Using PCA, the dimensionality of the face window is reduced to 40. The choice of the dimensionality is based on the work by Samaria [28].

It must be noted that here we treat the problem of face location and normalization as separate from feature extraction.

For the VAD, the first 30 frames were used. Session 1 of the database was used for generating the client models, while Sessions 2 and 3 were used to obtain opinions resulting from impostor and true claims. Two utterances, from Sessions 2 and 3 separately, each from 8 fixed persons (4 male and 4 female) were used for simulating impostor accesses against the remaining 35 persons. Hence for each Session there were 560 impostor and 70 true claims.

For non-adaptive opinion fusion, opinions from Session 2 were used by an exhaustive search procedure to find the best weights for the lowest EER. The weights were then fixed to find the performance on Session 3.

For the proposed adaptive opinion fusion, the a priori weights were found as for non-adaptive opinion fusion. N_{noise} was set to 30. One mixture for λ_{noise} proved sufficient in preliminary experiments. The sigmoïd parameters a and b [in Eq. (25)] were obtained by observing how q [Eq. (24)] decreases as the SNR is lowered on Session 2.

For Wark's adaptive opinion fusion, Session 2 was used to find the a priori confidences. For all of the fusion techniques, Session 3 was used to evaluate the performance. Results were obtained for non-corrupted speech as well as for corrupted speech were the SNR was varied from 28 to -8 dB. The decision threshold was set to obtain EER performance. The rest of the experimental setup is similar to Section 5.1. The results are shown in Figs. 9–11.

6.3. Discussion

In Fig. 9 we can observe that on the VidTIMIT database the performance of the face modality is better than the speech modality. Moreover, the performance of the speech modality rapidly degrades as the SNR falls below 4 dB.

Fig. 10 shows that the performance of the concatenation fusion approach stays relatively constant while the SNR is lowered. However, for all SNRs the performance is worse than the face modality. This is in contrast to the non-adaptive opinion fusion approach, where the performance is significantly better up to SNR of 12 dB. As expected, the performance then rapidly degrades, becoming worse than the face modality as the SNR falls below 4 dB.

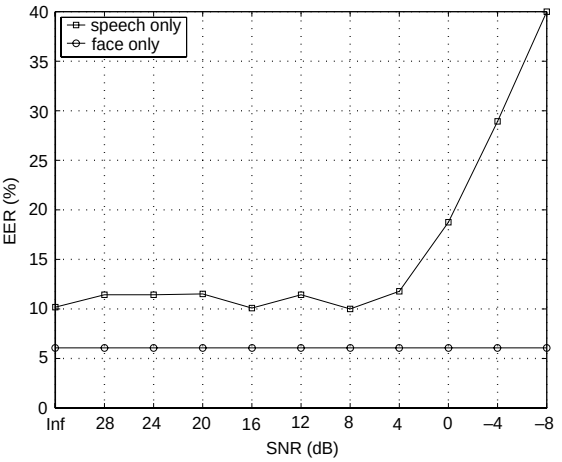

Fig. 9. Performance of individual modalities.

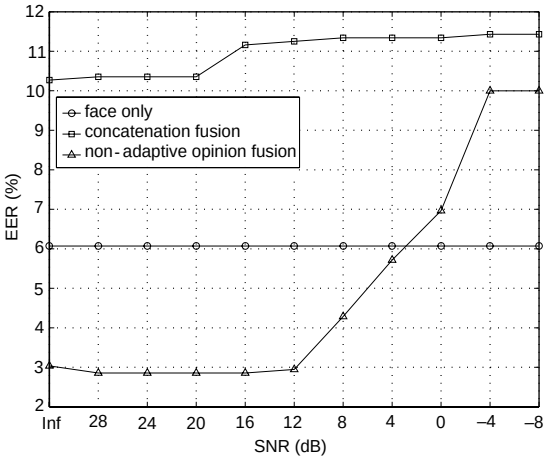

Fig. 10. Performance of non-adaptive fusion approaches.

In Fig. 11 we can see that the performance of Wark's adaptive opinion fusion becomes worse than the face modality as the SNR falls below 4 dB. Wark's approach assumes that under noisy conditions, the distance between a given opinion for an impostor claim and the corresponding model of opinions for impostor claims will decrease. However, we have found the distance to be relatively constant. Hence the a posteriori confidences (κ) for impostor claims changed relatively little as the SNR was lowered, leading to poor performance.

Performance of the proposed adaptive fusion, which explicitly measures the quality of the speech signal, is relatively constant up to SNR of 12 dB. While for lower SNRs it rapidly degrades, it is never worse than the face modality. These results thus support the use of the proposed approach.

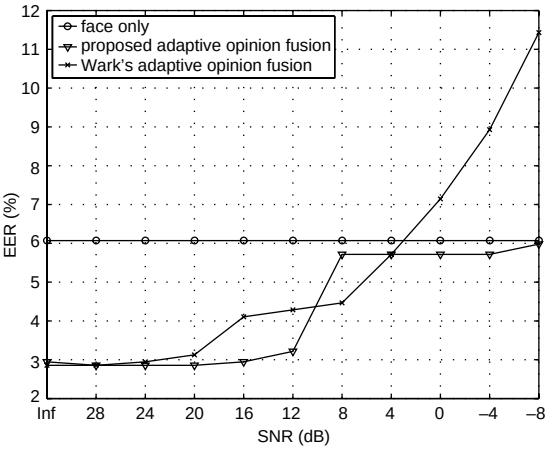

Fig. 11. Performance of adaptive fusion approaches.

7. Conclusion

We have demonstrated that use of MACV features (which utilize information from the source part of the speech signal) significantly improves the robustness of a speech based, text independent verification system.

We have also proposed an adaptive fusion technique for integration of audio and visual information in a multi-modal verification system. The proposed technique explicitly measures the quality of the speech signal, adjusting the amount of contribution of the speech modality to the final verification decision accordingly. Results on the VidTIMIT database indicate that the proposed approach outperforms an existing adaptive fusion technique proposed by Wark [9] as well as two popular non-adaptive methods. For a wide range of audio SNRs, the performance of the multi-modal system utilizing the proposed technique is always better than the performance of the face modality.

References

[1] D. O'Shaughnessy, Speech Communications: Human and Machine, 2nd Edition, IEEE Press, New York, 2000.

[2] D. Reynolds, Speaker identification and verification using Gaussian mixture speaker models, Speech Commun. 17 (1995) 91–108.

[3] F.K. Soong, A.E. Rosenberg, On the use of instantaneous and transitional spectral information in speaker recognition, Proc. IEEE Trans. Acoust. Speech Signal Process. 36 (1988) 871–879.

[4] S. Furui, Cepstral analysis technique for automatic speaker verification, IEEE Trans. Acoust. Speech Signal Process. 29 (1981) 254–272.

[5] B. Wildermoth, K.K. Paliwal, Use of voicing and pitch information for speaker recognition, Proceedings of the Eighth Australian International Conference on Speech Science and Technology, Canberra, 2000, pp. 324–328.

[6] L. Hong, A. Jain, Integrating faces and fingerprints for personal identification, IEEE Trans. Pattern Anal. Mach. Intell. 20 (1998) 1295–1306.
[7] A. Ross, A. Jain, J-Z. Qian, Information fusion in biometrics, Proceedings of the Third Audio- and Video-based Biometric Person Authentication, Halmstad, 2001, pp. 354–359.
[8] S. Ben-Yacoub, Y. Abdeljaoued, E. Mayoraz, Fusion of face and speech data for person identity verification, Proc. IEEE Trans. Neural Networks 10 (1999) 1065–1074.
[9] T. Wark, Multi-modal speech processing for automatic speaker recognition, Ph.D. Thesis, Queensland University of Technology, Brisbane, 2000.
[10] B.C.J. Moore, Frequency analysis and masking, in: D.A. Eddins, D.M. Green (Eds.), Hearing, Academic Press, USA, 1995.
[11] J. Picone, Signal modeling techniques in speech recognition, Proc. IEEE 79 (1991) 1214–1247.
[12] D. Reynolds, A Gaussian mixture modeling approach to text-independent speaker identification, Technical Report 967, Lincoln Laboratory, Massachusetts Institute of Technology, 1993.
[13] R. Balchandran, V. Ramanujam, R. Mammone, Channel estimation and normalization by coherent spectral averaging for robust speaker verification, Proceedings of the Sixth European Conference on Speech Communication and Technology, Budapest, 1999, pp. 755–758.
[14] H. Gish, M. Schmidt, Text-independent speaker identification, IEEE Signal Process. Mag. 11 (1994) 18–32.
[15] T.W. Parsons, Voice and Speech Processing, McGraw-Hill, New York, 1976.
[16] L.R. Rabiner, R.W. Schafer, Digital Signal Processing of Speech Signals, Prentice-Hall, Englewood Cliffs, NJ, 1978.
[17] L.R. Rabiner, On the use of autocorrelation analysis for pitch detection, IEEE Trans. Acoust. Speech Signal Process. 25 (1977) 24–33.
[18] M. Turk, A. Pentland, Eigenfaces for recognition, J. Cognitive Neurosci. 3 (1991) 71–86.
[19] T.K. Moon, Expectation-maximization algorithm, IEEE Signal Process. Mag. 13 (1996) 47–60.
[20] S. Furui, Recent advances in speaker recognition, Pattern Recognition Lett. 18 (1997) 859–872.
[21] P. Silsbee, A. Bovik, Computer lipreading for improved accuracy in automatic speech recognition, IEEE Trans. Speech Audio Process. 4 (1996) 337–351.
[22] C. Sanderson, K.K. Paliwal, Adaptive multi-modal person verification system, First IEEE Pacific-Rim Conference on Multimedia, Sydney, 2000, pp. 210–213.
[23] C. Jankowski, A. Kalyanswamy, S. Basson, J. Spitz, NTIMIT: a phonetically balanced, continuous speech telephone bandwidth speech database, Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Albuquerque, 1990, pp. 109–112.
[24] J.A. Haigh, Voice activity detection for conversational analysis, Masters Thesis, University of Wales, 1994.
[25] D. Reynolds, T. Quatieri, R. Dunn, Speaker verification using adapted Gaussian mixture models, Digital Signal Process. 10 (2000) 19–41.
[26] L-F. Chen, H-Y. Liao, J-C. Lin, C-C. Han, Why recognition in a statistics-based face recognition system should be based on the pure face portion: a probabilistic decision-based proof Pattern Recognition 34 (2001) 1393–1403.
[27] R.C. Gonzales, R.E. Woods, Digital Image Processing, Addison-Wesley, Reading, MA, 1993.
[28] F. Samaria, Face recognition using hidden Markov models, Ph.D. Thesis, University of Cambridge, 1994.

About the Author—CONRAD SANDERSON received the B.Eng. (Hons) degree from Griffith University, Brisbane, Australia in 1996. He has worked at the Advanced Telecommunication Research (ATR) Laboratories, Kyoto, Japan, and is presently a Ph.D. student at the Signal Processing Laboratory at Griffith University. His current research interests include identity authentication, face and speech recognition, and multi-modal information fusion.

About the Author—KULDIP K. PALIWAL received the B.S. degree from Agra University, India in 1969, M.S. degree from Aligarh University, India, in 1971 and Ph.D. degree from Bombay University, India, in 1978. Since 1993, he has been a Professor (Chair, Communication/Information Engineering) at the Griffith University, Brisbane, Australia. He has worked at a number of organizations including Tata Institute of Fundamental Research, Bombay, India, Norwegian Institute of Technology, Trondheim, Norway, University of Keele, UK, AT& T Bell Laboratories, Murray Hill, New Jersey, USA and Advanced Telecommunication Research (ATR) Laboratories, Kyoto, Japan. He has co-edited two books: *Speech Coding and Synthesis* (Elsevier, 1995) and *Speech and Speaker Recognition*: *Advanced Topics* (Kluwer, 1996). He has published more than 100 papers in international journals. He is a recipient of the 1995 IEEE Signal Processing Society Senior Award. He has been an Associate Editor of the IEEE Transactions on Speech and Audio Processing, and IEEE Signal Processing Letters. His current research interests include speech processing, image coding and neural networks.