ELSEVIER

Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT®

Pattern Recognition Letters 24 (2003) 3089–3099

Pattern Recognition Letters

www.elsevier.com/locate/patrec

Structurally noise resistant classifier for multi-modal person verification

Conrad Sanderson [a,b,*], Kuldip K. Paliwal [b]

[a] IDIAP, Rue du Simplon 4, Martigny CH-1920, Switzerland

[b] Signal Processing Laboratory, School of Microelectronic Engineering Griffith University, Nathan, Qld. 4111, Australia

Received 27 June 2002; received in revised form 27 June 2003

Abstract

In this letter we propose a piece-wise linear (PL) classifier for use as the decision stage in a two-modal verification system, comprised of a face and a speech expert. The classifier utilizes a fixed decision boundary that has been specifically designed to account for the effects of noisy audio conditions. Experimental results on the VidTIMIT database show that in clean conditions, the proposed classifier is outperformed by a traditional weighted summation decision stage (using both fixed and adaptive weights). Using white Gaussian noise to corrupt the audio data resulted in the PL classifier obtaining better performance than the fixed approach and similar performance to the adaptive approach. Using a more realistic noise type, namely "operations room" noise from the NOISEX-92 corpus, resulted in the PL classifier obtaining better performance than both the fixed and adaptive approaches. The better results in this case stem from the PL classifier not making a direct assumption about the type of noise that causes the mismatch between training and testing conditions (unlike the adaptive approach). Moreover, the PL classifier has the advantage of having a fixed (non-adaptive, thus simpler) structure.

Keywords: Multi-modal; Identity verification; Fusion; Noise resistance

1. Introduction

Recently there has been a lot of interest in multi-modal biometric person verification systems (Ben-Yacoub et al., 1999; Dugelay et al., 2002). A biometric verification (or authentication) system verifies the identity of a claimant based on the person's physical attributes, such as their voice, face or fingerprints. Apart from security applications (e.g., access control), verification systems are also useful in forensic work (where the task is whether a given biometric sample belongs to a given suspect) and law enforcement applications (Atkins, 2001; Woodward, 1997).

A multi-modal verification system is usually comprised of several *modality experts* (e.g., speech

[*] Corresponding author. Tel.: +41-27-721-7743/61-738756578; fax: +41-27-721-7712/61-738755198.
E-mail address: conradsand@ieee.org (C. Sanderson).

0167-8655/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/S0167-8655(03)00168-5

and face experts). Each expert provides an opinion on a claim, which, for mathematical convenience, is in the $[0,1]$ interval. The opinions from N_{E} modality experts then form an N_{E}-dimensional opinion vector, which is used by a *decision stage* to make the final accept or reject verdict. The decision stage is often a binary classifier discriminating between true claimant and impostor classes (Ben-Yacoub et al., 1999).

Multi-modal systems fall into two categories: non-adaptive and adaptive. While non-adaptive multi-modal systems exhibit lower error rates and are more robust to environmental conditions than mono-modal systems, their performance can still significantly degrade when one of the experts is processing noise corrupted information (e.g., speech with ambient noise) (Wark, 2000; Sanderson and Paliwal, 1999). In adaptive multi-modal systems, the contribution of the noise-affected expert is varied according to current environmental conditions, in an attempt to decrease the performance degradation (Wark, 2000; Sanderson and Paliwal, 2003).

In this letter we propose a structurally noise resistant piece-wise linear (PL) classifier for use in a non-adaptive system. In contrast to an adaptive system, where the decision boundary is effectively adjusted to take into account noisy conditions, the proposed classifier utilizes a fixed decision boundary that has been specifically designed to account for the effects of noisy audio conditions. This approach has the advantage of having a simpler structure than an adaptive approach.

The rest of this letter is organized as follows. In Sections 2 and 3 the speech and face experts are described, respectively. Section 4 describes the method used to map the experts' opinions to the $[0,1]$ interval. In Section 5 the traditional weighted summation decision stage is described, as well as a method to adjust the weights so the contribution of the speech expert is decreased is noisy conditions. The non-adaptive and adaptive approaches are used to compare the performance of the proposed piece-wise linear classifier, which is described in Section 6. Section 7 is devoted to experiments evaluating the proposed classifier. Finally, the paper is concluded in Section 8.

2. Speech expert

The speech expert is comprised of two main components: speech feature extraction and a Gaussian mixture model (GMM) based opinion generator. 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 each frame, a 37-dimensional feature vector is extracted, comprised of mel frequency cepstral coefficients (MFCCs), which represent the instantaneous Fourier spectrum (Reynolds, 1994), their corresponding deltas (which represent transitional spectral information) (Soong and Rosenberg, 1988) and maximum auto-correlation values (which represent pitch and voicing information) (Wildermoth and Paliwal, 2000). Cepstral mean subtraction (Furui, 1981) was applied to MFCCs.

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. We remove these segments using a parametric voice activity detector (VAD), based on the work by Haigh (1994); the details of the specific VAD implementation are presented in (Sanderson, 2002).

2.1. GMM-based opinion generator

The distribution of feature vectors for each person is modeled by a GMM. Given a set of training vectors, an N_{G}-Gaussian GMM is trained using the expectation maximization (EM) algorithm (Dempster et al., 1977; Duda et al., 2001).

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

$$\mathscr{L}(X|\lambda_{\mathrm{C}}) = \frac{1}{N_{\mathrm{V}}} \sum_{i=1}^{N_{\mathrm{V}}} \log p(\vec{x}_i|\lambda_{\mathrm{C}}) \tag{1}$$

where

$$p(\vec{x}|\lambda) = \sum_{j=1}^{N_{\mathrm{G}}} m_j \mathcal{N}(\vec{x}; \vec{\mu}_j, \boldsymbol{\Sigma}_j) \tag{2}$$

$$\lambda = \{m_j, \vec{\mu}_j, \boldsymbol{\Sigma}_j\}_{j=1}^{N_G} \tag{3}$$

Here, $\mathcal{N}(\vec{x}; \vec{\mu}, \boldsymbol{\Sigma})$ is a D-dimensional Gaussian function with mean $\vec{\mu}$ and diagonal covariance matrix $\boldsymbol{\Sigma}$:

$$\mathcal{N}(\vec{x}; \vec{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{D}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left[\frac{-1}{2}(\vec{x} - \vec{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\vec{x} - \vec{\mu})\right] \tag{4}$$

λ_{C} is the parameter set for person C, N_{G} is the number of Gaussians and m_j is the weight for Gaussian j (with constraints $\sum_{j=1}^{N_G} m_j = 1$ and $\forall j : m_j \geqslant 0$).

The likelihood of the claimant being an impostor can be found via the use of a composite model, [1] comprised of several GMMs for other clients. The client models in such a composite are referred to as background models (Reynolds, 1995) or cohort models (Furui, 1997). Given N_{B} background models, the impostor likelihood is found using:

$$\mathcal{L}(X|\lambda_{\overline{\mathrm{C}}}) = \log\left[\frac{1}{N_{\mathrm{B}}}\sum_{b=1}^{N_{\mathrm{B}}} \exp \mathcal{L}(X|\lambda_b)\right] \tag{5}$$

An opinion on the claim is then found using:

$$o = \mathcal{L}(X|\lambda_{\mathrm{C}}) - \mathcal{L}(X|\lambda_{\overline{\mathrm{C}}}) \tag{6}$$

The opinion reflects the likelihood 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). In a mono-modal system, the opinion can be thresholded to achieve the final verification decision.

The background model set contains models which are the "closest" as well as the "farthest" from the client model (Reynolds, 1995). 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 Eq. (6) where both terms would contain similar likelihoods, leading to an unreliable opinion on the claim. In this letter we have utilized the method described by Reynolds (1995) to select the background models for each client.

3. Face expert

The face expert is similar to the speech expert: the GMM-based opinion generator is the same, but a different feature extraction method is used. Here, we have utilized principal component analysis (PCA) to extract features from *closely cropped* (Chen et al., 2001) frontal face images, as follows. Given a face image matrix [2] F of size $X \times Y$, 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 $\{\vec{f}_i\}_{i=1}^{N_{\mathrm{P}}}$ 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:

$$\vec{g}_i = \vec{f}_i - \vec{f}_\mu, \quad i = 1, 2, \ldots, N_{\mathrm{P}} \tag{7}$$

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

$$\boldsymbol{C} = \boldsymbol{G}\boldsymbol{G}^{\mathrm{T}} \tag{8}$$

Let us construct matrix $\boldsymbol{U} = [\,\vec{u}_1 \quad \vec{u}_2 \quad \ldots \quad \vec{u}_D\,]$, containing D eigenvectors of $\boldsymbol{C}$ with largest corresponding eigenvalues. Here, $D < N_{\mathrm{P}}$. A feature vector $\vec{x}$ of dimensionality D is then derived from a face vector $\vec{f}$ using:

$$\vec{x} = \boldsymbol{U}^{\mathrm{T}}(\vec{f} - \vec{f}_\mu) \tag{9}$$

i.e., face vector $\vec{f}$ is decomposed in terms of D eigenvectors, [3] known as "eigenfaces" (Turk and Pentland, 1991). The PCA technique is basically used here for reducing the dimensionality from XY to D.

[1] It must be noted that the universal background model (Reynolds et al., 2000) can also be used to find $\mathcal{L}(X|\lambda_{\overline{\mathrm{C}}})$.

[2] The face images used in our experiments have 64 columns (X) and 56 rows (Y).

[3] In our experiments, we use $D = 40$ (choice based on the work of Samaria (1994)).

4. Mapping opinions to the [0,1] interval

The experiments reported throughout this letter utilize the following method (inspired by Jourlin et al. (1997)) of mapping the output of each expert to the $[0,1]$ interval. The original opinion of expert i, $o_{i,\mathrm{orig}}$, is mapped using a sigmoïd:

$$o_i = \frac{1}{1+\exp[-\tau_i(o_{i,\mathrm{orig}})]} \tag{10}$$

where

$$\tau_i(o_{i,\mathrm{orig}}) = \frac{o_{i,\mathrm{orig}} - (\mu_i - 2\sigma_i)}{2\sigma_i} \tag{11}$$

where, for expert i, μ_i and σ_i are the mean and the standard deviation of original opinions for true claims, respectively. Assuming that the original opinions for true and impostor claims follow Gaussian distributions $\mathcal{N}(o_{i,\mathrm{orig}}; \mu_i, \sigma_i^2)$ and $\mathcal{N}(o_{i,\mathrm{orig}}; \mu_i - 4\sigma_i, \sigma_i^2)$ respectively, 95% of the values 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 (Duda et al., 2001) (see also Fig. 1). Eq. (11) maps the opinions to the $[-2,2]$ interval, which corresponds to the approximately linear portion of the sigmoïd in Eq. (10). The sigmoïd is necessary to take care of situations where the assumptions do not hold entirely.

5. Weighted summation decision stage

A straightforward way to reach a verification decision given several expert opinions is via weighted summation, followed by thresholding (Wark et al., 1999). The opinions of N_{E} experts are first fused as follows:

$$f = \sum_{i=1}^{N_{\mathrm{E}}} w_i o_i \tag{12}$$

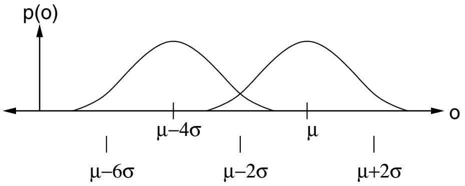

Fig. 1. Graphical interpretation of the assumptions used in Section 4.

where o_i is the opinion of the ith expert (in the $[0,1]$ interval), with corresponding weight w_i (also in the $[0,1]$ interval). The weights have a $\sum_{i=1}^{N_{\mathrm{E}}} w_i = 1$ constraint. The verification decision is then reached as follows: 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). Eq. (12) can be modified to:

$$F(\vec{o}) = \vec{w}^{\mathrm{T}}\vec{o} - t \tag{13}$$

where $\vec{w}^{\mathrm{T}} = [w_i]_{i=1}^{N_{\mathrm{E}}}$ and $\vec{o}^{\mathrm{T}} = [o_i]_{i=1}^{N_{\mathrm{E}}}$. The decision is accordingly modified to: the claim is accepted when $F(\vec{o}) \geqslant 0$; the claim is rejected when $F(\vec{o}) < 0$.

It can be seen that Eq. (13) is a form of a linear discriminant function (Duda et al., 2001), indicating that the procedure of weighted summation and thresholding creates a linear decision boundary in N_{E}-dimensional space. Thus weighted summation fusion is equivalent to a classifier which uses a linear decision boundary to separate the true claimant and impostor classes.

The weights can be selected to reflect the reliability and discrimination ability of each expert. Thus when fusing opinions from a speech and a face expert, it is possible to decrease the contribution of the speech expert when working in low audio signal-to-noise-ratio (SNR) conditions.

A weight update method has been recently presented by Sanderson and Paliwal (2003); it is summarized as follows. Every time a speech utterance is recorded, it is preceded by a short segment which contains only ambient noise. From each training utterance, MFCC feature vectors from the noise segment are used to construct a global noise GMM, λ_{noise}. Given a test speech utterance, N_{noise} MFCC feature vectors, $\{\vec{x}_i\}_{i=1}^{N_{\mathrm{noise}}}$, representing the noise segment, are used to estimate the utterance's quality by measuring the mismatch from λ_{noise} as follows:

$$q = \frac{1}{N_{\mathrm{noise}}} \sum_{i=1}^{N_{\mathrm{noise}}} \log p(\vec{x}_i | \lambda_{\mathrm{noise}}) \tag{14}$$

The larger the difference between the training and testing conditions, the lower q is going to be. q is then mapped to the $[0,1]$ interval using a sigmoïd:

$$q_{\mathrm{map}} = \frac{1}{1+\exp[-a(q-b)]} \tag{15}$$

where a and b describe the shape of the sigmoïd. The values of a and b are selected so that q_{map} is close to one for clean training utterances and close to zero for training utterances artificially corrupted with noise (thus this adaptation method is dependent on the noise type that caused the mismatch).

Let us assume that the face expert is the first expert and that the speech expert is the second expert. Given an a priori weight $w_{2,\text{a priori}}$ for the speech expert (found for clean conditions), the adapted weight for the speech expert is found using:

$$w_2 = q_{\text{map}} w_{2,\text{a priori}} \tag{16}$$

Since we are using a two modal system, there is a $\sum_{i=1}^{2} w_i = 1$ constraint on the weights; moreover, $\forall i : w_i \geqslant 0$. Thus the corresponding weight for the face expert is found using: $w_1 = 1 - w_2$.

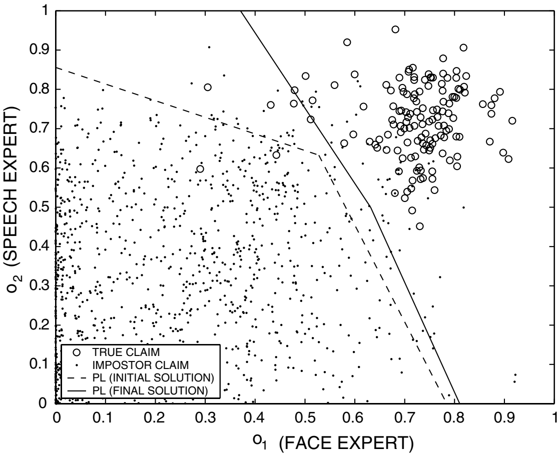

Fig. 2. Initial and final decision boundaries used by the PL classifier and distribution of opinion vectors for true & impostor claims using clean speech.

6. Structurally noise resistant piece-wise linear classifier

6.1. Motivation

For a given claim, let us construct an opinion vector $\vec{o} = [o_1 \;\; o_2]^{\mathrm{T}}$, where o_1 is the opinion of the face expert and o_2 is the opinion of the speech expert. Moreover, let us refer to the distribution of opinion vectors for true claims and impostor claims as the true claimant and impostor opinion distributions, respectively.

The opinion distributions for clean and noisy audio conditions are shown in Figs. 2 and 3, respectively. In this case the speech signal was corrupted with additive white Gaussian noise, simulating ambient noise.

As can be observed, the main effect of noisy audio conditions is the movement of the mean of the true claimant opinion distribution toward the o_1 axis. This movement can be explained by analyzing Eq. (6). Let us suppose a true claim has been made. In clean conditions $\mathscr{L}(X|\lambda_{\mathrm{C}})$ will be high while $\mathscr{L}(X|\lambda_{\overline{\mathrm{C}}})$ will be low, causing o_2 (the opinion of the speech expert) to be high. When the speech expert is processing noisy speech signals, there is a mismatch between training and testing conditions, causing the feature vectors to drift away from the feature space described by the true claimant model (λ_{C}). This in turn causes $\mathscr{L}(X|\lambda_{\mathrm{C}})$ to decrease. If $\mathscr{L}(X|\lambda_{\overline{\mathrm{C}}})$ decreases by the same amount as $\mathscr{L}(X|\lambda_{\mathrm{C}})$, then o_2 is relatively unchanged. However, to model possible impostors, the parametric model representing $\lambda_{\overline{\mathrm{C}}}$ (see Eq. (5)) may cover a wide area of the feature space. Thus while the feature vectors may have drifted away from the feature space described by the true claimant model, they may still be "inside" the space described by the impostor model, causing $\mathscr{L}(X|\lambda_{\overline{\mathrm{C}}})$ to decrease by a smaller amount, which in turn causes o_2 to decrease.

Let us now suppose that an impostor claim has been made. In clean conditions $\mathscr{L}(X|\lambda_{\mathrm{C}})$ will be low while $\mathscr{L}(X|\lambda_{\overline{\mathrm{C}}})$ will be high, causing o_2 to be low. The true claimant model does not represent the impostor feature space, indicating that $\mathscr{L}(X|\lambda_{\mathrm{C}})$ should be consistently low for impostor claims in noisy conditions. As described above, the parametric model representing $\lambda_{\overline{\mathrm{C}}}$ may cover a wide area of the feature space, thus even though the features have drifted due to mismatched conditions, they may still be "inside" the space described by the impostor model. This indicates that $\mathscr{L}(X|\lambda_{\overline{\mathrm{C}}})$ should remain relatively high in noisy conditions, which in turn indicates that the

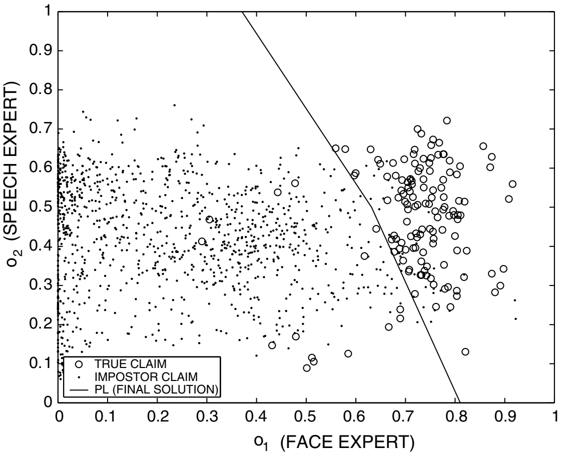

Fig. 3. Final decision boundaries used by the PL classifier and distribution of opinion vectors for true & impostor claims using noisy speech (white noise, SNR = −8 dB).

impostor opinion distribution should change relatively little due to noisy conditions.

While Figs. 2 and 3 were obtained by corrupting the speech signals with additive white Gaussian noise, we would expect a similar movement of the mean of the true claim opinion distribution for other noise types. Generally any noise types alters the features obtained, which would cause $\mathscr{L}(X|\lambda_C)$ to decrease, and as explained above, this leads to a decrease of o_2. Indeed, experiments in Section 7 show that the above argumentation also applies to a more realistic noise type.

6.2. Classifier definition

Let us describe the PL classifier as a discriminant function composed of two linear discriminant functions:

$$g(\vec{o}) = \begin{cases} a(\vec{o}) & \text{if } o_2 \geqslant o_{2,\text{int}} \\ b(\vec{o}) & \text{otherwise} \end{cases} \tag{17}$$

where $\vec{o} = [o_1 \ o_2]^{\mathrm{T}}$ is a two-dimensional opinion vector,

$$a(\vec{o}) = m_1 o_1 - o_2 + c_1 \tag{18}$$

$$b(\vec{o}) = m_2 o_1 - o_2 + c_2 \tag{19}$$

and $o_{2,\text{int}}$ is the threshold for selecting whether to use $a(\vec{o})$ or $b(\vec{o})$. Fig. 4 shows an example of the decision surface. The verification decision is reached as follows. The claim is accepted when $g(\vec{o}) \leqslant 0$ (i.e. true claimant); the claim is rejected when $g(\vec{o}) > 0$ (i.e. impostor).

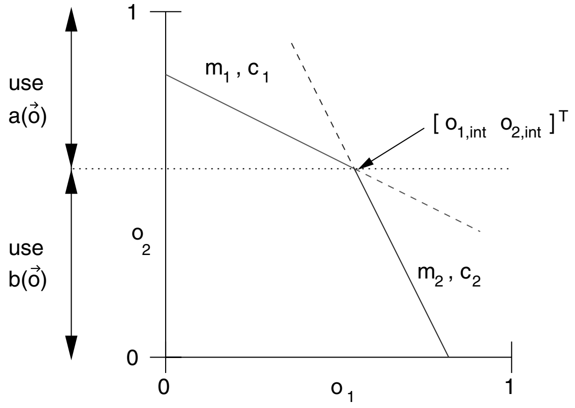

Fig. 4. Example decision surface of the PL classifier.

The first segment of the decision boundary can be described by $a(\vec{o}) = 0$, which reduces Eq. (18) to:

$$0 = m_1 o_1 - o_2 + c_1 \tag{20}$$

hence,

$$o_2 = m_1 o_1 + c_1 \tag{21}$$

If we assume o_2 is a function of o_1, Eq. (21) is simply the description of a line (Swokowski, 1991), where m_1 is the gradient and c_1 is the value at which the line intercepts the o_2 axis. Similar argument can be applied to the description of the second segment of the decision boundary. Given m_1, c_1, m_2 and c_2, we can find $o_{2,\text{int}}$ as follows. The two lines intersect at a single point $\vec{o}_{\text{int}} = [o_{1,\text{int}} \ o_{2,\text{int}}]^{\mathrm{T}}$; moreover, when the two lines intersect, $a(\vec{o}_{\text{int}}) = b(\vec{o}_{\text{int}}) = 0$. Hence

$$o_{2,\text{int}} = m_1 o_{1,\text{int}} + c_1 \tag{22}$$

and

$$o_{2,\text{int}} = m_2 o_{1,\text{int}} + c_2 \tag{23}$$

which leads to:

$$o_{1,\text{int}} = \frac{c_1 - c_2}{m_2 - m_1} \tag{24}$$

$$o_{2,\text{int}} = m_2 \left(\frac{c_1 - c_2}{m_2 - m_1} \right) + c_2 \tag{25}$$

6.3. Structural constraints and training

As described in Section 6.1, the main effect of noisy audio conditions is the movement of the mean of the true claim opinion distribution toward the o_1 axis. We would like to obtain a decision surface which minimizes the increase of verification errors due to this movement. Structurally, this requirement translates to a decision surface that is as steep as possible; moreover, we would like the classifier to be trained for equal error rate (EER) performance. This in turn translates to the following constraints on the parameters of the PL classifier:

1. Both lines must exist in valid 2D opinion space (where the opinion from each expert is in the $[0,1]$ interval) indicating that their intersect is constrained to exist in valid 2D opinion space.
2. Gradients for both lines have to be as large as possible.
3. The EER criterion must be satisfied.

Let $\lambda_{\mathrm{PL}} = \{m_1, c_1, m_2, c_2\}$ be the set of PL classifier parameters. Given an initial solution, described in Section 6.4, the downhill simplex optimization method (Nelder and Mead, 1965; Press et al., 1992) can be used to find the final parameters. The following function is minimized:

$$\varepsilon(\lambda_{\mathrm{PL}}) = \epsilon_1(\lambda_{\mathrm{PL}}) + \epsilon_2(\lambda_{\mathrm{PL}}) + \epsilon_3(\lambda_{\mathrm{PL}}) \tag{26}$$

where $\epsilon_1(\lambda_{\mathrm{PL}})$ through $\epsilon_3(\lambda_{\mathrm{PL}})$ (defined below) represent constraints 1–3 described above, respectively.

$$\epsilon_1(\lambda_{\mathrm{PL}}) = \gamma_1 + \gamma_2 \tag{27}$$

where

$$\gamma_j = \begin{cases} |o_{j,\mathrm{int}}| & \text{if } o_{j,\mathrm{int}} < 0 \text{ or } o_{j,\mathrm{int}} > 1 \\ 0 & \text{otherwise} \end{cases} \tag{28}$$

where $o_{1,\mathrm{int}}$ and $o_{2,\mathrm{int}}$ are found using Eqs. (24) and (25), respectively,

$$\epsilon_2(\lambda_{\mathrm{PL}}) = \left|\frac{1}{m_1}\right| + \left|\frac{1}{m_2}\right| \tag{29}$$

and finally

$$\epsilon_3(\lambda_{\mathrm{PL}}) = \left|\frac{\mathrm{FA\%}}{100\%} - \frac{\mathrm{FR\%}}{100\%}\right| \tag{30}$$

where FA% and FR% is the false acceptance rate and false rejection rate, respectively.

6.4. Initial solution of PL parameters

The initial solution for λ_{PL} (required by the downhill simplex optimization) is based on the impostor opinion distribution. Let us assume that the distribution can be described by a 2D Gaussian function with a diagonal covariance matrix [see Eq. (4)], indicating that it can be characterized by $\{\mu_1, \mu_2, \sigma_1, \sigma_2\}$, where μ_j and σ_j is the mean and standard deviation in the jth dimension, respectively. Under the Gaussian assumption, 95% of the values for the jth dimension lie in the $[\mu_j - 2\sigma_j, \mu_j + 2\sigma_j]$ interval. Let us use this property to define three points in 2D opinion space (shown graphically in Fig. 5):

$$P_1 = (x_1, y_1) = (\mu_1, \mu_2 + 2\sigma_2) \tag{31}$$

$$P_2 = (x_2, y_2) = \left(\mu_1 + 2\sigma_1 \cos\left[\frac{\pi}{4}\right], \mu_2 + 2\sigma_2 \sin\left[\frac{\pi}{4}\right]\right) \tag{32}$$

$$P_3 = (x_3, y_3) = (\mu_1 + 2\sigma_1, \mu_2) \tag{33}$$

Thus the gradient (m_1) and the intercept (c_1) for the first line can be found using:

$$m_1 = \frac{y_2 - y_1}{x_2 - x_1} \tag{34}$$

$$c_1 = y_1 - m_1 x_1 \tag{35}$$

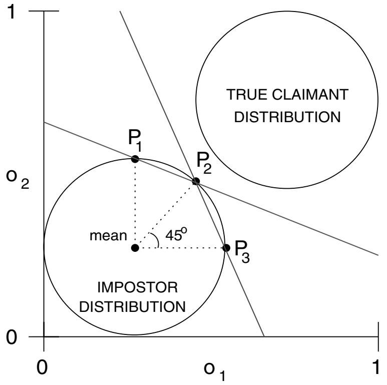

Fig. 5. Points used in the initial solution of PL classifier parameters.

Similarly, the gradient (m_2) and the intercept (c_2) for the second line can be found using:

$$m_2 = \frac{y_3 - y_2}{x_3 - x_2} \tag{36}$$

$$c_2 = y_2 - m_2 x_2 \tag{37}$$

The initial solution for real data is shown in Fig. 2.

7. Evaluation

7.1. VidTIMIT audio-visual database

The VidTIMIT database (Sanderson, 2002) is comprised of video and corresponding audio recordings of 43 people (19 female and 24 male), reciting short sentences selected from the NTIMIT corpus (Jankowski et al., 1990). It was recorded in 3 sessions, with a mean delay of 7 days between Session 1 and 2, and 6 days between Session 2 and 3.

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 mean duration of each sentence is 4.25 s, or approximately 106 video frames (using 25 fps).

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 512×384 pixels. The corresponding audio is stored as a mono, 16 bit, 32 kHz WAV file.

7.2. Experiments

Session 1 was used for training the speech and face experts. Each expert used eight-Gaussian client models. To find the performance, Sessions 2 and 3 were used for obtaining expert opinions of known impostor and true claims. Four utterances, each from 8 fixed persons (4 male and 4 female), were used for simulating impostor accesses against the remaining 35 persons. As per Reynolds (1995), 10 background person models were used for the impostor likelihood calculation. For each of the remaining 35 persons, their four utterances were used separately as true claims. In total there were 1120 impostor and 140 true claims.

In the first set of experiments, speech signals were corrupted by additive white Gaussian noise, with the SNR varying from 28 to −8 dB. Opinion mapping parameters (Section 4) were found on clean test data. Based on manual observation of plots of speech signals from the VidTIMIT database, N_{noise} was set to 30 for the adaptive weight adjustment method [see Eq. (14)]. As per Sanderson and Paliwal (2003), λ_{noise} was comprised of a single Gaussian. The sigmoïd parameters a and b [in Eq. (15)] were obtained by observing how q in Eq. (14) decreased as the SNR was lowered (using white Gaussian noise) on utterances in Session 1 (i.e., training utterances). The resulting value of q_{map} in Eq. (15) was close to one for clean utterances and close to zero for utterances with an SNR of −8 dB.

In the second set of experiments, speech signals were corrupted by adding "operations room" noise from the NOISEX-92 corpus (Varga et al., 1992); the "operations room" noise contains background speech as well as machinery sounds. For the adaptive weight adjustment method, the same parameters were used as found for the white noise case.

Performance of the following configurations was found: face expert alone, speech expert alone, weighted summation fusion with fixed & adaptive weights and the proposed piece-wise linear classifier. In multi-modal cases, the face expert provided the first opinion (o_1) while the speech expert provided the second opinion (o_2) when forming the opinion vector $\vec{o} = [o_1 \;\; o_2]^{\text{T}}$.

As a common starting point, classifier parameters (for all approaches) were selected to obtain performance as close as possible to EER on clean test data (following the popular practice in the speaker verification area of using EER as a measure of expected performance (Doddington et al., 2000; Furui, 1997)). The parameters for the weighted summation decision stage were found via an exhaustive search procedure. Given the common starting point, the performance in noisy conditions was then found in terms of false acceptance rate (FA%) and false rejection rate (FR%) and combined into one number:

$$\mathrm{TE} = \mathrm{FA}\% + \mathrm{FR}\% \tag{38}$$

where TE stands for total error. Results are presented in Figs. 6–8. It must be noted that results for noisy conditions cannot be reported in terms of EER; doing so would amount to adjusting classifier parameters to achieve EER performance, which can be interpreted as a non-causal adaptation method.

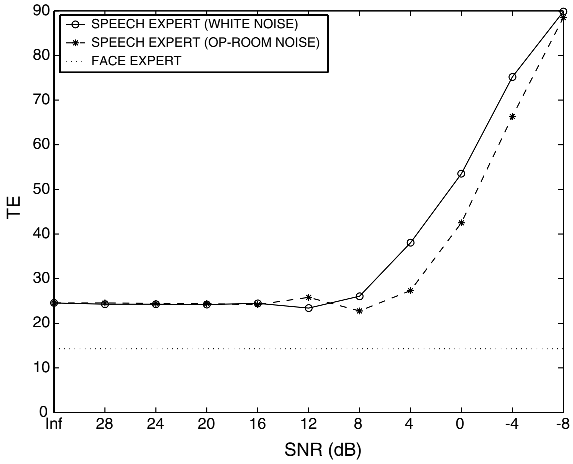

Fig. 6. Performance of the speech and face experts.

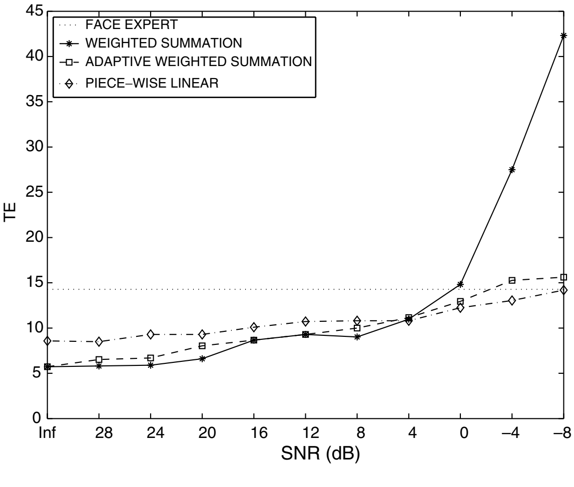

Fig. 7. Performance of the PL classifier compared to fixed and adaptive weighted summation decision stage (white noise).

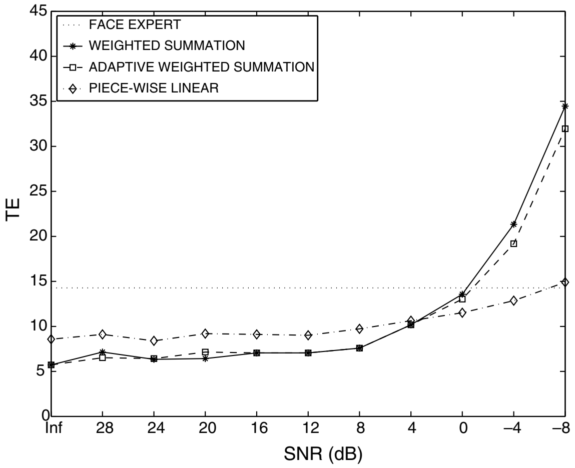

Fig. 8. Performance of the PL classifier compared to fixed and adaptive weighted summation decision stage (operations room noise).

The distribution of opinion vectors for clean and noisy audio data (as well as the decision boundary used by the PL classifier) is shown in Figs. 2 and 3, respectively. While Fig. 3 only shows the effects of white noise, we have observed similar effects when using the "operations room" noise.

7.3. Discussion

As can be observed in Figs. 2 and 3, the decision boundary used by the PL classifier effectively takes into account the movement of opinion vectors due to noisy audio conditions. In clean and low noise conditions (both noise types) the weighted summation decision stage (using both fixed and adaptive weights) outperforms the PL classifier.

For the case of white noise, the PL classifier obtains better performance in high noise conditions (SNR $\leqslant$ 0) than the fixed approach and has similar performance as the adaptive approach. For the case of "operations room" noise, the weight update algorithm shows its limitation of being dependent on the noise type; the algorithm was configured to operate with white noise and was unable to handle the "operations room" noise; the resulting performance of the adaptive approach is very similar to the fixed weight approach. This is in contrast to the PL classifier, which obtains better performance in high noise conditions.

8. Conclusion

In this letter we have proposed a piece-wise linear (PL) classifier for use as the decision stage in a two-modal verification system, comprised of a face and a speech expert. The classifier utilizes a fixed decision boundary that has been specifically designed to account for the effects of noisy audio conditions. Experimental results on the multi-session VidTIMIT database show that in clean conditions, the proposed classifier is outperformed by a traditional weighted summation decision stage (using both fixed and adaptive weights). Using white Gaussian noise to corrupt the audio data resulted in the PL classifier obtaining better performance than the fixed approach and similar performance to the adaptive approach. Using a more realistic noise type, namely "operations room" noise from the NOISEX-92 corpus, resulted in the PL classifier obtaining better performance than both the fixed and adaptive approaches. The better results in this case stem from the PL classifier not making a direct assumption about the type of noise that caused the mismatch between training and testing conditions (unlike the weight update algorithm used in the weighted summation decision stage). Moreover, the PL classifier has the advantage of having a fixed (non-adaptive, thus simpler) structure.

References

Atkins, W., 2001. A testing time for face recognition technology. Biomet. Technol. Today 9 (3), 8–11.

Ben-Yacoub, S., Abdeljaoued, Y., Mayoraz, E., 1999. Fusion of face and speech data for person identity verification. IEEE Trans. Neural Networks 10 (5), 1065–1074.

Chen, L.-F., Liao, H.-Y., Lin, J.-C., Han, C.-C., 2001. 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 (7), 1393–1403.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. Royal Statistical Soc., Ser. B 39 (1), 1–38.

Dugelay, J.-L., Junqua, J.-C., Kotropoulos, C., Kuhn, R., Perronnin, F., Pitas, I., 2002. Recent advances in biometric person authentication. In: Proc. Internat. Conf. Acoustics, Speech Signal Processing, Orlando, vol. IV, pp. 4060–4063.

Doddington, G.R., Przybocki, M.A., Martin, A.F., Reynolds, D.A., 2000. The NIST speaker recognition evaluation—Overview, methodology, systems, results, perspective. Speech Commun. 31 (2–3), 225–254.

Duda, R.O., Hart, P.E., Stork, D.G., 2001. Pattern Classification. John Wiley & Sons, USA.

Furui, S., 1981. Cepstral analysis technique for automatic speaker verification. IEEE Trans. Acoustics, Speech Signal Process. 29 (2), 254–272.

Furui, S., 1997. Recent advances in speaker recognition. Pattern Recognition Lett. 18 (9), 859–872.

Haigh, J.A., 1994. Voice Activity Detection for Conversational Analysis. Masters Thesis, University of Wales, United Kingdom.

Jankowski, C., Kalyanswamy, A., Basson, S., Spitz, J., 1990. NTIMIT: A phonetically balanced, continuous speech telephone bandwidth speech database. In: Proc. Internat. Conf. Acoustics, Speech Signal Process., Albuquerque, vol. 1, pp. 109–112.

Jourlin, P., Luettin, J., Genoud, D., Wassner, H., 1997. Acoustic-labial speaker verification. Pattern Recognition Lett. 18 (9), 853–858.

Nelder, J.A., Mead, R., 1965. A simplex method for function minimization. Computer J. 7 (4), 308–313.

Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P., 1992. Numerical Recipes in C: The Art of Scientific Computing. Cambridge University Press, Cambridge.

Reynolds, D.A., 1994. Experimental evaluation of features for robust speaker identification. IEEE Trans. Speech Audio Process. 2 (4), 639–643.

Reynolds, D.A., 1995. Speaker identification and verification using Gaussian mixture speaker models. Speech Commun. 17 (1–2), 91–108.

Reynolds, D., Quatieri, T., Dunn, R., 2000. Speaker verification using adapted Gaussian mixture models. Digital Signal Process. 10 (1–3), 19–41.

Samaria, F., 1994. Face Recognition Using Hidden Markov Models. Ph.D. Thesis, University of Cambridge.

Sanderson, C., 2002. Automatic Person Verification Using Speech and Face Information. Ph.D. Thesis, Griffith University, Queensland, Australia. (see http://adt.caul.edu.au).

Sanderson, C., Paliwal, K.K., 1999. Multi-modal person verification system based on face profiles and speech. In: Proc. 5th Internat. Symp. Signal Process. Appl., Brisbane, vol. 2, pp. 947–950.

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

Soong, F.K., Rosenberg, A.E., 1988. On the use of instantaneous and transitional spectral information in speaker recognition. IEEE Trans. Acoustics, Speech Signal Process. 36 (6), 871–879.

Swokowski, E.W., 1991. Calculus, fifth ed. PWS-Kent, USA.

Turk, M., Pentland, A., 1991. Eigenfaces for recognition. J. Cognitive Neurosci. 3 (1), 71–86.

Varga, A., Steeneken, H.J.M., Tomlinson, M., Jones, D., 1992. The NOISEX-92 Study on the Effect of Additive Noise on Automatic Speech Recognition. Technical Report, Defence Evaluation and Research Agency (DERA), Speech Research Unit, Malvern, United Kingdom.

Wark, T., Sridharan, S., Chandran, V., 1999. Robust speaker verification via fusion of speech and lip modalities. In: Proc. Internat. Conf. Acoustics, Speech Signal Process., Phoenix, vol. 6, pp. 3061–3064.

Wark, T., 2000. Multi-modal Speech Processing for Automatic Speaker Recognition. Ph.D. Thesis, Queensland University of Technology, Brisbane, Australia.

Wildermoth, B., Paliwal, K.K., 2000. Use of voicing and pitch information for speaker recognition. In: Proc. 8th Australian Internat. Conf. Speech Sci. Technol. Canberra, pp. 324–328.

Woodward, J.D., 1997. Biometrics: Privacy's foe or privacy's friend? Proc. IEEE 85 (9), 1480–1492.