Noise resistant audio-visual verification via structural constraints

NOISE RESISTANT AUDIO-VISUAL VERIFICATION VIA STRUCTURAL CONSTRAINTS

Conrad Sanderson *

IDIAP
PO Box 592, CH-1920 Martigny, Switzerland

Kuldip K. Paliwal

School of Microelectronic Engineering
Griffith University, QLD 4111, Australia

ABSTRACT

In this paper we propose a piece-wise linear 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 show that in clean conditions the proposed classifier is outperformed by a traditional weighted summation decision stage (using both fixed and adaptive weights); however, in high noise conditions the classifier obtains better performance than the fixed approach and has similar performance as the adaptive approach, with the advantage of having a fixed (non-adaptive) structure.

1. INTRODUCTION

Recently there has been a lot of interest in multi-modal biometric person verification systems [1]. 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 [16].

A multi-modal verification system is usually comprised of several *modality experts* (e.g., speech 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 [1].

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) [10]. 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 [11].

In this paper we propose a structurally noise resistant piecewise 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 conditions. This approach has the advantage of having a simpler structure than an adaptive approach.

*Financial support by the Swiss National Center of Competence in Research (NCCR) on Interactive Multimodal Information Management (IM)2.

The rest of the paper is organized as follows. In Sections 2 and 3 the speech and face experts are described, respectively. In Section 4 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 proposed PL classifier is described in Section 5. Section 6 is devoted to experiments comparing the proposed classifier against the traditional weighted summation decision stage (in both adaptive and non-adaptive configurations).

2. SPEECH EXPERT

The speech expert is comprised of two main components: speech feature extraction and a Gaussian Mixture Model (GMM) classifier. 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 (MFCC) [7], their corresponding deltas [9] and Maximum Auto-Correlation Values (which represent pitch and voicing information) [15].

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 [2, 4].

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:

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

where
$$p(\vec{x}|\lambda) = \sum_{j=1}^{N_M} m_j \, \mathcal{N}(\vec{x}; \vec{\mu}_j, \mathbf{\Sigma}_j) \quad (2)$$

and
$$\lambda = \{m_j, \vec{\mu}_j, \mathbf{\Sigma}_j\}_{j=1}^{N_M} \quad (3)$$

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 $\mathcal{N}(\vec{x}; \vec{\mu}, \mathbf{\Sigma})$ is a multi-variate Gaussian function with mean $\vec{\mu}$ and diagonal covariance matrix $\mathbf{\Sigma}$ [4]. Given a set $\{\lambda_b\}_{b=1}^{B}$ of B background person models for person C, the average log likelihood of the claimant being an impostor is found using:

$$\mathcal{L}(X|\lambda_{\overline{C}}) = \log \left[\frac{1}{B} \sum_{b=1}^{B} \exp \mathcal{L}(X|\lambda_b) \right] \quad (4)$$

The set of background person models is found using the method described in [8]. An opinion on the claim is found using:

$$o = \mathcal{L}(X|\lambda_C) - \mathcal{L}(X|\lambda_{\overline{C}}) \quad (5)$$

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). Before using the opinion in a multi-modal system, mapping to the $[0, 1]$ interval is usually performed [11].

3. FACE EXPERT

The face expert is similar to the speech expert. It differs in the feature extraction method: Principal Component Analysis (PCA)

[13] is employed to extract features from frontal face images. Given a face image matrix F of size $X \times Y$ (in our experiments we use 64×56), we construct a vector representation by concatenating all the columns of F to form a column vector $\vec{f}$ of dimensionality XY. A feature vector $\vec{x}$ of dimensionality D is then derived from a face vector $\vec{f}$ using:

$$\vec{x} = \boldsymbol{U}^T(\vec{f} - \vec{f}_\mu) \tag{6}$$

where $\boldsymbol{U}$ contains D eigenvectors (with largest corresponding eigenvalues) of the training data covariance matrix, and $\vec{f}_\mu$ is the mean of training face vectors. Typically, $D = 40$.

4. WEIGHTED SUMMATION DECISION STAGE

A straightforward way to reach a verification decision given several expert opinions is via weighted summation, followed by thresholding [14]. The opinions of N_E experts are first fused as follows:

$$f = \sum_{i=1}^{N_E} w_i o_i \tag{7}$$

where o_i is the opinion of the *i*-th expert (in the [0,1] interval), with corresponding weight w_i (also in the [0,1] interval). The weights have a $\sum_{i=1}^{N_E} w_i = 1$ constraint. The verification decision is then reached as follows: given a threshold t, the claim is accepted when $f \geq t$ (i.e., true claimant); the claim is rejected when $f < t$ (i.e., impostor). Eqn. (7) can be modified to:

$$F(\vec{o}) = \vec{w}^{\,T}\vec{o} - t \tag{8}$$

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

It can be seen that Eqn. (8) is a form of a linear discriminant function [4], indicating that the procedure of weighted summation and thresholding creates a linear decision boundary in N_E-dimensional space which discriminates between the true claimant and impostor classes.

4.1. Adaptivity

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 SNR conditions. A weight update method presented in [11] 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, $\{\vec{x}_i\}_{i=1}^{N_{\text{noise}}}$, representing the noise segment, are used to estimate the utterance's quality by measuring the mismatch from λ_{noise} as follows:

$$q = \frac{1}{N_{\text{noise}}} \sum_{i=1}^{N_{\text{noise}}} \log p(\vec{x}_i|\lambda_{\text{noise}}) \tag{9}$$

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:

$$q_{\text{map}} = \frac{1}{1 + \exp[-a(q-b)]} \tag{10}$$

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 $w_{2,\text{apriori}}$ for the speech expert (found for clean conditions), the adapted weight for the speech expert is found using:

$$w_2 = q_{\text{map}}\, w_{2,\text{apriori}} \tag{11}$$

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

5. STRUCTURALLY NOISE RESISTANT PIECE-WISE LINEAR CLASSIFIER

5.1. Motivation

For a given claim, let us construct an opinion vector $\vec{o} = [\, o_1 \; o_2 \,]^T$, 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. 1 and 2, respectively. In noisy conditions, the speech signal was corrupted with additive white Gaussian noise, simulating ambient noise.

As can be observed, the main effect of noisy 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 Eqn. (5). Let us suppose a true claim has been made; in clean conditions $\mathcal{L}(X|\lambda_C)$ will be high while $\mathcal{L}(X|\lambda_{\overline{C}})$ 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 causes $\mathcal{L}(X|\lambda_C)$ to decrease. If $\mathcal{L}(X|\lambda_{\overline{C}})$ decreases by the same amount as $\mathcal{L}(X|\lambda_C)$, then o_2 is relatively unchanged. However, to model possible impostors, the parametric model representing $\lambda_{\overline{C}}$ [see Eqn. (4)] 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 $\mathcal{L}(X|\lambda_{\overline{C}})$ 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 $\mathcal{L}(X|\lambda_C)$ will be low while $\mathcal{L}(X|\lambda_{\overline{C}})$ will be high, causing o_2 to be low. The true claimant model does not represent the impostor feature space, indicating that $\mathcal{L}(X|\lambda_C)$ should be consistently low for impostor claims in noisy conditions. As described above, the parametric model representing $\lambda_{\overline{C}}$ 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 $\mathcal{L}(X|\lambda_{\overline{C}})$ should remain relatively high in noisy conditions, which in turn indicates that the impostor opinion distribution should change relatively little due to noisy conditions.

While Figs. 1 and 2 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 $\mathcal{L}(X|\lambda_C)$ to decrease, and as explained above, this leads to a decrease of o_2.

5.2. Classifier Definition

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

$$g(\vec{o}) = \begin{cases} a(\vec{o}) & \text{if } o_2 >= o_{2,int} \\ b(\vec{o}) & \text{otherwise} \end{cases} \tag{12}$$

where $\vec{o} = [\, o_1 \; o_2 \,]^T$ is a 2-dimensional opinion vector,

$$a(\vec{o}) = m_1\, o_1 - o_2 + c_1 \tag{13}$$

$$b(\vec{o}) = m_2\, o_1 - o_2 + c_2 \tag{14}$$

and $o_{2,int}$ is the threshold for selecting whether to use $a(\vec{o})$ or $b(\vec{o})$. Fig. 3 shows an example of the decision surface. The verification decision is reached as follows. The claim is accepted when $g(\vec{o}) \leq 0$ (i.e., true claimant); the claim is rejected when $g(\vec{o}) > 0$ (i.e., impostor).

The first segment of the decision boundary can be described by $a(\vec{o}) = 0$, which reduces Eqn. (13) to:

$$0 = m_1\, o_1 - o_2 + c_1 \quad (15)$$

$$\text{hence, } o_2 = m_1\, o_1 + c_1 \quad (16)$$

If we assume o_2 is a function of o_1, Eqn. (16) is simply the description of a line [12], 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 $o_{2,int}$ as follows. The two lines intersect at a single point $\vec{o}_{int} = [\, o_{1,int}\, o_{2,int}\,]^T$; moreover, when the two lines intersect, $a(\vec{o}_{int}) = b(\vec{o}_{int}) = 0$. Hence,

$$o_{2,int} = m_1\, o_{1,int} + c_1 \quad (17)$$

$$\text{and } o_{2,int} = m_2\, o_{1,int} + c_2 \quad (18)$$

which leads to:

$$o_{1,int} = \frac{c_1 - c_2}{m_2 - m_1} \quad (19)$$

$$o_{2,int} = m_2 \left(\frac{c_1 - c_2}{m_2 - m_1} \right) + c_2 \quad (20)$$

5.3. Structural Constraints and Training

As described in Section 5.1, the main effect of noisy 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 $\lambda_{\text{PL}} = \{m_1, c_1, m_2, c_2\}$ be the set of PL classifier parameters. Given an initial solution (described in Section 5.4), the downhill simplex optimization method [5, 6] can be used to find the final parameters. The following function is minimized:

$$\varepsilon(\lambda_{\text{PL}}) = \epsilon_1(\lambda_{\text{PL}}) + \epsilon_2(\lambda_{\text{PL}}) + \epsilon_3(\lambda_{\text{PL}}) \quad (21)$$

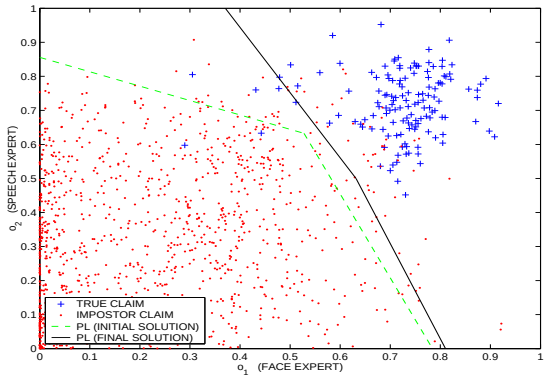

Fig. 1. Initial and final decision boundaries used by PL classifier and distribution of opinion vectors for true & impostor claims using clean speech

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

$$\epsilon_1(\lambda_{\text{PL}}) = \gamma_1 + \gamma_2 \quad (22)$$

$$\text{where } \gamma_j = \begin{cases} |o_{j,int}| & \text{if } o_{j,int} < 0 \text{ or } o_{j,int} > 1 \\ 0 & \text{otherwise} \end{cases} \quad (23)$$

where $o_{1,int}$ and $o_{2,int}$ are found using Eqns. (19) and (20), respectively,

$$\epsilon_2(\lambda_{\text{PL}}) = \left| \frac{1}{m_1} \right| + \left| \frac{1}{m_2} \right| \quad (24)$$

$$\text{and finally } \epsilon_3(\lambda_{\text{PL}}) = \left| \frac{\text{FA\%}}{100\%} - \frac{\text{FR\%}}{100\%} \right| \quad (25)$$

where FA% and FR% is the False Acceptance rate and False Rejection rate, respectively.

5.4. Initial Solution of PL Parameters

The initial solution for λ_{PL} 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, 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 j-th dimension, respectively. Under the Gaussian assumption, 95% of the values for the j-th dimension lie in the $[\mu_j - 2\sigma_j, \mu_j + 2\sigma_j]$ interval [4]. Let us use this property to define three points in 2D opinion space (shown graphically in Fig. 4):

$$P_1 = (x_1, y_1) = (\mu_1\ ,\ \mu_2 + 2\sigma_2) \quad (26)$$

$$P_2 = (x_2, y_2) = \left(\mu_1 + 2\sigma_1 \cos\left[\frac{\pi}{4}\right]\ ,\ \mu_2 + 2\sigma_2 \sin\left[\frac{\pi}{4}\right]\right) \quad (27)$$

$$P_3 = (x_3, y_3) = (\mu_1 + 2\sigma_1\ ,\ \mu_2) \quad (28)$$

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

$$m_1 = \frac{y_2 - y_1}{x_2 - x_1} \quad (29)$$

$$c_1 = y_1 - m_1 x_1 \quad (30)$$

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

$$m_2 = \frac{y_3 - y_2}{x_3 - x_2} \quad (31)$$

$$c_2 = y_2 - m_2 x_2 \quad (32)$$

The initial solution for real data is shown in Fig. 1.

6. EXPERIMENTS

6.1. VidTIMIT Audio-Visual Database

The VidTIMIT database [11], is comprised of video and corresponding audio recordings of 43 people, reciting short sentences. It was recorded in 3 sessions; the mean duration of each sentence is 4.25 seconds, or approx. 106 video frames. For more information on the database, please see *http://www.idiap.ch/~sanders/vidtimit/*

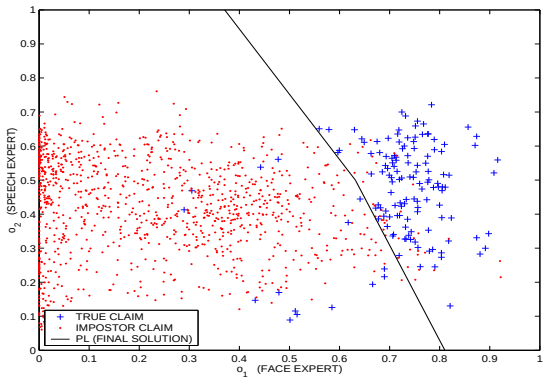

Fig. 2. Final decision boundaries used by PL classifier and distribution of opinion vectors for true & impostor claims using noisy speech (SNR=−8dB)

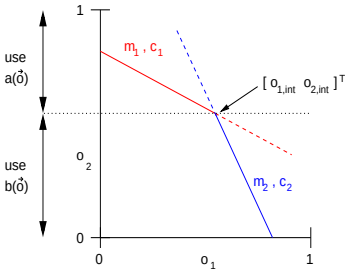

Fig. 3. Example decision surface of the PL classifier

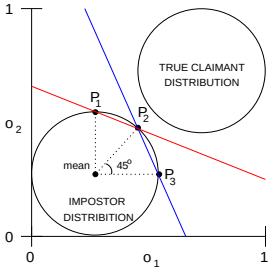

Fig. 4. Points used in the initial solution of PL classifier parameters

6.2. Experimental Setup

Session 1 was used for training the speech and face experts. Each expert used 8 mixture 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 in [8], 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.

Speech signals were corrupted by additive white Gaussian noise, with the SNR varying from 28 to -8 dB. Opinions of the experts were mapped to the $[0, 1]$ interval using the method described in [11]. 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 Eqn. (9)]. As in [11], λ_{noise} was comprised of a single mixture. The sigmoïd parameters a and b [in Eqn. (10)] were obtained by observing how q in Eqn. (9) decreased as the SNR was lowered on utterances in Session 1 (i.e., training utterances). The resulting value of q_{map} in Eqn. (10) was close to one for clean utterances and close to zero for utterances with an SNR of -8 dB.

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 $\vec{o} = [\, o_1 \; o_2 \,]^T$.

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 standard practice in the speaker verification area of using EER as a measure of expected performance [3]). 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:

$$\text{TE} = \text{FA\%} + \text{FR\%} \tag{33}$$

where TE stands for Total Error. Results are presented in Fig. 5. 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.

The distribution of opinion vectors for clean and noisy data (as well as the decision boundary used by the PL classifier) is shown in Figs. 1 and 2, respectively.

6.3. Discussion and Conclusions

As can be observed in Figs. 1 and 2, the decision boundary used by the PL classifier effectively takes into account the movement of opinion vectors due to noisy conditions. In clean and low noise conditions the weighted summation decision stage (using both fixed and adaptive weights) outperforms the PL classifier. However, in high noise conditions (SNR $\leq$ 0) the PL classifier obtains better performance than the fixed approach and has similar performance as the adaptive approach, with the advantage of having a fixed (non-adaptive) structure. Moreover, unlike the weight update algorithm used in the adaptive approach, the PL classifier does not make a direct assumption about the type of noise that caused the mismatch between training and testing conditions.

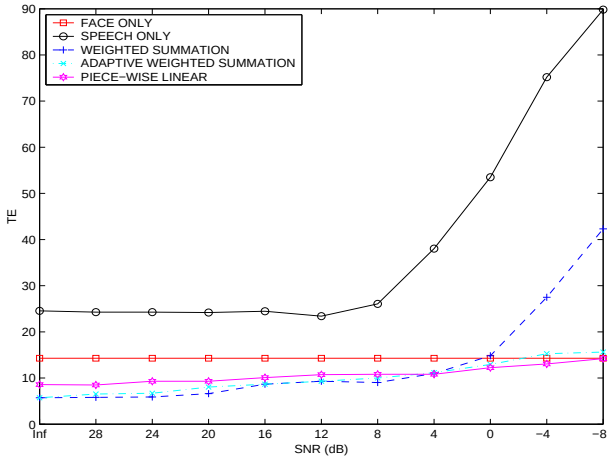

Fig. 5. Performance of the PL classifier compared to fixed and adaptive weighted summation decision stage

7. REFERENCES

[1] S. Ben-Yacoub et al, "Fusion of Face and Speech Data for Person Identity Verification", 10 (5), 1999, 1065-1074.

[2] A.P. Dempster et al, "Maximum likelihood from incomplete data via the EM algorithm", *J. Royal Statistical Soc., Ser. B* 39 (1), 1977, 1-38.

[3] G.R. Doddington et al, "The NIST speaker recognition evaluation - Overview, methodology, systems, results, perspective", *Speech Communication* 31 (2-3), 2000, 225-254.

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

[5] J.A. Nelder and R. Mead, "A simplex method for function minimization", *The Computer Journal* 7 (4), 1965, 308-313.

[6] W.H. Press et al, *Numerical Recipes in C: the Art of Scientific Computing*, Cambridge University Press, Cambridge, 1992.

[7] D.A. Reynolds, "Experimental Evaluation of Features for Robust Speaker Identification", *IEEE Trans. Speech and Audio Processing* 2 (4), 1994, 639-643.

[8] D.A. Reynolds, "Speaker Identification and Verification Using Gaussian Mixture Speaker Models", *Speech Communication* 17 (1-2), 1995, 91-108.

[9] 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 Proc.* 36 (6), 1988, 871-879.

[10] C. Sanderson and K.K. Paliwal, "Multi-Modal Person Verification System Based on Face Profiles and Speech", *Proc. 5th Intern. Symposium on Signal Proc. and its Applic.*, Brisbane, 1999, pp. 947-950.

[11] C. Sanderson and K.K. Paliwal, "Noise Compensation in a Person Verification System Using Face and Multiple Speech Features", *Pattern Recognition* 36 (2), 2003.

[12] E.W. Swokowski, *Calculus (5th ed.)*, PWS-Kent, USA, 1991.

[13] M. Turk and A. Pentland, "Eigenfaces for Recognition", *Journal of Cognitive Neuroscience* 3 (1), 1991, 71-86.

[14] T. Wark et al, "Robust Speaker Verification via Fusion of Speech and Lip Modalities", *Proc. ICASSP*, Phoenix, 1999, pp. 3061-3064.

[15] 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.

[16] J.D. Woodward, "Biometrics: Privacy's Foe or Privacy's Friend?", *Proceedings of the IEEE* 85 (9), 1997, 1480-1492.