IDIAP RESEARCH REPORT

ON THE USE OF SPEECH AND FACE INFORMATION FOR IDENTITY VERIFICATION

Conrad Sanderson (*) (a) (b) Kuldip K. Paliwal (b)

IDIAP–RR 04-10

MARCH 2004

(INITIAL VERSION: SEPTEMBER 2002)

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

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

(*) conradsand@ieee.org
(a) IDIAP, Rue du Simplon 4, CH-1920 Martigny, Switzerland
(b) School of Microelectronic Engineering, Griffith University, Queensland 4111, Australia

IDIAP Research Report 04-10

ON THE USE OF SPEECH AND FACE INFORMATION FOR IDENTITY VERIFICATION

Conrad Sanderson Kuldip K. Paliwal

MARCH 2004
(INITIAL VERSION: SEPTEMBER 2002)

Abstract. This report first provides a review of important concepts in the field of information fusion, followed by a review of important milestones in audio-visual person identification and verification. Several recent adaptive and non-adaptive techniques for reaching the verification decision (i.e., to accept or reject the claimant), based on speech and face information, are then evaluated in clean and noisy audio conditions on a common database; it is shown that in clean conditions most of the non-adaptive approaches provide similar performance and in noisy conditions most exhibit a severe deterioration in performance; it is also shown that current adaptive approaches are either inadequate or utilize restrictive assumptions. A new category of classifiers is then introduced, where the decision boundary is fixed but constructed to take into account how the distributions of opinions are likely to change due to noisy conditions; compared to a previously proposed adaptive approach, the proposed classifiers do not make a direct assumption about the type of noise that causes the mismatch between training and testing conditions.

This report is an extended and revised version of [60].

Keywords: biometrics, information fusion, identity verification, multi-modal, noise resistance.

Contents

List of Figures

Acronyms

EER	Equal Error Rate
ERM	Empirical Risk Minimization
FA	False Acceptance
FA%	False Acceptance rate
fps	frames per second
FR	False Rejection
FR%	False Rejection rate
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
MFCCs	Mel-Frequency Cepstral Coefficients
PCA	Principal Component Analysis
PL	Piece-wise Linear
SNR	Signal to Noise Ratio
SRM	Structural Risk Minimization
SVM	Support Vector Machine
TE	Total Error (defined as TE = FA%+FR%)
UBM	Universal Background Model
VAD	Voice Activity Detector

1 Introduction

A biometric verification (or authentication) system verifies the identity of a claimant based on measures such as the person's face, voice, iris or fingerprints. Apart from various forms of access control (e.g. border control, access to information), verification systems can also be useful in forensic work (where the task is whether a given biometric sample belongs to a given suspect) and law enforcement applications [2, 46, 81]. Recently there has been a lot of interest in multi-modal verification systems [9, 10, 23]; in such systems biometric information from two or more sources is utilized.

The aim of this article is to first provide a review of important concepts in the field of information fusion, which then leads to a review of literature pertaining to audio-visual person identification and verification (Sections 2 and 3, respectively). In the second part of the article we evaluate several recent non-adaptive and adaptive techniques for reaching the verification decision (using speech and face information) in noisy audio conditions on a common database (Sections 4 and 5). We shown that current adaptive approaches are either inadequate or utilize restrictive assumptions. A new category of post-classifiers (which utilize outputs from *modality experts*) is then introduced in Section 6, where the decision boundary is fixed but constructed to take into account the effects of noisy conditions; this approach has the advantage of being simpler than adaptive techniques and able to handle noisy conditions which a previously proposed adaptation technique cannot.

The reader may also be interested in the following articles which cover other important aspects in biometrics (such as front-end signal processing, hiding biometric data, privacy and security issues): [11, 35, 79, 81].

2 Review of Information Fusion Techniques

Broadly speaking, the term *information fusion* encompasses any area which deals with utilizing a combination of different sources of information, either to generate one representational format, or to reach a decision. This includes: consensus building, team decision theory, committee machines, integration of multiple sensors, multi-modal data fusion, combination of multiple experts/classifiers, distributed detection and distributed decision making. It is a relatively new research area, with pioneering publications tracing back to early 1980s [8, 47, 67, 68].

When looking from the point of decision making, there are several motivations for using information fusion:

- Utilizing complementary information (e.g., audio and video) can reduce error rates.
- Use of multiple sensors (i.e., redundancy) can increase reliability.
- Cost of implementation can be reduced by using several cheap sensors rather than one expensive sensor.
- Sensors can be physically separated, allowing the acquisition of information from different points of view.

Humans utilize information fusion every day; some examples are: use of both eyes, seeing and touching the same object, or seeing and hearing a person talk (which improves intelligibility in noisy situations [64]). Several species of snakes combine infrared information with visual information [34, 43].

This section is a review of the most important and common approaches to information fusion. In literature information fusion is often divided into several categories: sensor data level fusion, feature level fusion, score fusion and decision fusion [31, 34, 57]. However, it is more intuitive to classify it into three main categories: *pre-mapping fusion*, *midst-mapping fusion* and *post-mapping fusion*, as shown in Fig. 1. In *pre-mapping fusion*, information is combined before any use of classifiers or experts; in *midst-mapping fusion*, information is combined during mapping from sensor-data/feature space into opinion/decision space, while in *post-mapping fusion*, information is combined after mapping from sensor-data/feature space into opinion/decision space (here the mapping is accomplished by an ensemble of experts or classifiers; while a

classifier provides a hard decision, an expert provides an opinion (e.g., in the [0,1] interval) on each possible decision).

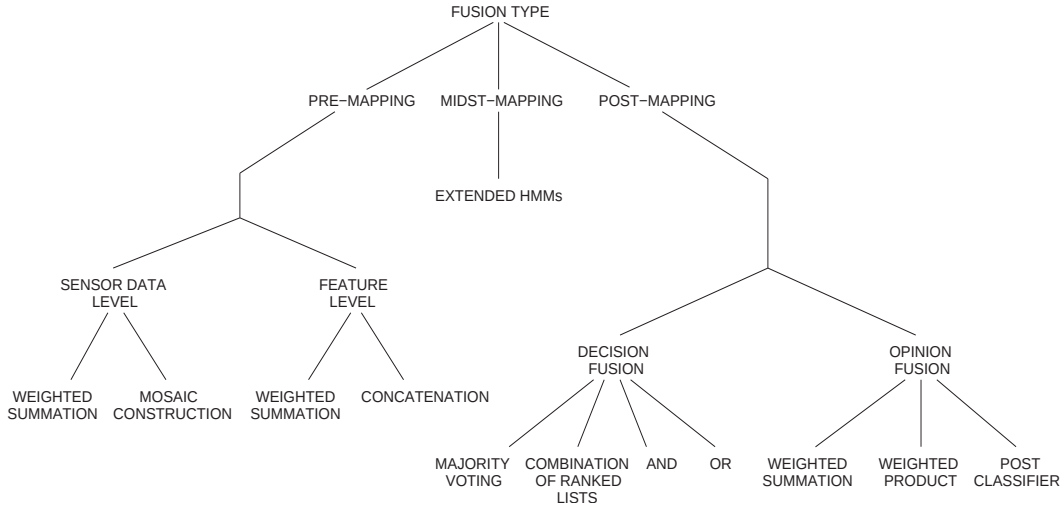

Figure 1: Non-exhaustive tree of fusion types

In *pre-mapping fusion*, there are two main sub-categories: sensor data level fusion and feature level fusion. In *post-mapping fusion*, there are also two main sub-categories: decision fusion and opinion fusion. It must be noted that in some works (e.g., [31, 34, 74]) the term "decision fusion" also encompasses opinion fusion; however, since each expert provides an opinion and not a decision, sub-typing opinion fusion under "decision fusion" is incorrect.

Silsbee and Bovik [64] refer to *pre-mapping fusion* and *post-mapping fusion* as *pre-categorical integration* and *post-categorical integration*, respectively; Wark [78] refers to *pre-mapping fusion* as *input level* or *early fusion* and *post-mapping fusion* as *classifier level* or *late fusion*. Ross and Jain [57] refer to *opinion fusion* as *score fusion*.

In order to aid understanding, the following description of fusion methods is presented in the general context of class identification. Wherever necessary, comments are included to elucidate a fusion approach in terms of the verification application. This section leads onto the review of important milestones in the field of information fusion in audio-visual person recognition (Section 3).

2.1 Pre-mapping Fusion: Sensor Data Level

In sensor data level fusion [31], the raw data from sensors is combined. Depending on the application, there are two main methods to accomplish this: weighted summation and mosaic construction. For example, weighted summation can be employed to combine visual and infra-red images into one image, or, in the form of an average operation, to combine the data from two microphones (to reduce the effects of noise); it must be emphasized that the data must first be commensurate, which can be accomplished by mapping to a common interval. Mosaic construction can be employed to create one image out of images provided by several cameras, where each camera is observing a different part of the same object [34].

2.2 Pre-mapping Fusion: Feature Level

In feature level fusion, features extracted from data provided by several sensors (or from one sensor but using different feature extraction techniques [49]) are combined. If the features are commensurate, the

combination can be accomplished by a weighted summation (e.g., features extracted from data provided by two microphones). If the features are not commensurate, feature vector concatenation can be employed [4, 31, 42, 57], where a new feature vector can be constructed by concatenating two or more feature vectors (e.g., to combine audio and visual features).

There are three downsides to the feature vector concatenation approach. The first is that there is no explicit control over how much each vector contributes to the final decision. The second downside is that the separate feature vectors must be available at the same frame rate (i.e., the feature extraction must be synchronous), which is a problem when combining speech and visual feature vectors[1]. The third downside is the dimensionality of the resulting feature vector, which can lead to the "curse of dimensionality" problem [22]. Due to the above problems, in many cases the post-mapping fusion approach is preferred (described in Sections 2.4 and 2.5).

2.3 Midst-Mapping Fusion

Compared to other fusion techniques presented in this paper, midst-mapping fusion is a relatively new and more complex concept; here several information streams are processed concurrently while mapping from feature space into opinion/decision space. Midst-mapping fusion can be used for exploitation of temporal synergies between the streams (e.g., speech signal and video of lip movements), with the ability to avoid problems present in vector concatenation (such as the "curse of dimensionality" and the requirement of matching frame rates). Examples of this type of fusion are extended Hidden Markov Models (adapted to handle multiple streams of data [9, 50, 52]), which have been shown useful for text-dependent person verification [9, 44, 77].

2.4 Post-Mapping Fusion: Decision Fusion

In decision fusion [31, 34], each classifier in an ensemble of classifiers provides a hard decision. The classifiers can be of the same type but working with different features (e.g., audio and video data), non-homogeneous classifiers working with the same features, or a hybrid of the previous two types. The decisions can be combined by majority voting, combination of ranked lists, or using AND & OR operators.

The inspiration behind the use of non-homogeneous classifiers with the same features stems from the belief that each classifier (due to different internal representation) may be "good" at recognizing a particular set of classes while being "bad" at recognizing a different set of classes; thus a combination of classifiers may overcome the "bad" properties of each classifier [32, 41].

2.4.1 Majority Voting

In majority voting [27, 34, 53], a consensus is reached on the decision by having a majority of the classifiers declaring the same decision. For a two class classification task, the number of classifiers must be odd and greater than two (to prevent ties).

2.4.2 Ranked List Combination

In ranked list combination [3, 32, 53], each classifier provides a ranked list of class labels, with the top entry indicating the most preferred class and the bottom entry indicating the least preferred class. The ranked lists can then be combined via various means [32], possibly taking into account the reliability and discrimination ability of each classifier. The decision is then usually reached by selecting the top entry in the combined ranked list.

[1] For example, speech feature vectors are usually extracted at a rate of 100 per second [48], while visual features are constrained by the video camera's frame rate (25 fps in the PAL standard and 30 fps in the NTSC standard [69]).

2.4.3 AND Fusion

In AND fusion [43, 73], a decision is reached only when all the classifiers agree. As such, this type of fusion is quite restrictive. For multi-class problems no decision may be reached, thus it is mainly useful in situations where one would like to detect the presence of an event/object, with a low false acceptance bias (in a person verification scenario, where we would like to detect the presence of a true claimant, this translates to a high False Rejection rate (FR%) and low False Acceptance rate (FA%)).

2.4.4 OR Fusion

In OR fusion [43, 73], a decision is made as soon as one of the classifiers makes a decision. In comparison to AND fusion, this type of fusion is very relaxed, providing multiple possible decisions in multi-class problems. Since in most multi-class problems this is undesirable, OR fusion is mainly useful where one would like to detect the presence of an event/object with a low false rejection bias (in a person verification scenario, where we would like to detect the presence of a true claimant, this translates to a low FR% and high FA%).

2.5 Post-Mapping Fusion: Opinion Fusion

In opinion fusion [31, 34, 57, 74] (also referred to as score fusion), an ensemble of experts provides an opinion on each possible decision. Since non-homogeneous experts can be used (e.g., where one expert provides its opinion in terms of distances while another in terms of a likelihood measure), the opinions are usually required to be commensurate before further processing. This can be accomplished by mapping the output of each expert to the $[0, 1]$ interval[2], where 0 indicates the lowest opinion and 1 the highest opinion. It must be noted that while the term non-homogeneous usually implies a different expert structure, it is sufficient for a set of experts to be considered non-homogeneous if they are using different features (e.g., audio and video features, or different features extracted from one modality [49]).

In ranked list combination fusion (which doesn't require the mapping step) the rank itself could be considered to indicate the opinion of the classifier. However, compared to opinion fusion, some information regarding the "goodness" of each possible decision is lost.

Opinions can be combined using weighted summation or weighted product approaches (described in Sections 2.5.1 and 2.5.2, respectively) before using a classification criterion, such as the MAX operator (which selects the class with the highest opinion), to reach a decision. Alternatively, a post-classifier (Section 2.5.3) can be used to directly reach a decision. In the former approach, each expert can be considered to be an elaborate discriminant function, working on its own section of the feature space [22].

The inherent advantage of weighted summation and product fusion over feature vector concatenation and decision fusion is that the opinions from each expert can be weighted. 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 SNR conditions (this type of fusion is known as *adaptive fusion*). The weights can also be optimized to satisfy a given criterion (e.g., to obtain EER performance).

2.5.1 Weighted Summation Fusion

In weighted summation, the opinions regarding class j from N_E experts are combined using:

$$f_j = \sum_{i=1}^{N_E} w_i o_{i,j} \tag{1}$$

[2]The mapping can be performed via a sigmoid; see Section 4.4 for more information.

where $o_{i,j}$ is the opinion from the i-th expert and w_i is the corresponding weight in the $[0, 1]$ interval, with the constraint $\sum_{i=1}^{N_E} w_i = 1$. When all the weights are equal, Eqn. (1) reduces to an arithmetic mean operation. The weighted summation approach is also known as *linear opinion pool* [6] and *sum rule* [5, 41].

2.5.2 Weighted Product Fusion

The opinions can be interpreted as posterior probabilities in the Bayesian framework [13]. Assuming the experts are independent, the opinions regarding class j from N_E experts can be combined using a product rule:

$$f_j = \prod_{i=1}^{N_E} o_{i,j} \tag{2}$$

Moreover, to account for varying discrimination ability and reliability of each expert, weighting is introduced:

$$f_j = \prod_{i=1}^{N_E} (o_{i,j})^{w_i} \tag{3}$$

When all the weights are equal, Eqn. (3) reduces to a geometric mean operation. The weighted product approach is also known as *logarithmic opinion pool* [6] and *product rule* [5, 41].

There are two downsides to weighted product fusion: the first is that one expert can have a large influence over the fused opinion - for example, an opinion close to zero from one expert sets the fused opinion also close to zero. The second downside is that the independence assumption is only strictly valid when each expert is using independent features.

2.5.3 Post-Classifier

Since the opinions produced by the experts indicate the "likelihood" of a particular class, the opinions can be considered as features in "likelihood space". The opinions from N_E experts regarding N_C classes form a $N_E N_C$-dimensional opinion vector, which is used by a classifier to make the final decision. We shall refer to such a classifier as a *post-classifier*[3]. It must be noted that the opinions do not necessarily have to be commensurate, as it is the post-classifier's job to provide adequate mapping from the "likelihood space" to class label space.

The obvious downside of this approach is that the resultant dimensionality of the opinion vector is dependent on the number of experts as well as the number of classes, which can be quite large in some applications. However, in a verification application, the dimensionality of the opinion vector is usually only dependent on the number of experts [10]. Each expert provides only one opinion, indicating the likelihood that a given claimant is the true claimant (thus a low opinion suggests that the claimant is an impostor, while a high opinion suggests that the claimant is the true claimant). The post-classifier then provides a decision boundary in N_E-dimensional space, separating the impostor and true claimant classes[4].

2.5.4 Special Case of Equivalence of Weighted Summation and Post-Classifier Approaches

In a normal verification application, there are only two classes (i.e., true claimants and impostors) and each expert provides only one opinion (i.e., high opinion suggests a true claimant while a low opinion suggests an impostor). Once the fused score is obtained using the weighted summation approach the accept/reject decision

[3] In the identification scenario, the described post-classifier is a natural extension of the approach presented in [7]. In the verification scenario it has been implemented by Ben-Yacoub *et al.* [10] as a binary classifier.

[4] see Fig. 6 for example decision boundaries.

can be 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. (1) can thus be modified to:

$$F(\mathbf{o}) = \mathbf{w}^T \mathbf{o} - t \tag{4}$$

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

It can be seen that Eqn. (4) is a form of a linear discriminant function [22], indicating that the procedure of weighted summation followed by thresholding creates a linear decision boundary in N_E-dimensional space. Thus in the verification application, weighted summation fusion is equivalent to a post-classifier which uses a linear decision boundary to separate the true claimant and impostor classes.

2.6 Hybrid Fusion

For certain applications, it may be necessary to combine various fusion techniques due to practical considerations. For example, Hong and Jain [33] used a fingerprint expert and a frontal face expert; a hybrid fusion scheme involving a ranked list and opinion fusion was used: opinions of the face expert for the top n identities were combined with the opinions of the fingerprint expert for the corresponding identities using a form of the product approach. This hybrid approach was used to take into account the relative computational complexity of the fingerprint expert (i.e., the fingerprint expert was significantly slower than the face expert).

3 Important Milestones in Audio-Visual Person Recognition

This section provides an overview of the most important contributions in the field of audio-visual person recognition; it is assumed that the reader is familiar with the concepts presented in Section 2. We concentrate on the verification task while briefly touching on the identification task. Almost all of the work reviewed here used different databases and/or different experimental setup (e.g., experts and performance measures), thus any direct comparison between the numerical results would be meaningless. Numerical figures are only shown in the first few cases to demonstrate that using fusion increases performance. Moreover, no thorough description of the various experts used is provided, as it is beyond the scope of this section.

The review is split into two areas: non-adaptive (Section 3.1) and adaptive (Section 3.2) approaches. In non-adaptive approaches, the contribution of each expert is priorly fixed. In adaptive approaches, the contribution of at least one expert is varied according to its reliability and discrimination ability in the presence of some environmental condition; for example, the contribution of a speech expert can be decreased when the audio SNR is lowered.

3.1 Non-Adaptive Approaches

Fusion of audio and visual information has been applied to automatic person recognition in pioneering papers by Chibelushi *et al.* [18] in 1993 and Brunelli *et al.* [12, 13] in 1995.

Chibelushi *et al.* [18] combined information from speech and still face profile images using a form of weighted summation fusion:

$$f = w_1 o_1 + w_2 o_2 \tag{5}$$

where o_1 and o_2 are the opinions from the speech and face profile experts, respectively, with corresponding weights w_1 and w_2. Each 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). Since there are constraints on the weights ($\sum_{i=1}^{2} w_i = 1$ and $\forall i : w_i \geq 0$), Eqn. (5) reduces to:

$$f = w_1 o_1 + (1 - w_1) o_2 \tag{6}$$

The verification decision was reached via thresholding the fused opinion, f. When using the speech expert alone (i.e., $w_1 = 1$), an Equal Error Rate (EER) of 3.4% was achieved, while when using the face profile expert alone (i.e., $w_1 = 0$), an EER of 3.0% was obtained. Using an optimal weight and threshold (in the EER sense) the EER was reduced to 1.5%.

Brunelli *et al.* [12] combined the opinions from a face expert (which utilized geometric features obtained from static frontal face images) and a speech expert using the weighted product approach:

$$f = (o_1)^{w_1} \times (o_2)^{(1-w_1)} \tag{7}$$

When the speech expert was used alone (i.e., $w_1 = 1$), an identification rate of 51% was obtained, while when the face expert was used alone (i.e., $w_1 = 0$), an identification rate of 92% was achieved. Using an optimal weight, the identification rate increased to 95%.

In [13], two speech experts (for static and delta features) and three face experts (for the eye, nose and mouth areas of the face) were used for person identification. The weighted product approach was used to fuse the opinions, with the weights found automatically via a heuristic approach. The static and dynamic feature experts obtained an identification rate of 77% and 71%, respectively. Combining the two speech experts increased the identification rate to 88%. The eye, nose and mouth experts obtained an identification rate of 80%, 77% and 83%, respectively. Combining the three facial experts increased the identification rate to 91%. When all five experts were used, the identification rate increased to 98%.

Dieckmann *et al.* [20] used three experts (frontal face expert, dynamic lip image expert and text-dependent speech expert). A hybrid fusion scheme involving majority voting and opinion fusion was utilized; two of the experts had to agree on the decision and the combined opinion had to exceed a pre-set threshold. The hybrid fusion scheme provided better performance than using the underlying experts alone.

Kittler *et al.* [40] used one frontal face expert which provided one opinion for one face image. Multiple images of one person were used to generate multiple opinions, which were then fused by various means, including averaging (a special case of weighted summation fusion). It was shown that error rates were reduced by up to 40% and that performance gains tended to saturate after using five images (however, no results were provided for using more than six images). The results suggest that using a video sequence of the face, rather than one image, provides superior performance.

In further work, Kittler *et al.* [41] attempted to provide theoretical foundations for common fusion approaches such as the summation and product methods. However, by the authors' own admission, the foundations utilized assumptions which are "unrealistic in most applications". Experimental results for combining the opinions from three experts (two face experts (frontal and profile) and a text-dependent speech expert) showed that the summation approach outperformed the product approach.

Luettin [42] investigated the combination of speech and (visual) lip information using feature vector concatenation. In order to match the frame rates of both feature sets, speech information was extracted at 30 fps instead of the usual 100 fps. In text-dependent configuration, the fusion process resulted in a minor performance improvement, however, in text-independent configuration, the performance slightly decreased; this suggests that feature vector concatenation in this case is unreliable.

Jourlin *et al.* [38, 39] used a form of weighted summation fusion to combine the opinions of two experts: a text-dependent speech expert and a text-dependent lip expert. Using an optimal weight, fusion led to better performance than using the underlying experts alone.

Abdeljaoued [1] proposed to use a Bayesian post-classifier to reach the verification decision. Formally, the decision rule is expressed as:

$$\text{chosen class} = \begin{cases} C_1 & \text{if} \quad \prod_{i=1}^{N_E} p(o_i|\lambda_{i,\text{true}}) > \prod_{i=1}^{N_E} p(o_i|\lambda_{i,\text{imp}}) \\ C_2 & \text{otherwise} \end{cases} \tag{8}$$

where C_1 and C_2 are true claimant and impostor classes, respectively, N_E is the number of experts, while $\lambda_{i,\text{true}}$ and $\lambda_{i,\text{imp}}$ are, for the i-th expert, the parametric models of the distribution of opinions for true claimant and impostor claims, respectively[5]. Due to precision issues in a computational implementation, it is more convenient to use a summation rather than a series of multiplications. Since $\log(\cdot)$ is a monotonically increasing function, the decision rule can be modified to:

$$\text{chosen class} = \begin{cases} C_1 & \text{if} \quad \sum_{i=1}^{N_E} \log p(o_i|\lambda_{i,\text{true}}) > \sum_{i=1}^{N_E} \log p(o_i|\lambda_{i,\text{imp}}) \\ C_2 & \text{otherwise} \end{cases} \tag{9}$$

To allow adjustment of FA% and FR%, the above decision rule is in practice modified by introducing a threshold:

$$\text{chosen class} = \begin{cases} C_1 & \text{if} \quad \sum_{i=1}^{N_E} \log p(o_i|\lambda_{i,\text{true}}) - \sum_{i=1}^{N_E} \log p(o_i|\lambda_{i,\text{imp}}) > t \\ C_2 & \text{otherwise} \end{cases} \tag{10}$$

Abdeljaoued used three experts and showed that use of the above classifier (with Beta distributions) provided lower error rates than when using the experts alone.

Ben-Yacoub *et al.* [10] investigated the use of several binary classifiers for opinion fusion using a post-classifier. The investigated classifiers were: Support Vector Machine (SVM), Bayesian classifier (using Beta distributions), Fisher's Linear Discriminant, Decision Tree and Multi Layer Perceptron (MLP). Three experts were used: a frontal face expert and two speech based experts (text-dependent and text-independent). It was found that the SVM classifier (using a polynomial kernel) and the Bayesian classifier provided the best results.

Verlinde [74] also investigated various binary classifiers for opinion fusion as well as the majority voting and AND & OR fusion methods (which fall in the decision fusion category). Three experts were used: frontal face expert, face profile expert and a text-independent speech expert. In the case of decision fusion, each expert acted like a classifier and provided a hard decision rather than an opinion. The investigated classifiers were: Decision Tree, MLP, Logistic Regression (LR) based classifier, Bayesian classifier using Gaussian distributions, Fisher's Linear Discriminant and various forms of the *k*-Nearest Neighbour classifier. Verlinde found that the LR based classifier (which created a linear decision surface) provided the lowest overall error rates as well as being the easiest to train. Verlinde also attempted to develop a piece-wise linear classifier but obtained poor results.

Wark *et al.* [75] used the weighted summation approach to combine the opinions of a speech expert and a lip expert (both text-independent). The performance of the speech expert was deliberately decreased by adding varying amounts of white noise to speech data (where the SNR varied from 50 to 10 dB). Experimental results showed that although the performance of the system was always better than using the speech expert alone, it significantly decreased as the noise level increased. Depending on the values of the weights (which were priorly selected), the performance in high noise levels was actually worse than using the lip expert alone (a condition referred to as *catastrophic fusion* [78]). The authors proposed a statistically inspired method of priorly selecting weights (described below) which resulted in good performance in clean conditions and never fell below the performance of the lip expert in noisy conditions; however, the performance in noisy conditions was shown not to be optimal and no results were reported for SNR levels below 10 dB; moreover, the performance (for each noise level) was found using only 30 true claimant tests and 210 impostor tests.

The weight for the speech expert was found as follows:

$$w_1 = \frac{\zeta_2}{\zeta_1 + \zeta_2} \tag{11}$$

[5] In our experiments we utilize Gaussian Mixture Models to model the distribution of opinions; see Section 4.2 for more information.

where

$$\zeta_i = \sqrt{\frac{\sigma_{i,true}^2}{N_{true}} + \frac{\sigma_{i,imp}^2}{N_{imp}}} \tag{12}$$

where, for the *i*-th expert, ζ_i is the standard error [16] of the difference between sample means $\mu_{i,true}$ and $\mu_{i,imp}$ of opinions for true and impostor claims, respectively, $\sigma_{i,true}^2$ and $\sigma_{i,imp}^2$ are the corresponding variances, while N_{true} and N_{imp} is the number of opinions for true and impostor claims, respectively. Wark *et al.* referred to ζ_i as a prior confidence. Since there are constraints on the weights ($\sum_{i=1}^{2} w_i = 1$ and $\forall i : w_i \geq 0$), the weight for the lip expert is $1 - w_1$.

Wark *et al.* assumed that the standard error gives relative indication of the discrimination ability of an expert. The less variation there is in the opinions for known true and impostor claims, the lower the standard error; thus a low standard error indicates better performance.

Multi-Stream Hidden Markov Models (MS-HMMs) (a form of midst-mapping fusion) were evaluated for the task of text-dependent audio-visual person identification in [77]. The audio stream was comprised of a sequence of vectors containing Mel Frequency Cepstral Coefficients (MFCCs) [55] and their deltas [65], while the video stream was comprised of a sequence of feature vectors describing lip contours. Due to the nature of the MS-HMM implementation the frame rate of the video features had to match the frame rate of the audio features (accomplished by up-sampling). Experiments on a small audio-visual database showed that for high SNRs the performance was comparable to that of an audio-only HMM system (which outperformed the video-only HMM system), while at low SNRs the multi-stream system obtained significantly better performance than the audio-only system and exceeded the performance of the video-only system. No comparison was given against a system utilizing pre-mapping or post-mapping fusion (e.g., utilizing two separate experts and opinion fusion).

Bengio [9] addressed several limitations of previous MS-HMM systems, allowing the two streams to be temporarily desynchronized (since related events in the streams may start and/or end at different points, e.g., lip movement can start before speech is heard) and have different frame rates (thus up-sampling is no longer required). Experiments on a small audio-visual database (using two feature streams similar to the audio and video streams described for [77], above) showed that while at a relatively high SNR the performance was worse than a text-independent audio-only system, the performance was better at lower SNRs; moreover, the proposed system had higher performance (and was more robust) than a text-dependent HMM system based on feature vector concatenation.

3.2 Adaptive Approaches

Wark *et al.* [76] extended the work presented in [75] (see above) by proposing a heuristic method to adjust the weights. Experimental results showed that although the performance significantly decreased as the noise level increased, it was always better than using the speech expert alone. However, in high noise levels, equal weights (non-adaptive) were shown to provide better performance. A major disadvantage of the method is that the calculation of the weights involved finding the opinion of the speech expert for all possible claims (i.e., for all persons enrolled in the system), thus limiting the approach to systems with a small number of clients due to practical considerations (i.e., time taken to verify a claim). Moreover, similar experimental limitations were present as described for [75] (above).

In further work [78], Wark proposed another heuristic technique of weight adjustment (described below). In a text-dependent configuration, the system provided performance which was always better than using the lip expert alone. However, in a text-independent configuration, the performance in low SNR conditions was worse than using the lip expert alone.

The weight for the speech expert was found as follows:

$$w_1 = \left[\frac{\zeta_2}{\zeta_1 + \zeta_2}\right]\left[\frac{\kappa_1}{\kappa_1 + \kappa_2}\right] \tag{13}$$

where $\frac{\zeta_2}{\zeta_1+\zeta_2}$ was found using Eqn. (12) during training and

$$\kappa_i = \frac{|\mathcal{M}(o_i)_{i,true} - \mathcal{M}(o_i)_{i,imp}|}{\mu_{i,true}} \tag{14}$$

was found during testing. Wark referred to κ_i as the posterior confidence. For the *i*-th expert, $\mathcal{M}(o_i)_{i,true} = \frac{(o_i-\mu_{i,true})^2}{\sigma^2_{i,true}}$ is the one-dimensional squared Mahalanobis distance [22] between opinion o_i and the model of opinions for true claims. Here, $\mu_{i,true}$ and $\sigma^2_{i,true}$ are the mean and variance of opinions for true claims, respectively; they are found during training.

Similarly, $\mathcal{M}(o_i)_{i,imp} = \frac{(o_i-\mu_{i,imp})^2}{\sigma^2_{i,imp}}$ is the one-dimensional squared Mahalanobis distance between opinion o_i and the model of opinions for impostor claims. Here, $\mu_{i,imp}$ and $\sigma^2_{i,imp}$ are the mean and variance of opinions for impostor claims, respectively; they are found during training.

Under clean conditions, the distance between a given opinion for a true claim and the model of opinions for true claims 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 used empirical evidence to argue that under noisy conditions, the distances should decrease, hence κ_i should decrease.

We recently proposed [61] a weight adjustment method which is summarized as follows. Every time a speech utterance is recorded, it is usually preceded by a short segment which contains only ambient noise. From each training utterance, Mel Frequency Cepstral Coefficients (MFCCs) [48, 55] from the noise segment are used to construct a global noise Gaussian Mixture Model (GMM), λ_{noise}. Given a test speech utterance, N_{noise} MFCC feature vectors, $\{\mathbf{x}_i\}_{i=1}^{N_{\text{noise}}}$, representing the noise segment, are used to estimate the utterance's quality by measuring the mismatch from λ_{noise} as follows:

$$q = \frac{1}{N_{\text{noise}}} \sum_{i=1}^{N_{\text{noise}}} \log p(\mathbf{x}_i|\lambda_{\text{noise}}) \tag{15}$$

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 sigmoid:

$$q_{\text{map}} = \frac{1}{1 + \exp[-a(q-b)]} \tag{16}$$

where a and b describe the shape of the sigmoid. The values of a and b are manually 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 (which is found on clean data [to achieve, for example, EER performance]), the adapted weight for the speech expert is found using:

$$w_2 = q_{\text{map}} w_{2,\text{apriori}} \tag{17}$$

Since we are using a two modal system the corresponding weight for the face expert is found using: $w_1 = 1 - w_2$. We shall refer to this weight adjustment method as the *mismatch detection* method.

4 Performance of Non-Adaptive Approaches in Noisy Audio Conditions

In this section we evaluate the performance of feature vector concatenation fusion and several non-adaptive opinion fusion methods (weighted summation fusion, Bayesian and SVM post-classifiers), for combining face and speech information under the presence of audio noise.

4.1 VidTIMIT Audio-Visual Database

The VidTIMIT database [59] is comprised of video and corresponding audio recordings of 43 people (19 female and 24 male), reciting short sentences selected from the NTIMIT corpus [36]. 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 seconds, or approximately 106 video frames.

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 (columns $\times$ rows); the corresponding audio is stored as a mono, 16 bit, 32 kHz WAV file.

4.2 Speech Expert

The speech expert is comprised of two main components: speech feature extraction and a Gaussian Mixture Model (GMM) 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 (MFCC), which reflect the instantaneous Fourier spectrum [48, 55], their corresponding deltas (which represent transitional spectral information) [65] and Maximum Auto-Correlation Values (which represent pitch and voicing information) [80]. Cepstral mean subtraction was applied to MFCCs [24, 55]. The sequence of feature vectors is then processed by a parametric Voice Activity Detector (VAD) [29, 30], which removes feature vectors that are considered to represent silence or background noise.

The distribution of feature vectors for each person is modeled by a GMM. Given a claim for person C's identity and a set of feature vectors $X = \{\mathbf{x}_i\}_{i=1}^{N_V}$ supporting the claim, the average log-likelihood of the claimant being the true claimant is found with:

$$\mathcal{L}(X|\lambda_C) = \frac{1}{N_V} \sum_{i=1}^{N_V} \log p(\mathbf{x}_i|\lambda_C) \tag{18}$$

where

$$p(\mathbf{x}|\lambda) = \sum_{j=1}^{N_G} m_j \, \mathcal{N}(\mathbf{x}; \mu_j, \mathbf{\Sigma}_j) \tag{19}$$

$$\lambda = \{m_j, \mu_j, \mathbf{\Sigma}_j\}_{j=1}^{N_G} \tag{20}$$

Here λ_C is the parameter set[6] for client C, N_G is the number of Gaussians, m_j is the weight for Gaussian j (with constraints $\sum_{j=1}^{N_G} m_j = 1$ and $\forall\, j : m_j \geq 0$). $\mathcal{N}(\mathbf{x}; \mu, \mathbf{\Sigma})$ is a multi-variate Gaussian function with mean

[6] We use the terms *parameter set* and *model* interchangeably.

μ and diagonal covariance matrix $\mathbf{\Sigma}$:

$$\mathcal{N}(\mathbf{x};\mu,\mathbf{\Sigma}) = \frac{1}{(2\pi)^{\frac{D}{2}}|\mathbf{\Sigma}|^{\frac{1}{2}}} \exp\left[\frac{-1}{2}(\mathbf{x}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right] \tag{21}$$

where D is the dimensionality of $\mathbf{x}$. Given the average log-likelihood of the claimant being an impostor, $\mathcal{L}(X|\lambda_{\overline{C}})$, an opinion on the claim is found using:

$$\mathcal{O}(X|\lambda_C,\lambda_{\overline{C}}) = \mathcal{L}(X|\lambda_C) - \mathcal{L}(X|\lambda_{\overline{C}}) \tag{22}$$

The verification decision is reached as follows: given a threshold t, the claim is accepted when $\mathcal{O}(X|\lambda_C,\lambda_{\overline{C}}) \geq t$ and rejected when $\mathcal{O}(X|\lambda_C,\lambda_{\overline{C}}) < t$. 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 mono-modal systems, the opinion can be thresholded to achieve the final verification decision.

4.2.1 Estimation of Model Parameters (Training)

First, a Universal Background Model (UBM) is trained using the Expectation Maximization (EM) algorithm [19, 22][7]; as it is a good representation of the general population [56], it is also used to find the average log-likelihood of the claimant being an impostor, i.e.:

$$\mathcal{L}(X|\lambda_{\overline{C}}) = \mathcal{L}(X|\lambda_{ubm}) \tag{23}$$

The parameters (λ) for each client model are then found by using the client's training data and adapting the UBM using a form of Maximum *a Posteriori* adaptation [26, 56].

4.3 Face Expert

The face expert is similar to the speech expert; the main difference is in the feature extraction method. Here we use the common Principal Component Analysis (PCA) technique [70] (also known as eigenfaces), which is holistic in nature (that is, one face image yields one feature vector)[8].

Before facial feature extraction can occur, the face must first be located [17]. 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[9]. Using the distance between the eyes as a size measure, an affine transformation is used [28] to adjust the size of the image, resulting in the distance between the eyes to be the same for each person. Finally a 64×56 pixel (columns $\times$ rows) face window, containing the eyes and the nose (the most invariant face area to changes in the expression and hair style) is extracted from the image.

PCA based feature extraction is performed as follows. A given size normalized face image is represented by a matrix containing grey level pixel values; the matrix is then converted to a face vector, $\mathbf{v}$, by concatenating all the columns; a D-dimensional feature vector, $\mathbf{x}$, is then obtained by:

$$\mathbf{x} = \mathbf{U}^T(\mathbf{v} - \mathbf{v}_\mu) \tag{24}$$

[7] We used 20 iterations of EM algorithm; Reynolds [54] showed that the EM algorithm generally converges in 10 to 15 iterations, with further iterations resulting in only very minor improvements.

[8] Non-holistic (local) face features can also be effectively used with the GMM opinion generator [15, 62, 63].

[9] A "mother" prototype face was constructed by averaging manually extracted and size normalized faces from clients (non-impostors) in the VidTIMIT database; prototype faces of various sizes were constructed by applying an affine transform to the "mother" prototype face.

where $\mathbf{U}$ contains D eigenvectors (corresponding to the D largest eigenvalues) of the training data covariance matrix, and $\mathbf{v}_\mu$ is the mean of training face vectors. In our experiments we use training images from all clients (i.e. excluding impostors) find $\mathbf{U}$ and $\mathbf{v}_\mu$; moreover, $D = 20$. Preliminary experiments showed that while $D = 30$ obtained optimal face verification, the performance was not improved further with the use of fusion; since in this paper we wish to evaluate how noisy audio conditions degrade fusion performance, we deliberately detuned the face expert so that fusion had a positive effect on performance in clean conditions.

4.4 Mapping Opinions to the [0,1] Interval

The experiments reported throughout this paper utilize the following method (inspired by [38]) of mapping the output of each expert to the $[0, 1]$ interval.

The original opinion of expert i, $o_{i,\text{orig}}$, is mapped to the $[0, 1]$ interval using a sigmoid:

$$o_i = \frac{1}{1 + \exp[-\tau_i(o_{i,\text{orig}})]} \tag{25}$$

where

$$\tau_i(o_{i,\text{orig}}) = \frac{o_{i,\text{orig}} - (\mu_i - 2\sigma_i)}{2\sigma_i} \tag{26}$$

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 $\mathcal{N}(o_{i,\text{orig}}; \mu_i, {\sigma_i}^2)$ and $\mathcal{N}(o_{i,\text{orig}}; \mu_i - 4\sigma_i, {\sigma_i}^2)$ 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 [22] (see also Fig. 2). Eqn. (26) maps the opinions to the $[-2, 2]$ interval, which corresponds to the approximately linear portion of the sigmoid in Eqn. (25). The sigmoid is necessary to take care of situations where the assumptions do not hold entirely.

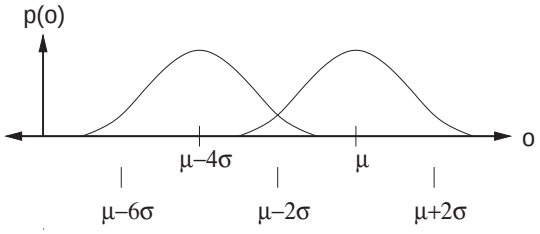

Figure 2: Graphical interpretation of the assumptions used in Section 4.4.

4.5 Support Vector Machine Post-Classifier

The Support Vector Machine (SVM) [71] has been previously used by Ben-Yacoub *et al.* [10] as a post-classifier. While an in-depth description of SVM is beyond the scope of this section, important points are summarized; for more detail, the reader is referred to [14].

The SVM is based on the principle of Structural Risk Minimization (SRM) as opposed to Empirical Risk Minimization (ERM) used in classical learning approaches. Under ERM, without testing on a separate data set, it is unknown which decision surface would have a good generalization capability. For the case of the SVM, the decision surface has to satisfy a structural requirement which is thought to obtain the best generalization capability. For example, let us assume we have a set of training vectors belonging to two completely separable classes and we seek a linear decision surface that separates the classes. Let us define the term *margin* as the sum of distances from the decision surface (in the space implied by the employed kernel, see below) to the closest points of the two classes; we interpret the meaning of the margin as a measure of generalization capability. Thus using the SRM principle, the optimal decision surface has the maximum *margin*.

The SVM is inherently a binary classifier. Let us define a set S containing N_V opinion vectors (N_E-dimensional) belonging to two classes labeled as -1 and $+1$, indicating impostor and true claimant classes respectively:

$$S = \{ \ (\mathbf{o}_i, y_i) \mid \mathbf{o}_i \in \mathbb{R}^{N_E}, y_i \in \{-1, +1\} \ \}_{i=1}^{N_V} \tag{27}$$

The SVM uses the following function, which implements the optimal decision surface in SRM sense [71], to map a given vector to its label space (i.e., -1 or $+1$):

$$f(\mathbf{o}) = \text{sign}\left(\sum_{i=1}^{N_V} \alpha_i y_i K(\mathbf{o}_i, \mathbf{o}) + b\right) \tag{28}$$

where vectors $\mathbf{o}_i$ with corresponding $\alpha_i > 0$ are known as *support vectors*. $K(\mathbf{d}, \mathbf{e})$ is a symmetric kernel function, subject to Mercer's condition [14, 71]. $\alpha^T = [\alpha_i]_{i=1}^{N_V}$ is found by minimizing (via quadratic programming):

$$-\sum_{i=1}^{N_V} \alpha_i + \frac{1}{2}\alpha^T D \alpha \tag{29}$$

subject to constraints:

$$\alpha^T \mathbf{y} = 0 \tag{30}$$

$$\alpha_i \in [0, C] \ \forall \, i \tag{31}$$

where, $\mathbf{y}^T = [\ y_i \ ]_{i=1}^{N_V}$ and C is a large positive value (e.g., 1000); C is utilized to allow training with non-separable data. The elements of D are defined as:

$$D_{ij} = y_i y_j K(\mathbf{o}_i, \mathbf{o}_j) \tag{32}$$

The parameter b is found after α has been found [14]. The kernel function $K(\mathbf{d}, \mathbf{e})$ usually implements a dot product in a high dimensional space, $\mathbb{R}^h$ (where $h > N_E$), which can improve separability of the data [58]. Popular kernels used for pattern recognition problems are [14]:

$$K(\mathbf{d}, \mathbf{e}) = \mathbf{d}^T \mathbf{e} \tag{33}$$

$$K(\mathbf{d}, \mathbf{e}) = (\mathbf{d}^T \mathbf{e} + 1)^p \tag{34}$$

$$K(\mathbf{d}, \mathbf{e}) = \exp(-\frac{1}{\sigma^2}||\mathbf{d} - \mathbf{e}||^2) \tag{35}$$

Eqn. (33) is a dot product, which is referred to as the linear kernel, Eqn. (34) is a p-th degree polynomial, while Eqn. (35) is a gaussian kernel (where σ represents the standard deviation of the kernel).

The experiments reported in this section utilize the SVM engine developed by Joachims [37]. In a verification system there is generally more training data for the impostor class than the true claimant class; thus a misclassification on the impostor class (i.e., a FA error) has less contribution toward the EER than a misclassification on the true claimant class (i.e., a FR error). Hence standard SVM training, which in the non-separable case minimizes the *total* misclassification rate (subject to SRM constraints), is not compatible with the EER criterion. Fortunately, Joachims' SVM engine allows setting of an appropriate cost of making an error on either class; while this does not explicitly guarantee training for EER, the cost can be tuned manually until performance close to EER is obtained.

4.6 Experiments

The experiments were done on the VidTIMIT database (see Section 4.1); the speech and frontal face experts are described in Sections 4.2 and 4.3, respectively. For the speech expert, best results on clean test data[10] were obtained with 32-Gaussian client models. For the face expert, best results were obtained with one-Gaussian client models.

Session 1 was used as the training data. To find the performance, Sessions 2 and 3 were used for obtaining expert opinions of known impostor and true claims. Four utterances, each from eight fixed persons (four male and four female), were used for simulating impostor accesses against the remaining 35 persons. 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 resulting SNR varying from 12 to -8 dB; SNR of -8 dB was chosen as the end point as preliminary experiments showed that at this SNR the EER of the speech expert was close to chance level. In the second set of experiments, speech signals were corrupted speech signals were corrupted by adding "operations-room" noise from the NOISEX-92 corpus [72]; the "operations-room" noise contains background speech as well as machinery sounds. Again, the resulting SNR varied from 12 to -8 dB.

Performance of the following configurations was found: speech expert alone, face expert alone, feature vector concatenation, weighted summation fusion (equivalent to a post-classifier with a linear decision boundary), the Bayesian post-classifier and the SVM post-classifier. For the latter three approaches, the face expert provided the first opinion (o_1) while the speech expert provided the second opinion (o_2) when forming the opinion vector $\mathbf{o} = [\ o_1\ o_2\ ]^T$.

The parameters for weighted summation fusion were found via an exhaustive search procedure. For the Bayesian post-classifier, two Gaussians were used to model the distribution of opinion vectors (one Gaussian each for true claimant and impostor distributions); multiple Gaussians for each distribution, i.e. GMMs, were also evaluated but did not provide performance advantages. For the SVM post-classifier, the linear kernel [see Eqn. (33)] was used; other kernels were also evaluated but did not provide performance advantages.

As described in Section 2.2, the basic idea of the feature vector concatenation 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 fps while the standard frame rate for video is 25 fps (using off the shelf commercial PAL 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. It is also possible to decrease the frame rate of the speech features, but this would result in less speech information being available, decreasing performance [42]. Thus in the experiments reported in this section, the information loss is avoided by utilizing the former approach of artificially increasing the video frame rate. As done by the speech expert, the feature vectors resulting from feature vector concatenation were processed by the VAD (Section 4.2). Best results on clean data were obtained with one-Gaussian client models.

The equivalency described in Section 2.5.4 has several implications on the measurement of performance of multi-expert systems. In speech based verification systems, the Equal Error Rate (EER) is often used as a measure of expected performance [21, 25]. In a single expert configuration this amounts to selecting the appropriate posterior threshold so that the False Acceptance rate (FA%) is equal to the False Rejection rate (FR%); in a multi-expert scenario this translates to selecting appropriate posterior parameters for opinion mapping (Section 4.4) and for the post-classifier (in the weighted summation case the parameters are $\mathbf{w}$ and t). In a multi-expert *adaptive* system, the weights are automatically tuned in an attempt to account the current reliability of one or more experts (as in the system proposed by Wark [78]). Tuning the threshold to obtain

[10] By *clean data* we mean original data which has not been artificially corrupted with noise.

EER performance is equivalent to modifying one of the parameters of the post-classifier, which is in effect *further adaptation* of the post-classifier after observing the effect that the weights have on the distribution of f [Eqn. (1)] for true and impostor claims. Since this cannot be accomplished in real life, it is a fallacy to report the performance in *noisy conditions* in terms of EER for an *adaptive* multi-expert system.

Taking into account the above argumentation and to keep the presentation of results consistent between non-adaptive and adaptive systems, the results in this paper are reported in the following manner. The post-classifier is tuned for EER performance on clean test data (analogous to the popular practice of using the posterior threshold in single-expert systems [21, 25]); performance in clean and noisy conditions is then reported in terms of Total Error (TE), defined as:

$$\text{TE} = \text{FA\%} + \text{FR\%} \tag{36}$$

where the post-classifier parameters are fixed (in non-adaptive systems), or automatically varied (in adaptive systems). We note that posterior selection of parameters (for clean data) puts an optimistic bias on the results; however, since we wish to evaluate how noisy audio conditions degrade fusion performance, we would like to have an optimal starting point.

Performance of the face and speech experts is shown in Fig. 3; performance of the four multi-modal systems is shown in Fig. 4 for white noise, and in Fig. 5 for "operations-room" noise. Figures 6 and 7 show the distribution of opinion vectors in clean and noisy (SNR = -8 dB) conditions (white noise), respectively, with the decision boundaries used by the three post-classifier approaches.

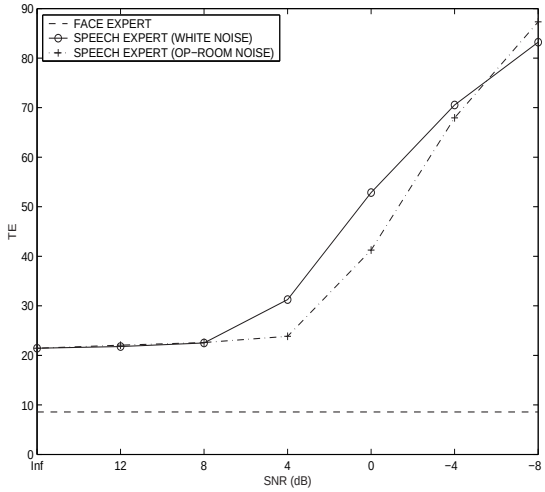

Figure 3: Performance of the speech and face experts.

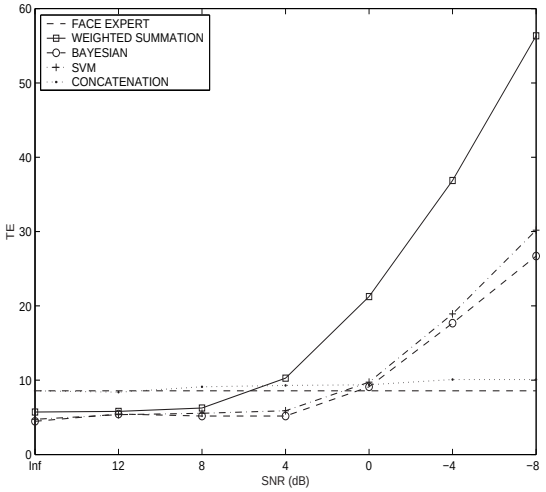

Figure 4: Performance of non-adaptive fusion techniques in the presence of white noise.

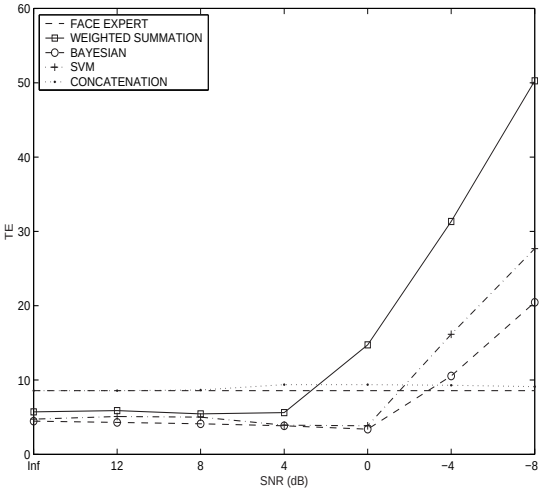

Figure 5: Performance of non-adaptive fusion techniques in the presence of operations-room noise.

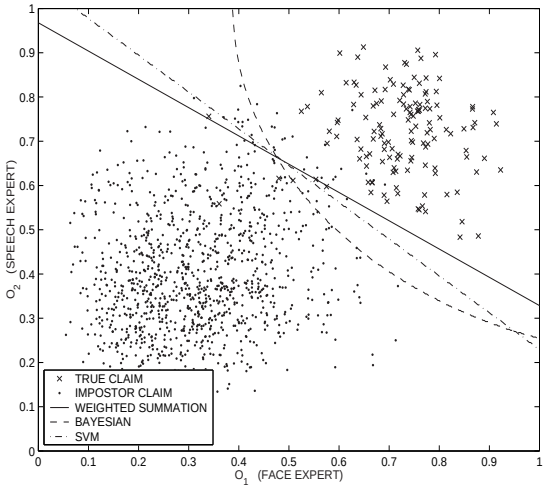

Figure 6: Decision boundaries used by fixed post-classifier fusion approaches and the distribution of opinion vectors for true and impostor claims (clean speech).

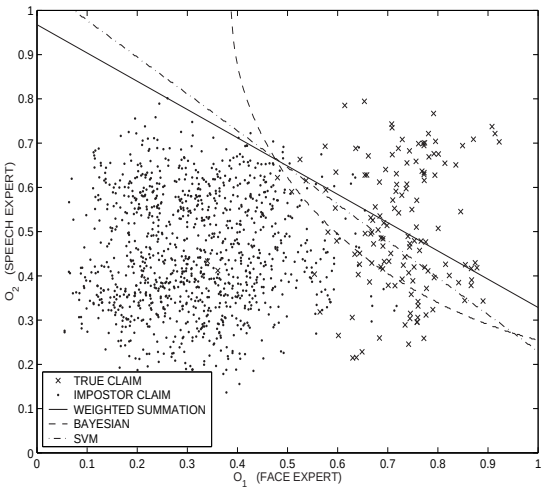

Figure 7: As per Fig. 6, but using noisy speech (corrupted with white noise, SNR = -8 dB).

4.7 Discussion

4.7.1 Effect of Noisy Conditions on Distribution of Opinion Vectors

For convenience, let us refer to the distribution of opinion vectors for true claims and impostor claims as the true claimant and impostor opinion distributions, respectively.

As can be observed in Figs. 6 and 7, the main effect of noisy conditions is the movement of the mean of the true claim opinion distribution towards the o_1 axis. This movement can be explained by analyzing Eqn. (22). 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 $\mathcal{L}(X|\lambda_C)$ 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, as $\lambda_{\overline{C}}$ is a good representation of the general population, it usually covers a wide area of the feature space (see Section 4.2). Thus while the feature vectors may have drifted away from the space described by the true claimant model, they may still be "inside" the space described by the anti-client 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 several impostor claims have 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 mentioned above, $\lambda_{\overline{C}}$ usually covers a wide area of the feature space, thus even though the features have drifted due to mismatched conditions, the may still be "inside" the space described by the anti-client 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. 6 and 7 show the effects of corrupting speech signals with additive white Gaussian noise, we have observed similar effects with the "operations-room" noise.

4.7.2 Effect of Noisy Conditions on Performance

In clean conditions, the weighted summation approach, SVM and Bayesian post-classifiers obtain performance better than either the face or speech expert. However, in high noise levels (SNR = -8 dB), all have performance worse than the face expert; this is expected since in all cases the decision mechanism uses fixed parameters.

All three approaches exhibit similar performance upto a SNR of 8 dB. As the SNR decreases further, the weighted summation approach is significantly more affected than the SVM and Bayesian post-classifiers. The differences in performance in noisy conditions can be attributed to the decision boundaries used by each approach, shown in Figs. 6 and 7; it can be seen that the weighted summation approach has a decision boundary which results in the most mis-classifications of true claimant opinion vectors in noisy conditions.

The performance of the feature concatenation fusion approach is relatively more robust than the three post-classifier approaches. However, for most SNRs the performance is worse than the face expert, suggesting that while in this case feature concatenation fusion is relatively robust to the effects of noise, it is not optimal. The relatively poor performance in clean conditions can be attributed to the VAD; the entire speech signal was classified as containing speech instead of only the speech segments, thus providing a significant amount of irrelevant (non-discriminatory) information when modeling and calculating opinions. Unlike the feature vectors obtained from the speech signal (which could contain either background noise or speech) each facial feature vector contained valid face information; since the speech and facial vectors were concatenated to form one feature vector, the VAD could not distinguish between feature vectors containing background noise and speech. As stated previously, best results were obtained with one-Gaussian client models (compared to 32-Gaussian client models for the speech-only expert), suggesting that when more Gaussians were used, they were used for modeling the non-discriminatory information; moreover, since one-Gaussian models are inherently less precise

than 32-Gaussian models, we would expect them to be more robust to changes in distribution of feature vectors; indeed the results suggest that this is occurring.

5 Performance of Adaptive Approaches in Noisy Audio Conditions

In this section we evaluate the performance of several adaptive opinion fusion methods described in Section 3.2, namely weighted summation fusion with Wark's weight selection and the *mismatch detection* weight adjustment method.

The experimental setup is similar to the one described in Section 4.6. Based on manual observation of plots of speech signals from the VidTIMIT database, N_{noise} was set to 30 for the mismatch detection method [see Eqn. (15)]. One Gaussian for λ_{noise} was sufficient in preliminary experiments. The sigmoid parameters a and b [in Eqn. (16)] were obtained by observing how q in Eqn. (15) 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 Eqn. (16) was close to one for clean utterances and close to zero for utterances with an SNR of -8 dB.

Performance of the adaptive systems is shown in Fig. 8 for white noise, and in Fig. 9 for "operations-room" noise.

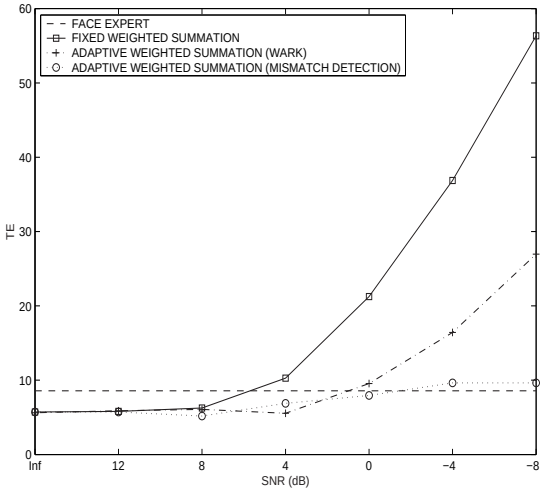

Figure 8: Performance of adaptive fusion techniques in the presence of white noise.

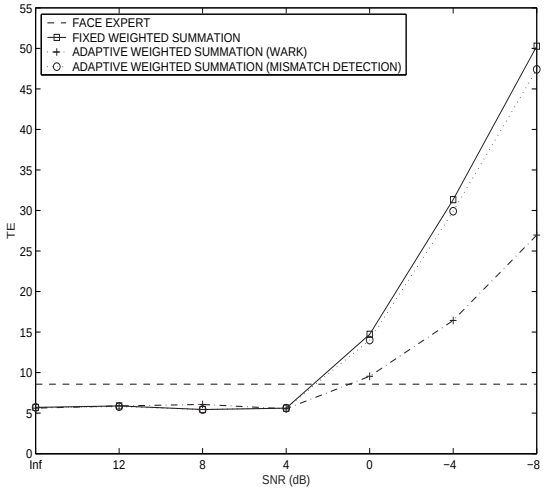

Figure 9: Performance of adaptive fusion techniques in the presence of operations-room noise.

5.1 Discussion

Wark's weight selection 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 [see Eqn. (14)]. However, the impostor distribution changed relatively little due to noisy conditions (as discussed in Section 4.7.1), thus Wark's posterior confidences (κ) for impostor claims changed relatively little as the SNR was lowered. However, Wark's approach appears to be more robust than the fixed weighted summation approach; this is not due to the posterior confidences (κ), but due to the decision boundary being steeper from the start (thus being able to partially take into account the movement of opinion vectors due to noisy conditions); the nature of decision boundary was largely determined by the prior confidences (ζ) found with Eqn. (12).

For the case of white noise, when the mismatch detection weight adjustment method is used in the weighted summation approach, the performance gently deteriorates as the SNR is lowered, becoming slightly worse than the performance of the face expert at an SNR of -4 dB. For the case of "operations-room" noise, the mismatch detection method 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, resulting in performance very similar to the fixed (non-adaptive) approach.

6 Structurally Noise Resistant Post-Classifiers

Partly inspired by the SRM principle used in SVM (see Section 4.5) and by the movement of opinion vectors due to presence of noise (see Section 4.7.1) a structurally noise resistant piece-wise linear (PL) post-classifier is developed (Section 6.1). As the name suggests, the decision boundary used by the post-classifier is designed so that the contribution of errors from the movement of opinion vectors is minimized; this is in comparison to standard post-classifier approaches, where the decision boundary is selected to optimize performance on clean data, with little or no regard to how the distributions of opinions may change due to noisy conditions. The Bayesian classifier presented in Section 3.1 is modified to introduce a similar structural constraint (Section 6.2). The performance of the two proposed post-classifiers is evaluated in Section 6.3.

6.1 Piece-Wise Linear Post-Classifier Definition

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

$$g(\mathbf{o}) = \begin{cases} a(\mathbf{o}) & \text{if } o_2 \geq o_{2,int} \\ \\ b(\mathbf{o}) & \text{otherwise} \end{cases} \tag{37}$$

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

$$a(\mathbf{o}) = m_1 o_1 - o_2 + c_1 \tag{38}$$

$$b(\mathbf{o}) = m_2 o_1 - o_2 + c_2 \tag{39}$$

and $o_{2,int}$ is the threshold for selecting whether to use $a(\mathbf{o})$ or $b(\mathbf{o})$; Figure 10 shows an example of the decision boundary. The verification decision is reached as follows: the claim is accepted when $g(\mathbf{o}) \leq 0$ (i.e. true claimant) and rejected when $g(\mathbf{o}) > 0$ (i.e. impostor).

The first segment of the decision boundary can be described by $a(\mathbf{o}) = 0$, which reduces Eqn. (38) to:

$$o_2 = m_1 o_1 + c_1 \tag{40}$$

If we assume o_2 is a function of o_1, Eqn. (40) is simply the description of a line [66], 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 $\mathbf{o}_{int} = [\ o_{1,int}\ o_{2,int}\ ]^T$; moreover, when the two lines intersect, $a(\mathbf{o}_{int}) = b(\mathbf{o}_{int}) = 0$. Hence

$$o_{2,int} = m_1 o_{1,int} + c_1 = m_2 o_{1,int} + c_2 \tag{41}$$

which leads to:

$$o_{1,int} = \frac{c_1 - c_2}{m_2 - m_1} \tag{42}$$

$$o_{2,int} = m_2 \left(\frac{c_1 - c_2}{m_2 - m_1} \right) + c_2 \tag{43}$$

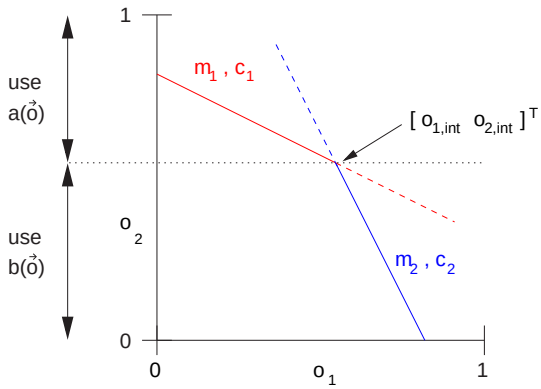

Figure 10: Example decision boundary of the PL classifier

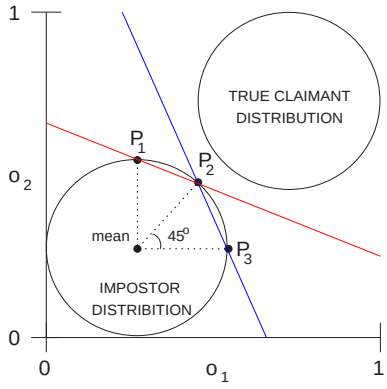

Figure 11: Points used in the initial solution of PL classifier parameters

6.1.1 Structural Constraints and Training

As described in Section 4.7.1, the main effect of noisy conditions is the movement of opinion vectors for true claims toward the o_1 axis. We would like to obtain a decision boundary which minimizes the increase of errors due to this movement. Structurally, this requirement translates to a decision boundary that is as steep as possible; moreover, to keep consistency with the experiments done in Sections 4 and 5, the classifier should be trained for 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 need to be as large as possible (so the decision boundary that is as steep 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 6.1.2, the downhill simplex optimization method [45, 51] can be used to find the final parameters. The following function is minimized:

$$\varepsilon(\lambda_{\text{PL}}) = \epsilon_1(\lambda_{\text{PL}}) + \epsilon_2(\lambda_{\text{PL}}) + \epsilon_3(\lambda_{\text{PL}}) \tag{44}$$

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

$$\epsilon_1(\lambda_{\text{PL}}) = \gamma_1 + \gamma_2 \tag{45}$$

$$\text{where } \gamma_j = \begin{cases} |o_{j,int}| & \text{if } o_{j,int} < 0 \text{ or } o_{j,int} > 1 \\ 0 & \text{otherwise} \end{cases} \tag{46}$$

where $o_{1,int}$ and $o_{2,int}$ are found using Eqns. (42) and (43), respectively,

$$\epsilon_2(\lambda_{\text{PL}}) = \left|\frac{1}{m_1}\right| + \left|\frac{1}{m_2}\right| \tag{47}$$

and finally

$$\epsilon_3(\lambda_{\text{PL}}) = \left|\frac{\text{FA\%}}{100\%} - \frac{\text{FR\%}}{100\%}\right| \tag{48}$$

6.1.2 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 Eqn.(21)], 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. Let us use this property to define three points in 2D opinion space (shown graphically in Fig. 11):

$$P_1 = (\ x_1\ ,\ y_1\ ) \quad = \quad (\ \mu_1\ ,\ \mu_2 + 2\sigma_2\ ) \tag{49}$$

$$P_2 = (\ x_2\ ,\ y_2\ ) \quad = \quad \left(\ \mu_1 + 2\sigma_1 \cos\left[\frac{\pi}{4}\right]\ ,\ \mu_2 + 2\sigma_2 \sin\left[\frac{\pi}{4}\right]\ \right) \tag{50}$$

$$P_3 = (\ x_3\ ,\ y_3\ ) \quad = \quad (\ \mu_1 + 2\sigma_1\ ,\ \mu_2\ ) \tag{51}$$

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

$$m_1 \quad = \quad \frac{y_2 - y_1}{x_2 - x_1} \tag{52}$$

$$c_1 \quad = \quad y_1 - m_1 x_1 \tag{53}$$

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

$$m_2 \quad = \quad \frac{y_3 - y_2}{x_3 - x_2} \tag{54}$$

$$c_2 \quad = \quad y_2 - m_2 x_2 \tag{55}$$

The initial solution for real data is shown in Fig. 14.

6.2 Modified Bayesian Post-Classifier

In Fig. 6 it can be seen that the decision boundary made by the Bayesian post-classifier (described in Section 3.1) envelops the true claimant opinion distribution. The downward movement of the vectors due to noisy conditions (discussed in Section 4.7.1) crosses the boundary and is the main cause of the error increases. If the decision boundary was forced to envelop the distribution of opinion vectors for impostor claims, the error increase would be reduced; this can be accomplished by modifying the decision rule described in (10) to use only the impostor likelihood (i.e., $\log p(o_i|\lambda_{i,\text{true}}) = 0\ \ \forall\ i$):

$$\text{chosen class} = \begin{cases} C_1 & \text{if} \quad -\sum_{i=1}^{N_E} \log p(o_i|\lambda_{i,\text{imp}}) > t \\ \\ C_2 & \text{otherwise} \end{cases} \tag{56}$$

where C_1 and C_2 are the true claimant and impostor classes, respectively.

Compared to the piece-wise linear classifier presented in Section 6.1, the modified Bayesian classifier avoids heuristics and is easily extendable to three or more experts.

6.3 Experiments and Discussion

The performance of the proposed PL and modified Bayesian post-classifiers is evaluated; the experimental setup is the same as described in Section 4.6, with the results for white noise shown in Fig. 12 and for "operations-room" noise in Fig. 13. Figures 14 and 15 show the distribution of opinion vectors in clean and noisy (SNR = -8 dB) conditions (white noise), respectively, with the decision boundaries used by the proposed approaches.

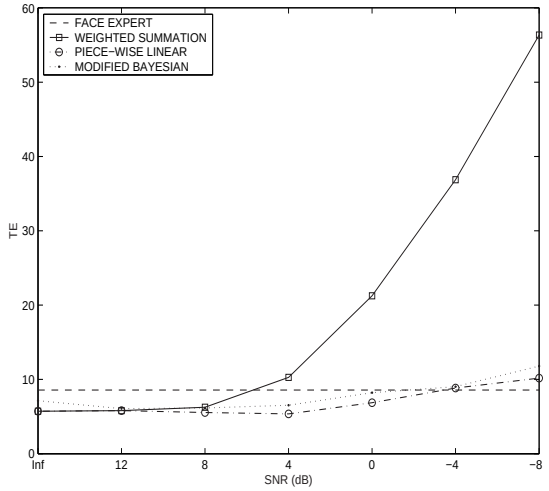

Figure 12: Performance of structurally noise resistant fusion techniques in the presence of white noise.

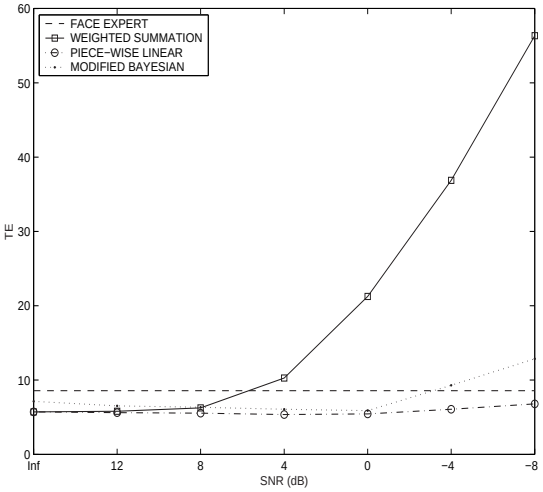

Figure 13: Performance of structurally noise resistant fusion techniques in the presence of operations-room noise.

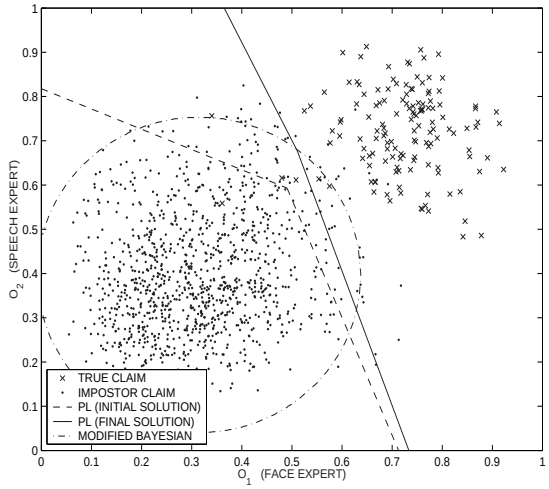

Figure 14: Decision boundaries used by structurally noise resistant fusion approaches and the distribution of opinion vectors for true and impostor claims (clean speech).

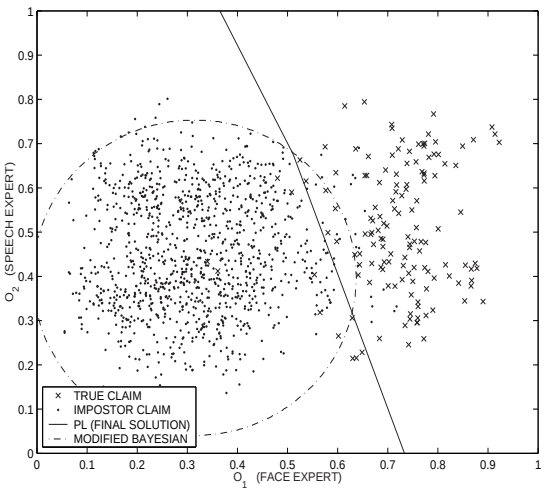

Figure 15: As per Fig. 14, but using noisy speech (corrupted with white noise, SNR = -8 dB).

As can be observed, the decision boundary used by the PL post-classifier effectively takes into account the movement of opinion vectors due to noisy conditions. Comparing Figs. 8 and 12 it can be seen that the proposed PL post-classifier has similar performance to the adaptive weighted summation approach, with the advantage of having a fixed (non-adaptive) structure; moreover, unlike the mismatch detection weight update algorithm used in the adaptive approach, the PL post-classifier does not make a direct assumption about the type of noise that caused the mismatch between training and testing conditions.

Due to the nature of the decision boundary, the performance of the modified Bayesian post-classifier is slightly worse than the PL post-classifier; however, unlike the PL post-classifier proposed here, the modified Bayesian post-classifier is easily extendable to three or more experts.

7 Conclusions and Future Work

This paper first provided a review of important concepts in the field of information fusion, followed by a review of important milestones in audio-visual person identification and verification. Several recent adaptive and non-adaptive techniques for reaching the verification decision (i.e, whether to accept or reject the claimant), based on speech and face information, were evaluated in clean and noisy audio conditions on a common database; it was shown that in clean conditions most of the non-adaptive approaches provide similar performance and in noisy conditions most exhibit deterioration in performance; moreover, it was shown that current adaptive approaches are either inadequate or utilize restrictive assumptions. A new category of classifiers was then introduced, where the decision boundary is fixed but constructed to take into account how the distributions of opinions are likely to change due to noisy conditions; compared to a previously proposed adaptive approach, the proposed classifiers do not make a direct assumption about the type of noise that causes the mismatch between training and testing conditions.

Future work will include a modification of the feature vector concatenation approach, so that only audio vectors classified as speech (by the Voice Activity Detector) are concatenated with corresponding face vectors; this should aid in significantly reducing the amount of irrelevant (non-discriminative) information that is currently being used during modeling and likelihood calculation (leading to the relatively poor performance of feature vector concatenation approach in clean conditions).

8 Acknowledgments

The initial version of this work was written while the first author was a student at Griffith University; revision was performed at IDIAP, with thanks to support by the Swiss National Science Foundation through the National Center of Competence in Research (NCCR) on Interactive Multimodal Information Management (IM2). The authors also thank the anonymous reviewers, Samy Bengio and Alexei Pozdnoukhov for helpful suggestions.

References

[1] Y. Abdeljaoued, Fusion of person authentication probabilities by Bayesian statistics, in: Proc. 2nd Int. Conf. Audio- and Video-based Biometric Person Authentication, Washington D.C., 1999, pp. 172-175.

[2] W. Atkins, A testing time for face recognition technology, Biometric Technology Today 9 (3) (2001) 8-11.

[3] B. Achermann, H. Bunke, Combination of classifiers on the decision level for face recognition, Technical Report IAM-96-002, Institut für Informatik und angewandte Mathematik, Universität Bern, 1996.

[4] A. Adjoudani, C. Benoît, Audio-visual speech recognition compared across two architectures, in: Proc. 4th European Conf. Speech Communication and Technology, Madrid, Spain, 1995, Vol. 2, pp. 1563-1567.

[5] L. A. Alexandre, A. C. Campilho, M. Kamel, On combining classifiers using sum and product rules, Pattern Recognition Letters 22 (2001) 1283-1289.

[6] H. Altiçay, M. Demirekler, An information theoretic framework for weight estimation in the combination of probabilistic classifiers for speaker identification, Speech Communication 30 (2000) 255-272.

[7] H. Altiçay, M. Demirekler, Comparison of different objective functions for optimal linear combination of classifiers for speaker identification, in: Proc. IEEE International Conf. Acoustics, Speech and Signal Processing, Salt Lake City, 2001, Vol. 1, pp. 401-404.

[8] Y. Barniv, D. Casasent, Multisensor image registration: Experimental verification, in: Proceedings of the SPIE 292 (1981) 160-171.

[9] S. Bengio, Multimodal authentication using asynchronous HMMs, in: Proc. 4th International Conf. Audio- and Video-based Biometric Person Authentication (AVBPA), Guildford, 2003, pp. 770-777.

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

[11] R.M. Bolle, J.H. Connell, N.K. Ratha, Biometric perils and patches, Pattern Recognition 35 (2002) 2727-2738.

[12] R. Brunelli, D. Falavigna, T. Poggio, L. Stringa, Automatic person recognition using acoustic and geometric features, Machine Vision & Applications 8 (1995) 317-325.

[13] R. Brunelli, D. Falavigna, Person identification using multiple cues, IEEE Trans. Pattern Analysis and Machine Intelligence 10 (1995) 955-965.

[14] C.J.C. Burges, A tutorial on support vector machines for pattern recognition, Data Mining and Knowledge Discovery 2 (1998) 121-167.

[15] F. Cardinaux, C. Sanderson, S. Bengio, Face verification using adapted generative models, in: 6th Int. Conf. Automatic Face and Gesture Recognition (AFGR), Seoul, 2004.

[16] E. Caulcott, Significance Tests, Routledge & Kegan Paul, London, 1973.

[17] 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.

[18] C.C. Chibelushi, F. Deravi, J.S. Mason, Voice and facial image integration for speaker recognition, in: IEEE International Symposium and Multimedia Technologies and Future Applications, Southampton, UK, 1993.

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

[20] U. Dieckmann, P. Plankensteiner, T. Wagner, SESAM: A biometric person identification system using sensor fusion, Pattern Recognition Letters 18 (1997) 827-833.

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

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

[23] J.-L. Dugelay, J.-C. Junqua, C. Kotropoulos, R. Kuhn, F. Perronnin, I. Pitas, Recent advances in biometric person authentication, In: Proc. International Conf. Acoustics, Speech and Signal Processing, Orlando, 2002, Vol. IV, pp. 4060-4063.

[24] S. Furui, Cepstral analysis technique for automatic speaker verification, IEEE Trans. Acoustics, Speech and Signal Processing 29 (2) (1981), 254-272.

[25] S. Furui, Recent advances in speaker recognition, Pattern Recognition Letters 18 (1997) 859-872.

[26] J.-L. Gauvain, C.-H. Lee, Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains, IEEE Trans. Speech and Audio Processing 2 (1994) 291-298.

[27] D. Genoud, F. Bimbot, G. Gravier, G. Chollet, Combining methods to improve speaker verification, in: Proc. 4th International Conf. Spoken Language Processing, Philadelphia, 1996, Vol. 3, pp. 1756-1759.

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

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

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

[31] D.L. Hall, J. Llinas, Multisensor data fusion, in: D. L. Hall and J. Llinas (Eds.), Handbook of Multisensor Data Fusion, CRC Press, USA, 2001, pp. 1-1 - 1-10.

[32] T.K. Ho, J.J. Hull, S.N. Srihari, Decision combination in multiple classifier systems, IEEE Trans. Pattern Analysis and Machine Intelligence 16 (1994) 66-75.

[33] L. Hong, A. Jain, Integrating Faces and Fingerprints for Personal Identification, IEEE Trans. Pattern Analysis and Machine Intelligence 20 (1998) 1295-1306.

[34] S.S. Iyengar, L. Prasad, H. Min, Advances in Distributed Sensor Technology, Prentice Hall PTR, New Jersey, 1995.

[35] A. Jain, U. Uludag, Hiding biometric data, IEEE Trans. Pattern Analysis and Machine Intelligence 25 (2003) 1494-1498.

[36] C. Jankowski, A. Kalyanswamy, S. Basson, J. Spitz, NTIMIT: a phonetically balanced, continuous speech telephone bandwidth speech database, in: Proc. International Conf. Acoustics, Speech and Signal Processing, Albuquerque, 1990, Vol. 1, pp. 109-112.

[37] T. Joachims, Making large-scale SVM learning practical, in: B. Schölkopf, C. Burges and A. Smola (Eds.), Advances in Kernel Methods - Support Vector Learning, MIT-Press, 1999.

[38] P. Jourlin, J. Luettin, D. Genoud, H. Wassner, Acoustic-labial speaker verification, Pattern Recognition Letters 18 (1997) 853-858.

[39] P. Jourlin, J. Luettin, D. Genoud, H. Wassner, Integrating acoustic and labial information for speaker identification and verification, in: Proc. 5th European Conf. Speech Communication and Technology, Rhodes, Greece, 1997, Vol. 3, pp. 1603-1606.

[40] J. Kittler, J. Matas, K. Johnsson, M. U. Ramos-Sánchez, Combining evidence in personal identity verification systems, Pattern Recognition Letters 18 (1997) 845-852.

[41] J. Kittler, M. Hatef, R.P.W. Duin, J. Matas, On combining classifiers, IEEE Trans. Pattern Analysis and Machine Intelligence 20 (1998) 226-239.

[42] J. Luettin, Visual Speech and Speaker Recognition, PhD Thesis, Department of Computer Science, University of Sheffield, 1997.

[43] R.C. Luo, M.G. Kay, Introduction, in: R.C. Luo and M.G. Kay (Eds.), Multisensor Integration and Fusion for Intelligent Machines and Systems, Ablex Publishing Corporation, Norwood, NJ, 1995, pp. 1-26.

[44] A.V. Nefian, L.H. Liang, T. Fu, X.X. Liu, A Bayesian approach to audio-visual speaker identification, in: Proc. 4th International Conf. Audio- and Video-based Biometric Person Authentication (AVBPA), Guildford, 2003, pp. 761-769.

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

[46] J. Ortega-Garcia, J. Bigun, D. Reynolds, J. Gonzales-Rodriguez, Authentication Gets Personal with Biometrics, IEEE Signal Processing Magazine, 21 (2004) 50-62.

[47] L.F. Pau, Fusion of multisensor data in pattern recognition, in: J. Kittler, K.S. Fu and L.F. Pau (Eds.), Pattern Recognition Theory and Applications (Proceedings of NATO Advanced Study Institute), D Reidel Publ., Dordrecht, Holland, 1982.

[48] J. Picone, Signal modeling techniques in speech recognition, Proceedings of the IEEE 81 (1993) 1215-1247.

[49] N. Poh, S. Bengio, Non-linear variance reduction techniques in biometric authentication, in: Proc. Workshop on Multimodal User Authentication, Santa Barbara, 2003, pp. 123-130.

[50] G. Potamianos, H. Graf, Discriminative training of HMM stream exponents for audio-visual speech recognition, in: Proc. International Conf. Acoustics, Speech and Signal Processing, Seattle, 1998, pp. 3733-3736.

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

[52] L. Rabiner, B.-H. Juang, Fundamentals of Speech Recognition, Prentice Hall, Englewood Cliffs, 1993.

[53] V. Radová, J. Psutka, An approach to speaker identification using multiple classifiers, in: Proc. IEEE Conf. Acoustics, Speech and Signal Processing, Munich, 1997, Vol. 2, pp. 1135-1138.

[54] D. Reynolds, A gaussian mixture modeling approach to text-independent speaker identification, Technical Report 967, Lincoln Laboratory, Massachusetts Institute of Technology, 1993.

[55] D. Reynolds, Experimental evaluation of features for robust speaker identification, IEEE Trans. Speech and Audio Processing 2 (1994) 639-643.

[56] D. Reynolds, T. Quatieri, R. Dunn, Speaker verification using adapted gaussian mixture models, Digital Signal Processing 1-3 (2000) 19-41.

[57] A. Ross, A. Jain, Information fusion in biometrics, Pattern Recognition Letters 24 (2003) 2115-2125.

[58] V. Roth, V. Steinhage, Nonlinear discriminant analysis using kernel functions, Technical Report Nr IAI-TR-99-7 (ISSN 0944-8535), University of Bonn, 1999.

[59] C. Sanderson, K.K. Paliwal, The VidTIMIT database, IDIAP Communication 02-06, Martigny, Switzerland, 2002.

[60] C. Sanderson, K.K. Paliwal, Information fusion and person verification using speech & face information, IDIAP Research Report 02-33, Martigny, Switzerland, 2002.

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

[62] C. Sanderson, K.K. Paliwal, Fast features for face authentication under illumination direction changes, Pattern Recognition Letters 24 (14) (2003) 2409-2419.

[63] C. Sanderson, S. Bengio, Statistical transformation techniques for face verification using faces rotated in depth, IDIAP Research Report 04-04, Martigny, Switzerland, 2004.

[64] P. Silsbee, A. Bovik, Computer lipreading for improved accuracy in automatic speech recognition, IEEE Trans. Speech and Audio Processing 4 (1996) 337-351.

[65] F.K. Soong, A.E. Rosenberg, On the use of instantaneous and transitional spectral information in speaker recognition, IEEE Trans. Acoustics, Speech and Signal Processing 36 (1988) 871-879.

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

[67] R.R. Tenney, N.R. Sandell Jr., Detection with distributed sensors, IEEE Trans. Aerospace and Electronic Systems 17 (1981) 98-101.

[68] R.R. Tenney, N.R. Sandell Jr., Strategies for distributed decision making, IEEE Trans. on Systems, Man and Cybernetics 11 (1981) 527-537.

[69] T. Thong, Y.C. Jenq, Hardware and architecture, in: S.K. Mitra and J.F. Kaiser (Eds.), Handbook for Digital Signal Processing, John Wiley & Sons, USA, 1993, pp. 721-781.

[70] M. Turk, A. Pentland, Eigenfaces for recognition, Journal of Cognitive Neuroscience 3 (1991) 71-86.

[71] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, New York, 1995.

[72] A. Varga, H.J.M. Steeneken, M. Tomlinson, D. Jones, 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, 1992.

[73] P.K. Varshney, Distributed Detection and Data Fusion, Springer-Verlag, New York, 1997.

[74] P. Verlinde, A contribution to multi-modal identity verification using decision fusion, PhD Thesis, Department of Signal and Image Processing, Telecom Paris, France, 1999.

[75] T. Wark, S. Sridharan, V. Chandran, Robust speaker verification via fusion of speech and lip modalities, in: Proc. International Conf. Acoustics, Speech and Signal Processing, Phoenix, 1999, Vol. 6, pp. 3061-3064.

[76] T. Wark, S. Sridharan, V. Chandran, Robust speaker verification via asynchronous fusion of speech and lip information, in: Proc. 2nd International Conf. Audio- and Video-based Biometric Person Authentication, Washington, D.C., 1999, pp. 37-42.

[77] T. Wark, S. Sridharan, V. Chandran, The use of temporal speech and lip information for multi-modal speaker identification via multi-stream HMM's [sic], in: Proc. International Conf. Acoustics, Speech and Signal Processing, Istanbul, 2000, pp. 2389-2392.

[78] T. Wark, Multi-modal speech processing for automatic speaker recognition, PhD Thesis, School of Electrical & Electronic Systems Engineering, Queensland University of Technology, Brisbane, 2000.

[79] J.L. Wayman, Digital Signal Processing in Biometric Identification: a Review, In: Proc. IEEE Int. Conf. Image Processing, Rochester, 2002, Vol. 1, pp. 37-40.

[80] B. Wildermoth, K.K. Paliwal, Use of voicing and pitch information for speaker recognition, in: Proc. 8th Australian International Conf. Speech Science and Technology, Canberra, 2000, pp. 324-328.

[81] J.D. Woodward, Biometrics: privacy's foe or privacy's friend?, Proceedings of the IEEE 85 (1997) 1480-1492.