Optimal Weighting of Bimodal Biometric Information with Specific Application to Audio-Visual Person Identification

Roland Hu, R.I. Damper

PII: S1566-2535(08)00042-0
DOI: 10.1016/j.inffus.2008.08.003
Reference: INFFUS 316

To appear in: *Information Fusion*

Received Date: 4 August 2006
Revised Date: 9 July 2007
Accepted Date: 7 August 2008

Please cite this article as: R. Hu, R.I. Damper, Optimal Weighting of Bimodal Biometric Information with Specific Application to Audio-Visual Person Identification, *Information Fusion* (2008), doi: 10.1016/j.inffus.2008.08.003

Optimal Weighting of Bimodal Biometric Information with Specific Application to Audio-Visual Person Identification

Roland Hu and R. I. Damper

Information: Signals, Images, Systems (ISIS) Research Group,
School of Electronics and Computer Science,
University of Southampton,
Southampton SO17 1BJ, UK.
Emails: {hh03r|rid}@ecs.soton.ac.uk

Abstract

A new method is proposed to estimate the optimal weighting parameter for combining audio (speech) and visual (face) information in person identification, based on estimating probability density functions (pdf's) for classifier scores under Gaussian assumptions. Performance comparisons with real and simulated data indicate that this method has advantages in reducing bias and variance of the estimation relative to other methods tried, so achieving a robust estimator of the optimal weighting parameter. Another contribution is that we propose the bootstrap method to compare performances of different algorithms for estimating the optimal weighting parameter, so providing a strict criterion in comparing algorithms of this kind. Using simulated data, for which the pdf is controlled and known, we show that the advantages of the method hold up when the underlying Gaussian assumption is violated. The main drawback is that we have to choose an adjustable parameter, and it is not clear how this should best be done.

Key words: Face recognition, speaker recognition, person identification, weighted sum rule, bootstrapping

1 Introduction

There is an increasing interest in biometric person identification for commercial, security, surveillance and other applications, but identification based on only one modality is unlikely to achieve acceptable performance for practical deployment. A potential way to overcome this is to combine information from more than one modality, and several important studies have confirmed this potential (Xu *et al.*,

1992; Kittler *et al.*, 1998; Toh and Yau, 2004). In this paper, we consider the particular situation of audio-visual person identification where there are just two sources of information: an audio signal (speech) and a video signal (face).

It is widely agreed that the audio and visual modalities can be combined at three different levels, is defined by Lucey *et al.* (2005) as early integration, middle integration and late integration, respectively. For early integration, the feature vectors of audio and visual signals are extracted separately, then vector concatenation is employed to form a new feature vector, finally, this new feature vector is used for recognition (Adjoudani and Benoît, 1995; Luettin, 1997). For late integration, the audio and visual classifiers are built separately, then fusion methods are implemented to combine the scores generated by the audio and visual classifiers (Kittler *et al.*, 1998; Ben-Yacoub *et al.*, 1999; Toh and Yau, 2004). Another level of integration, namely middle integration, is also frequently used for combining audio and visual modalities. Examples of this level of integration are multistream hidden Markov models (Bengio, 2003; Fu *et al.*, 2003; Lucey *et al.*, 2005).

However, the most frequently-used methods appearing in the literature are based on late integration. This is because of two reasons. First, compared with early integration and middle integration, late integration is simple. It does not take into account the correlation and interaction of audio and visual signals, thus circumventing the problem of synchronizing audio and visual signals. Instead, it treats these two modalities separately, obtaining two separate classifiers, then processing scores generated by these two classifiers. Second, late integration achieves commensurate, if not better, recognition rates compared with early integration and middle integration. Lucey *et al.* compared different approaches of the early integration, middle integration and late integration, and found that late integration is superior in terms of classifier flexibility and its ability to dampen independent errors coming from either modality.

Of all the approaches based on late integration, the simplest use some fixed fusion rule, e.g., the sum rule, product rule etc. (Kittler *et al.*, 1998; Duin, 2002). The scores generated by the audio and visual classifiers are combined by some fixed functions, and training the combined classifier is not needed. It has been shown that by using fixed rules, the performance of the person recognition system can be greatly improved (Kittler *et al.*, 1998; Erzin *et al.*, 2005). Kittler *et al.* (1998) attempted to build a theoretical framework for such fixed rules. Their experimental results for combining the scores from three experts (two face experts and a text-dependent speaker expert) showed that the sum rule outperformed the product rule. A small revision to the fixed rules is to assign weighting parameter(s) to each modality based on the performance of that modality: the so-called weighted sum rule and weighted product rule. Various studies have shown that weighted sum and product rules can perform better than fixed sum and product rules (Chibelushi *et al.*, 1993; Brunelli and Falavigna, 1995; Maison *et al.*, 1999; Wark, 2000; Sanderson and Paliwal, 2003). Although various methods are proposed to choose weighting

parameters, there is not unanimous agreement on how to do this. Another approach based on late integration is to regard the fusion problem as a pattern recognition problem. The scores generated by the audio and visual classifiers can be regarded as features. The combined classifier needs to be built to fit these features (or scores) into their correct classes. Several models which are frequently used in pattern recognition can be used to build the combined classifier. Ben-Yacoub *et al.* (1999) investigated support vector machines (SVMs), Bayesian classifiers, Fisher linear discriminants, C4.5 classifiers and multilayer perceptrons for audio-visual classifier combination, and found that SVM and Bayesian classifiers perform slightly better than the others.

In the current work, we propose a means to choose the weighting parameter for audio-visual person identification which is based on estimating the probability density functions (pdf's) for the classifier scores. We have argued elsewhere (Hu and Damper, 2006) that studying pdf's should be the first step in finding a good fusion algorithm. The proposed approach is compared with three other well-established techniques. Using the bootstrap method, we conclude that our approach can both reduce the bias and variance, thus achieving a better estimation for the optimal weighting parameter. Although the method is described and studied for person identification, it has the potential to be generalized to the verification case, as discussed in Section 8.

Our audio and visual classification techniques are deliberately very classical, in order to focus on the issue of combining classifiers. The speaker identifier is based on the Gaussian mixture model (Reynolds and Rose, 1995), and the face identifier is based on dynamic link architecture (Lades *et al.*, 1993). In this work, the two sources are not synchronized. That is, the video information is a static image of the speaker rather than an image sequence depicting lip movements etc. during speech production. This is done to simplify the problem at this stage and to allow us to concentrate on the issue of optimal weighting of the two sources of information.

The remainder of this paper is organized as follows. The construction of the speaker and face classifiers is briefly discussed in Sections 2 and 3, respectively. The proposed method is described in Section 4. The performance of the proposed method is then compared to that of three competitor methods using real speech and video data in Section 5. In Section 6, the statistical technique of bootstrapping is used to improve the estimates obtained. However, proper interpretation of the comparison is still uncertain because the actual value of the optimal weighting parameter is unknown. Hence, in Section 7 we repeat the comparison with simulated data for which the optimal weight is precisely known. Based on these comparisons, Section 8 concludes that the proposed method provides a robust estimator of the optimal weighting parameter for combining classifiers, and outlines our future work to generalize the method to person verification.

2 Speaker Identification

We use cepstral coefficients derived from a mel-frequency filterbank to represent the features for speaker identification. Speaker modelling is based on the Gaussian mixture model (GMM) introduced for speaker recognition by Reynolds and Rose (1995). A Gaussian mixture density is a weighted sum of M component densities:

$$p(\vec{x}|\lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x}) \tag{1}$$

where $\vec{x}$ is a D-dimensional random vector, $b_i(\vec{x})$ is the component density of the ith mixture and p_i is the weight of the ith mixture. Each component density is a D-variate Gaussian function of the form:

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(\vec{x} - \vec{\mu}_i)^T \Sigma_i^{-1}(\vec{x} - \vec{\mu}_i)\right\}$$

with mean vector $\vec{\mu}_i$ and covariance matrix Σ_i. The mixture weights satisfy the constraint $\sum_{i=1}^{M} p_i = 1$. The complete Gaussian mixture density is parameterized by the mean vectors, covariance matrices and mixture weights from all component densities. These parameters are collectively represented as the 3-tuple:

$$\lambda = \{p_i, \vec{\mu}_i, \Sigma_i\} \quad i = 1, 2, \ldots, M$$

GMMs for each speaker k are trained (i.e., the parameters of λ_k are estimated) using the EM (expectation-maximisation) algorithm (Dempster *et al.*, 1977).

Suppose there are K speakers to be identified. Then λ_k, $k = 1, 2, \ldots, K$, is the model corresponding to the kth enrolled speaker. The goal of speaker identification is to find the one among these K models that best matches the test data represented by a sequence of F frames, $X = \{\vec{x_1}, \vec{x_2}, \ldots, \vec{x_F}\}$. In making the decision, we use the following frame-base weighted likelihood distance measure between the test data and the kth speaker model:

$$d_k = \frac{1}{F} \sum_{f=1}^{F} \log p(\vec{x_f}|\lambda_k)$$

in which $p(\vec{x_t}|\lambda_k)$ is given by (1). The normalisation by F is necessary as each token will, in general, have a different length and, therefore, a different number of frames.

The task of a classifier is to assign an input sequence X to one of K classes: $\omega_1, \omega_2, \ldots, \omega_K$. In this paper, X represents the input of both audio and video modalities. We can then identify speaker s according to the rule:

$$\text{decide } X \in \omega_s \text{ if } s = \arg\max_i d_i$$

The GMM algorithm is applied after intra- and inter-word silence is automatically removed. A simple silence-removal technique based on combining information of sound intensity and zero crossing rate is used. This is a version of the algorithm due to Rabiner and Sambur (1975), originally designed for detection of the endpoints of isolated words but modified here for the removal of word-internal silence. This algorithm sets two sound intensity thresholds: an upper threshold I_1 and a lower threshold I_2 ($I_1 > I_2$). It also sets a zero crossing rate threshold, Z_1. First, the algorithm marks the data whose intensity is higher than the upper threshold I_1 as 'speech points'. Then it extends the boundary of the speech points to points which have higher intensity than the lower intensity threshold I_2. After this, the algorithm further extends the boundary of the speech points to those whose zero crossing rates exceed Z_1. All the other data, not marked as speech points, are removed as 'silence'. By appropriately setting these three thresholds, the algorithm can successfully remove silence in most cases. This is found to improve performance relative to retaining periods of silence (Hu and Damper, 2005).

3 Face Identification

The face identification system is based on the dynamic link architecture (DLA) of Lades *et al.* (1993). The input face image is represented as a set of nodes, each of which contains 40 Gabor wavelet coefficients. Following the authors of the original paper, we call a node and the 40 Gabor wavelet coefficients affiliated with it a 'jet', and we call all the jets in one image a 'graph'. The graph on a training image is defined as a 'model graph'; and the graph on a testing image is defined as an 'image graph'. During storage, all the model graphs are formed and are labelled with jets from a subgrid centred over the training images to be stored. During identification, matching takes place by the adaptive formation of an image graph to match best a given model graph. The matching process is based throughout on one-to-one links between jets in the model graph and the image graph. The process of image-graph formation is controlled by a cost function which favours similarity of corresponding jets and which penalizes metric deformation. The quality of different matches between a model graph and an image graph can be evaluated using the

following score function:

$$S(M, I) = \sum_{i=1}^{N_n} S_1(M_{n_i}, I_{n_i}) - \lambda \sum_{j=1}^{N_e} S_2(M_{e_j}, I_{e_j}) \tag{2}$$

where M_{n_i} represents the ith jet of the model graph; and I_{e_j} represents the jth edge of the image graph; N_n, N_e are the number of jets and edges, respectively.

The first term of the score function, $S_1(M_{n_i}, I_{n_i})$, measures the similarity of jets in the model graph and the image graph. Suppose the 40 Gabor wavelet coefficients affiliated with the ith jet of the model graph are represented by the vector $J(M_{n_i})$; and correspondingly we define $J(I_{n_i})$ as the wavelet coefficients affiliated with the ith jet of the image graph. Then,

$$S_1(M_{n_i}, I_{n_i}) = \frac{J(M_{n_i}) \cdot J(I_{n_i})}{\|J(M_{n_i})\| \|J(I_{n_i})\|}$$

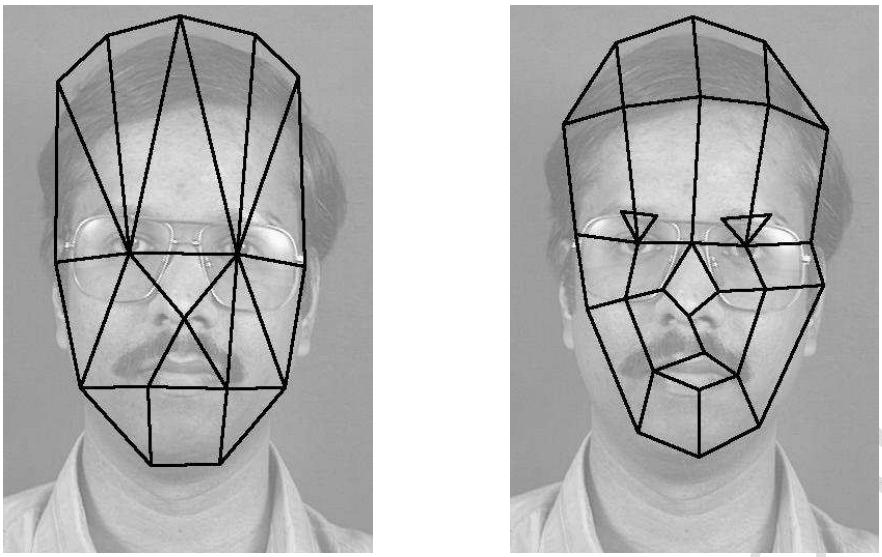
The second term, $S_2(M_{e_j}, N_{e_j})$, calculates the metric deformation between the model graph and the image graph. Suppose the jth edge connects the pth node and the qth node. We use $\vec{M}_p$ and $\vec{M}_q$ to represent the position vectors of the pth and qth nodes in the model graph; and $\vec{I}_p$ and $\vec{I}_q$ to represent the position vectors of the pth and qth nodes in the image graph. Then the second term can be written as:

$$S_2(M_{e_j}, N_{e_j}) = \|(\vec{M}_p - \vec{M}_q) - (\vec{I}_p - \vec{I}_q)\|$$

Because an ideal match of the model graph and the image graph should have large similarity values (S_1) and small distortion values (S_2), equation (2) will have a higher value for a good match and a lower value for a poor match. Suppose there are K classes for a face identification system, which corresponds to K model graphs $M_1, M_2, \ldots, M_K$ respectively. The task of a face identification system is to assign an input image graph I to one of these K classes according to the following rule:

$$\text{decide } X \in \omega_s \text{ if } s = \arg\max_i S(M_i, I)$$

Face recognition can be divided into two steps (Wiskott *et al.*, 1997). First, the face region of an image should be automatically determined. Second, the detected face region should be sent to an identification system and the identification result obtained. We call the first step 'face detection' and the second 'face identification'. In our work, we use a coarse graph which consists of 16 nodes for face detection. This coarse graph finds the best-fitting position (the position which maximizes the similarity score S_1) by scanning around the whole image. Then a more



(a) (b)

Fig. 1. Typical example of use of DLA for face detection and identification: (a) A coarse graph for face detection; (b) A complex graph for face identification.

complex graph is placed on the detected best-fitting position. Its nodes are then adjusted to improve the similarity score further, thus finding accurate positions of facial features. One can see from Figure 1 that, in general, the matching finds face features quite accurately. But mismatches occur: for example, the nodes in (b) are not exactly positioned on the two corners of the mouth. After having extracted the nodes on the testing image, identification is possible with relatively little computational effort by comparing an image graph to all model graphs by equation (2) and picking the one with the highest score. Refer to Lades *et al.* for more details of the DLA method.

4 Proposed Method

After obtaining identification scores of both the audio and video classifiers, the next step is to combine these with a view to obtaining better identification results. Some well-known simple fixed rules for combining the set of base classifiers, such as product rule, sum rule, maximum rule, minimum rule and median rule, are described by Kittler *et al.* (1998). However, fixed rules can be sub-optimal (Duin, 2002) and there exist rules which need a training set to adjust parameters so as to obtain better identification. One popular example is the weighted sum rule, which we use here.

Of course, an alternative to training (i.e., finding the weights empirically) is to try to determine the weights from theory, based on some assumptions. The obvious problem which arises is that, frequently, no analytically-soluble formulation can be

found, even with dramatically-strong simplifying assumptions. In this work, we present a method for estimating the (single) optimal weight for combining our two classifiers under Gaussian assumptions and compare it with results obtained using the actual audio-visual identification system, as well a selection of competitor systems described in the literature. We refer to the former method as 'proposed' and to the latter as 'empirical'. The reader should note that a form of 'training' is still required for the proposed method, as we have to estimate the moments of some Gaussian distributions from training data.

4.1 Theoretical Development

Suppose each of the audio and video classifiers consists of K discriminant functions, $f^1(X)$, $f^2(X)$,..., $f^K(X)$. The decision rule in terms of discriminant functions is:

$$\text{decide } X \in \omega_s \text{ if } s = \arg\max_i f^i(X) \tag{3}$$

Here, X represents the input of both audio and video modalities. We denote by $f_1^1(X)$, $f_1^2(X)$, ..., $f_1^K(X)$ the scores obtained from the video classifier (face identification), and by $f_2^1(X)$, $f_2^2(X)$, ..., $f_2^K(X)$ the scores obtained from the audio classifier (speaker identification). Our aim is to find a weighting parameter $\alpha \in [0, 1]$ to combine optimally these two sets of scores using the weighted sum rule. This gives a new set of score functions:

$$f_{\text{comb}}^k(X, \alpha) = \alpha f_1^k(X) + (1 - \alpha) f_2^k(X) \qquad k = 1, 2, \ldots, K \tag{4}$$

The notation $f_{\text{comb}}^k(X, \alpha)$ indicates that the combined scores depend not only on the input data X but also on the weighting parameter α. In what follows, however, we simplify the notation for discriminant functions by dropping arguments X and α, *except* when it is necessary to distinguish among different values of these arguments.

The weighting parameter α in equation (4) should be selected according to the relative reliability of the two classifiers. The most direct way to do this is to optimize α so as to maximize the identification rate on some training data (Maison *et al.*, 1999; Duin, 2002), but this carries the danger of over-fitting, so reducing the ability to generalize to unseen test data. Several methods can be used to prevent over-fitting. For example, Ney (1995) used the smoothed error rate as the cost function for optimizing the parameter α, and Brunelli and Falavigna (1995) used the normalized ratio of the first- to the second-best integrated score to calculate α. A common property of these two methods is that both use the information of the probability densities of the scores obtained by the two classifiers. In this paper, we

propose a method for choosing the weighting parameter α that directly maximizes the correct identification rate, i.e., the probability of correct identification by the combined system, from score distributions.

The first step is to normalize the scores of the training data. We use the so-called z-score normalisation technique, which is calculated using the arithmetic mean and standard deviation of the given data. Refer to Jain *et al.* (2005) for an overview of score normalisation techniques in multimodal biometric systems.

The z-score normalisation process can be divided into two steps. In the first step, all scores of both audio and video classifiers have their mean subtracted and the result is then divided by their variance:

$$\overline{f_m^k(X_i)} = \frac{f_m^k(X_i) - \mu_m}{\sigma_m} \text{ : with } \mu_m = \frac{\sum_{i=1}^{I}\sum_{k=1}^{K} f_m^k(X_i)}{I \times K} \text{ and } \sigma_m = \frac{\sum_{i=1}^{I}\sum_{k=1}^{K}\left(f_m^k(X_i) - \mu_m\right)^2}{I \times K} \tag{5}$$

Here, I is the number of training data, K is the number of classes, and $m \in \{1, 2\}$.

The second step of normalisation is to make the correct score (i.e., that for the correct person) zero. This gives us a known reference point from which to assess scores, and simplifies the derivation of an appropriate mathematical model under Gaussian assumptions—see below. If we set the weighting parameter α to a constant value, we can obtain the combined scores $\overline{f_{\text{comb}}^1}, \overline{f_{\text{comb}}^2}, \ldots, \overline{f_{\text{comb}}^K}$ by equations (4) and (5). The second step of the normalisation process is:

$$\text{if } X \in w_i \text{ then } F_{\text{comb}}^k = \overline{f_{\text{comb}}^k} - \overline{f_{\text{comb}}^i} \qquad k = 1, 2, \ldots, K \tag{6}$$

Equation (6) is used to make the correct score zero. We can see from the decision rule, equation (3), that these two steps of normalisation do not change the identification result because the new scores in (6) are obtained only by subtracting and dividing the same number from the original scores, which does not influence the rank of the scores.

After normalisation, the next step is to estimate the probability distribution of the scores. We assume that the values of the score functions are independent. That is:

$$P\left(F_{\text{comb}}^1, \ldots, F_{\text{comb}}^{i-1}, F_{\text{comb}}^{i+1}, \ldots, F_{\text{comb}}^K | X \in w_i\right) = \prod_{k=1, k\neq i}^{K} P\left(F_{\text{comb}}^k | X \in w_i\right)$$

The reason why $k \neq i$ is that, after the normalisation, F_{comb}^i always equals zero if $X \in w_i$. We denote the correct identification rate (the probability of correct

identification) when $X \in \omega_i$ as $C_i(\alpha)$. Since $F^i_{\text{comb}} \equiv 0$ when $X \in \omega_i$ after the normalisation process, we can calculate $C_i(\alpha)$ on the basis of equation (3) as:

$$C_i(\alpha) = \prod_{k=1,k\neq i}^{K} P\left(F^k_{\text{comb}} < 0 | X \in \omega_i\right) \tag{7}$$

4.2 Probability Density Estimation

To calculate the probability $P\left(F^k_{\text{comb}} < 0 | X \in \omega_i\right)$ for each $k = 1, 2, \ldots, K$, we first have to estimate the probability distribution $P\left(F^k_{\text{comb}} | X \in \omega_i\right)$ from the training data in the form of a Gaussian mixture model. But a problem of sparse data arises when we try to model the distribution this way. In essence, it is hard to estimate the density of a multi-modal data distribution reliably.

Our approach to this problem is to break the available training data up into 'sections', and to treat each of these as a unimodal Gaussian, and then to combine them. Suppose there are M training data available for deciding the weighting parameter α. Among these M files, there are M_1 files belonging to class ω_1, M_2 files belonging to class ω_2, ..., and finally M_K files belonging to class ω_K ($M_1 + M_2 + \ldots + M_K = M$).

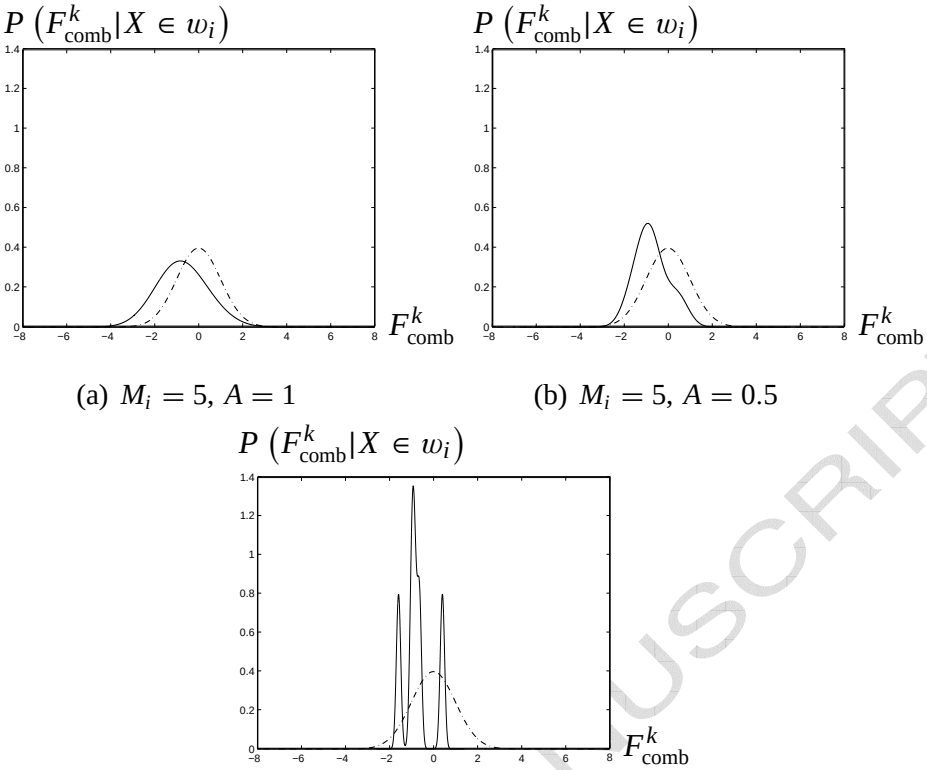
We denote the M_i training data belonging to class ω_i as $X_1, X_2, \ldots, X_{M_i}$. The Gaussian mixture is then:

$$P\left(F^k_{\text{comb}} | X \in \omega_i\right) = \frac{1}{M_i} \sum_{j=1}^{M_i} \frac{1}{\sqrt{2\pi}A} \exp\left(-\frac{(F^k_{\text{comb}} - \mu_{kj})^2}{2A^2}\right) \tag{8}$$

where A is a parameter controlling the variance(s).

The component means μ_{kj} are obtained as $\mu_{kj} = F^k_{\text{comb}}(X_j, \alpha)$, $j = 1, 2, \ldots, M_i$. From this, we see that the means of the mixture components are the scores of the training data. When A is large, the variance of each mixture component is large; when it is small, the variance is small. In the extreme case when A becomes zero, the probability density shrinks to a series of impulse functions.

Figure 2 demonstrates an example of estimating probability density functions using equation (8). The probability density function to be estimated is Gaussian with zero mean and standard deviation of one. It can be seen that in this specific example, the true density function is better estimated when A has a greater value, but this is not always the case. Other distributions may favour smaller rather than greater A. To estimate the probability density distribution using equation (8) with finite data, we have to choose a suitable value of A and it is not clear how this should be done.

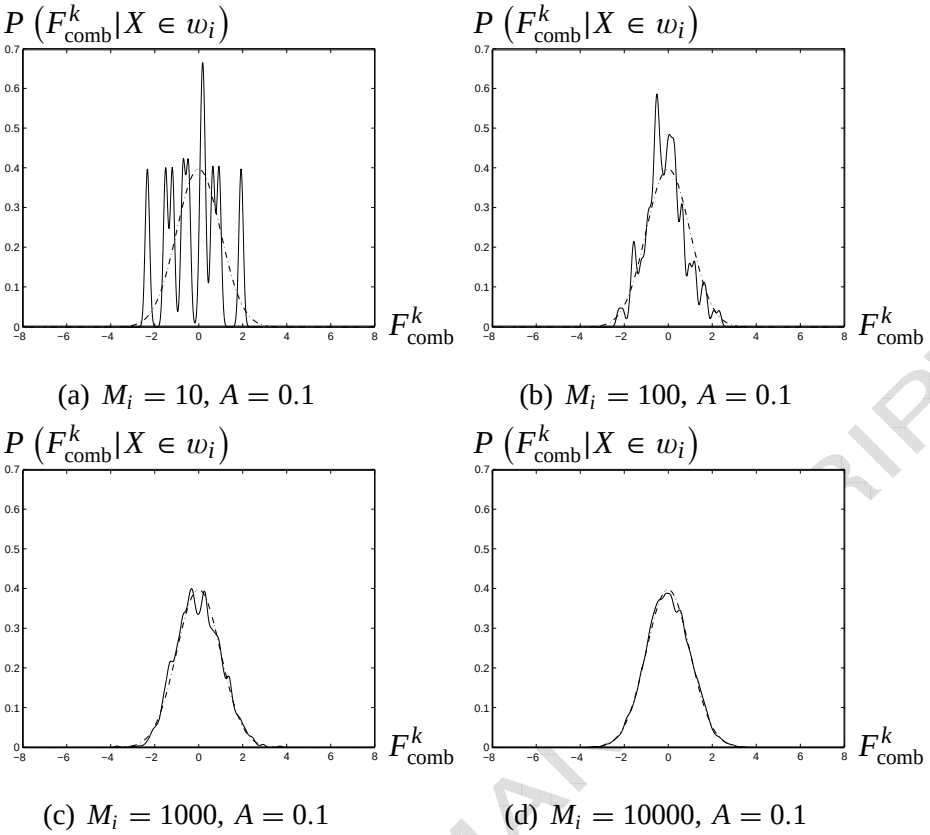


(a) $M_i = 5$, $A = 1$ (b) $M_i = 5$, $A = 0.5$

(c) $M_i = 5$, $A = 0.1$

Fig. 2. Probability density estimation using equation (8). The distribution to be estimated is Gaussian distribution with zero mean and standard deviation of one (as indicated by the dashed lines). In this example, $M_i = 5$, which means that five sample points are drawn from this distribution. Using equation (8), we can obtain the estimated distributions (solid lines) with $A = 1, 0.5$ and 0.1 in (a), (b) and (c), respectively.

However, Bishop (1995, pp. 54–55) proves that when the data are infinite, the expectation of the estimated probability density using the above method will converge to the true probability density. Figure 3 demonstrates the convergence procedure when the number of data increases.



(a) $M_i = 10$, $A = 0.1$ (b) $M_i = 100$, $A = 0.1$

(c) $M_i = 1000$, $A = 0.1$ (d) $M_i = 10000$, $A = 0.1$

Fig. 3. An example to illustrate how the estimated density function (solid lines) reaches the true density function (dashed lines) when increasing M_i. The parameter A is fixed at 0.1 and $M_i = 10, 100, 1000$ and 10000 in (a), (b), (c) and (d), respectively.

4.3 Estimating Correct Identification Rate

Using the estimated pdf, we can now calculate the probability that $F^k_{\text{comb}}(X) < 0$ as:

$$\begin{aligned} &P\left(F^k_{\text{comb}}(X,\alpha) < 0 \,|\, X \in w_i\right) \\ &\quad = \frac{1}{M_i}\sum_{j=1}^{M_i} \frac{1}{\sqrt{2\pi}\,A} \int_{-\infty}^{0} \exp\left(-\frac{(F^k_{\text{comb}}(X,\alpha)-\mu_{kj})^2}{2A^2}\right) d\left[F^k_{\text{comb}}(X,\alpha)\right] \\ &\quad = \frac{1}{M_i}\sum_{j=1}^{M_i} \Phi\left(-\frac{\mu_{kj}}{A}\right) \end{aligned}$$

where $\Phi(x)$ is the integral of the Gaussian distribution:

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx$$

From equation (7), we can finally obtain $C_i(\alpha)$, which is the correct identification rate for a specified α when $X \in w_i$, as:

$$C_i(\alpha) = \frac{1}{M_i^{K-1}} \prod_{k=1,k\neq i}^{K} \left(\sum_{j=1}^{M_i} \Phi\left(-\frac{\mu_{kj}}{A}\right) \right)$$

The overall correct identification rate, denoted $C(\alpha)$, is given as:

$$C(\alpha) = \sum_{i=1}^{K} C_i(\alpha) P(X \in w_i) \tag{9}$$

where $P(X \in w_i)$ can be estimated as $\frac{M_i}{M}$ with M_i equal to the number of training data that belong to class ω_i, and M equal to the total number of training data. Thus, we have transformed the problem of choosing weighting parameter α for combining two classifiers to a problem of maximising the correct identification rate $C(\alpha)$:

$$\text{decide } \alpha = \alpha_{\text{opt}} \text{ if } \alpha = \arg\max_{\alpha} C(\alpha)$$

Once the weighting parameter α is selected using our proposed method, we assume it does not change when it is applied to the test data. Such an assumption is based on a more general assumption that the training data and the test data are independently drawn from the same probability distribution. However, this assumption may not hold in practice, especially when unexpected environmental noise has dramatically changed the probability distribution of the test data. In this situation, it is preferable to use adaptive methods to adjust the weighting parameter(s) (Wark *et al.*, 1999; Wark, 2000; Sanderson and Paliwal, 2003). In this paper, we still assume that the probability distribution of the training data and test data is the same, so that the optimal weighting parameter remains the same for the training and test data, because our focus is on accurate estimation of the parameter under this condition.

5 Results Using Real Data

The database used to test the performance of the proposed method is XM2VTS (Messer *et al.*, 1999). This database, specifically intended for research into multi-modal person identification, is issued by the Centre for Vision, Speech and Signal Processing at the University of Surrey, UK. It contains 4 recordings of 295 subjects taken over a period of four months. Each recording contains a speaking head shot and a rotating head shot, although these are not used here. For each person, static facial images are also provided. The database contains high-quality color images,

32 kHz 16-bit sound files, video sequences and a 3D model. We use the data (video and audio) for 74 people (51 male, 23 female). Each person provides 24 speech files, which were recorded during 4 sessions (6 files for each session), and 8 static images (2 for each session).

The silence removal method for the speaker identification has been discussed in Section 2. We set the upper sound intensity threshold I_1 to be 0.5 times the average sound intensity of the speech file, and the lower intensity threshold I_2 to be 0.2 times this average intensity value. The zero crossing rate threshold Z_1 was set to the average zero crossing rate of the speech file. Careful examination of our results suggest that, in most cases, these settings are reasonable and correctly remove silence while retaining the speech.

For each test speech file, we randomly select three files which are not from the same session as the training set, then test this file with the trained GMM model. This train-3/test-1 strategy is applied to the 24 files for each speaker and 24 identification results are obtained. For training the Gaussian mixture model, we use mel-frequency cepstral coefficients as features (Davis and Mermelstein, 1980). The magnitude spectrum from a 20 ms short-time segment of speech is pre-emphasized and processed by a simulated mel-scale filterbank, then the log-energy filter outputs are cosine transformed to produce the cepstral coefficients. We use the first 20 coefficients, excluding the zeroth coefficient, plus the first 20 delta coefficients as the feature set. This process occurs every 10 ms, producing 100 feature vectors per second. Gaussian mixture density functions consisting of 64 component densities are used in this work.

In building the face classifier, for each image, we use the 6 images in the other 3 sessions as the training set, and then test that image. Such a strategy is applied to all 8 images of each person, obtaining 8 identification results. For each of the 4 sessions, we randomly select 2 of the 6 speaker identification results and then combine them with the 2 face identification results for that session.

Figure 4 shows the empirical correct identification rate $C_e(\alpha)$ as α increases from 0 to 1, with a 0.01 increment on each trial. This is done by first determining the individual scores of both the audio and video classifier, then calculating the combined scores using equation (4), and finally using these for identification. Let us first define the indicator function $T_k(X_i)$ as 0 when $X_i \notin \omega_k$, and 1 when $X_i \in \omega_k$:

$$T_k(X_i) = \begin{cases} 0 & : \quad X_i \notin \omega_k \\ 1 & : \quad X_i \in \omega_k \end{cases} \tag{10}$$

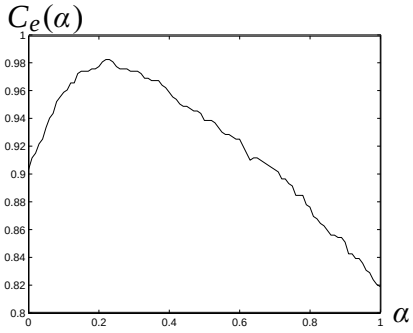


Fig. 4. The empirical correct identification rate using the test data, with α varying from 0 to 1.

Then $C_e(\alpha)$ is defined as follows:

$$C_e(\alpha) = \frac{1}{M}\sum_{i=1}^{M} T_{\hat{k}}(X_i), \text{ where } \hat{k} = \arg\max_{k=1}^{K} f^k_{\text{comb}}(X_i, \alpha) \tag{11}$$

The identification rate for the video classifier is 81.93%, and for the audio classifier it is 90.37%. The combined classifier achieves the highest identification rate (98.31%) when α equals 0.22. However, the empirical identification rate $C_e(\alpha)$ is not a very suitable function to determine the weighting parameter α because of its non-smooth nature, making it difficult to identify a clear peak corresponding to the optimum.

We can also obtain a similar curve, $C_{\text{prop}}(\alpha)$, by estimating the correct identification rate, as proposed in Section 4. Because for the proposed method the scores for both classifiers are normalized by equation (5), some adjustments need to be done to eliminate the effect of normalisation. Recall equation (4), the combination function, is as follows.

$$f^k_{\text{comb}}(X, \alpha) = \alpha f^k_1(X) + (1-\alpha) f^k_2(X) \qquad k = 1, 2, \ldots, K$$

If we replace the original scores $f^k_1(X)$ and $f^k_2(X)$ with the normalized scores $\overline{f^k_1(X)} = \frac{f^k_1(X)-\mu_1}{\sigma_1}$ and $\overline{f^k_2(X)} = \frac{f^k_2(X)-\mu_2}{\sigma_2}$, the weighting parameter α also needs to be changed correspondingly to obtain the same effect. Suppose α is changed to α', equation (4) can be rewritten as follows:

$$\begin{aligned}\overline{f^k_{\text{comb}}(X, \alpha')} &= \alpha' \overline{f^k_1(X)} + (1-\alpha')\overline{f^k_2(X)} \\ &= \alpha' \frac{f^k_1(X)-\mu_1}{\sigma_1} + (1-\alpha')\frac{f^k_2(X)-\mu_2}{\sigma_2} \qquad k = 1, 2, \ldots, K\end{aligned}$$

To obtain the same effect, we must have:

$$\frac{\alpha}{1-\alpha} = \frac{\frac{\alpha'}{\sigma_1}}{\frac{1-\alpha'}{\sigma_2}}$$

Thus, we obtain

$$\alpha' = \frac{\sigma_1 \alpha}{\sigma_2 + (\sigma_1 - \sigma_2)\,\alpha}$$

In order to make $C_{\text{prop}}(\alpha)$ comparable to $C_e(\alpha)$, we must define $C_{\text{prop}}(\alpha)$ as follows:

$$C_{\text{prop}}(\alpha) = C(\alpha') \qquad \alpha' = \frac{\sigma_1 \alpha}{\sigma_2 + (\sigma_1 - \sigma_2)\,\alpha}$$

where $C(\alpha')$ is defined as in equation (9).

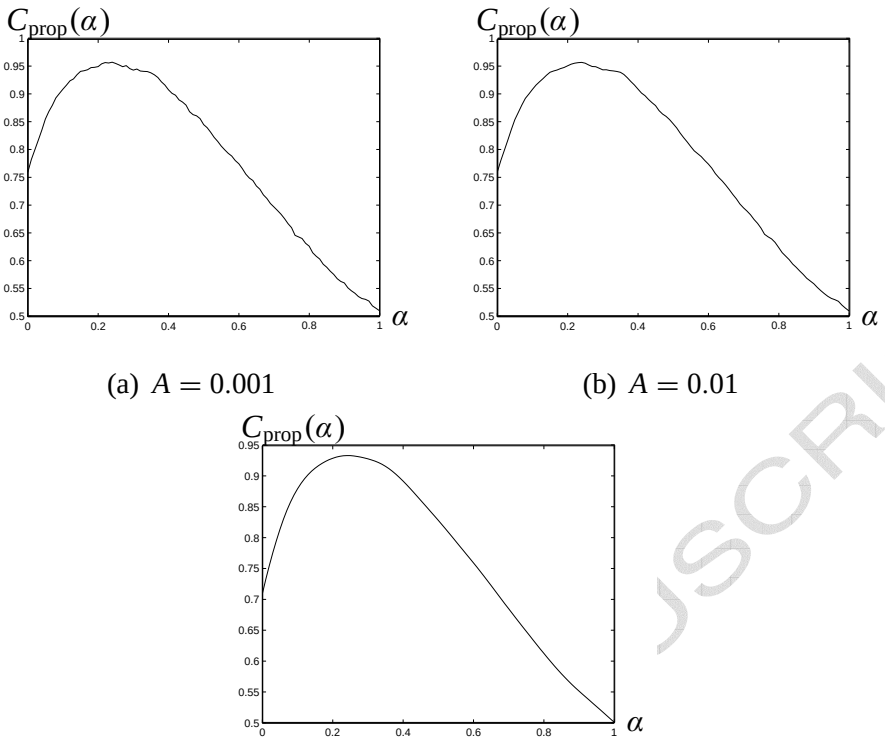
Figure 5 illustrates the obtained correct identification curves when $A = 0.001$, $A = 0.01$ and $A = 0.1$. When A takes a relatively large value, the curve is smoothed relative to the empirical correct identification curve, and the peak of $C_{\text{prop}}(\alpha)$ when α varies from 0 to 1 can be more clearly observed.

We can see from Figure 5 that the estimated correct identification rate is always smaller than the true correct identification rate. This is because the estimated score distribution as in equation (8) does not precisely reflect the true distribution. From Figure 6, we observe that even if all the scores for estimation are below 0, the estimated probability that the score is greater than zero is still 0.02.

That is, in this special case, the estimated identification rate is 2% smaller than the true correct identification rate. If we add up all these score distributions as in equations (7) and (9), we can also expect the estimated identification rate to be smaller than the true identification rate. We can further estimate that when A becomes larger, the estimated identification rate will be even smaller. This is why the values of $C_{\text{prop}}(\alpha)$ are different when A takes different values. But we can see from the above results that this problem does not interfere the process of deciding the weighting parameter α, because we only need to find the α that gives the maximum value of $C_{\text{prop}}(\alpha)$.

6 Further Results Using Bootstrapping

The empirical identification curve in Figure 4 shows that the maximal identification rate is achieved when $\alpha = 0.22$, while all the three identification rate curves in



(a) $A = 0.001$ (b) $A = 0.01$

(c) $A = 0.1$

Fig. 5. The correct identification rate $C_{\text{prop}}(\alpha)$ using the proposed method as α varies from 0 to 1: (a) $A = 0.001$; (b) $A = 0.01$: (c) $A = 0.1$.

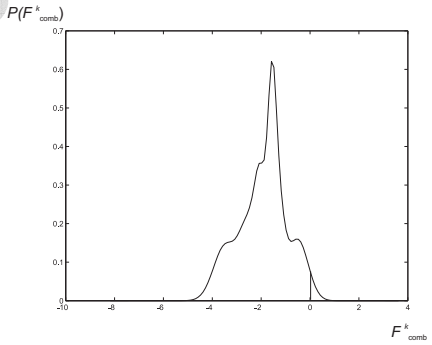


Fig. 6. The estimated probability density of $f^k_{\text{comb}}(X, \alpha)$ when $A = 0.5$ and $f^k_{\text{comb}}(X_1, \alpha) = -3.48$, $f^k_{\text{comb}}(X_2, \alpha) = -2.54$, $f^k_{\text{comb}}(X_3, \alpha) = -2.02$, $f^k_{\text{comb}}(X_4, \alpha) = -1.56$, $f^k_{\text{comb}}(X_5, \alpha) = -1.34$ and $f^k_{\text{comb}}(X_6, \alpha) = -0.50$. It indicates that even if all the scores for estimation are below 0, the estimated probability that the score is greater than 0 is 0.02.

Figure 5 using the proposed method indicate that the optimal α is 0.24. Our intuition told us that the estimation by the proposed method is more accurate because it provides a smooth curve, thus reducing the possibility of over-fitting. In this section, we use the bootstrap method to indicate that, compared with the empirical method and other frequently-used methods, the proposed method performs well in reducing the variance of the estimated optimal weighting parameter, thus suggesting a more accurate estimation.

The publication in 1979 of Bradley Efron's first article on bootstrap methods was a major event in statistics, at once synthesising some of the earlier resampling ideas and establishing a new framework for statistical analysis. It has been shown that bootstrap methods often perform better than traditional methods in many applications. The reader is referred to Davison and Hinkley (1997) for a detailed discussion.

The bootstrapping is performed as follows. As indicated in Section 5, there are 8 face identification results and 24 speaker identification results for each person. In each bootstrap process, we randomly select 8 speaker identification results out of these 24, then combine them with the 8 face identification results, and obtain estimates of the optimal α by both the empirical and proposed methods. The bootstrap process is repeated N times (i.e., repeating the sampling of 8 from 24 files), so obtaining N estimates of the optimal α, one for each process, which are represented as $\alpha_{\text{opt}}^1, \alpha_{\text{opt}}^2, \ldots, \alpha_{\text{opt}}^N$. The mean and variance of α_{opt} can be calculated as follows:

$$\overline{\alpha_{\text{opt}}} = \frac{1}{N}\sum_{i=1}^{N} \alpha_{\text{opt}}^i$$

$$\sigma_{\alpha_{\text{opt}}} = \frac{1}{N-1}\sum_{i=1}^{N} \left(\alpha_{\text{opt}}^i - \overline{\alpha_{\text{opt}}}\right)^2$$

We have tested four methods for choosing the optimal weighting parameter using the bootstrap method:

(1) the empirical method based on actual identification results;
(2) the proposed method based on pdf estimation;
(3) smoothed error rate estimation; and
(4) a genetic algorithm, as proposed by Lam and Suen (1995).

The first two have already been described. The smoothed error rate estimation method was first used by Ney (1995), and subsequently in audio-visual speaker identification by Maison *et al.* (1999). This method shares some similarities with our method (2) as proposed in this paper, which is also a smoothing technique for the correct identification curve. Instead of finding the value of α

that maximizes $C_e(\alpha)$, the smoothed error rate estimation method finds the α that maximizes $C_{\text{smooth}}(\alpha)$ defined as follows:

$$C_{\text{smooth}}(\alpha) = \frac{1}{M}\sum_{i=1}^{M}\sum_{k=1}^{k} T_k(X_i)\frac{\exp\left\{\eta f_{\text{comb}}^k(X_i,\alpha)\right\}}{\sum_{j=1}^{K}\exp\left\{\eta f_{\text{comb}}^j(X_i,\alpha)\right\}}$$

Here M is the total number of training data, K is the total number of classes and $T_k(X_i)$ has been defined in equation (10). We note here that it depends on choosing a parameter η that when large reduces the smoothed error rate to the empirical one.

Lam and Suen's method attempts to find the optimal weighting parameters using a genetic algorithm. This method assigns a weighting parameter to each classifier, making the fusion function as follows:

$$f_{\text{comb}}^k(X,\alpha_1,\alpha_2) = \alpha_1 f_1^k(X) + \alpha_2 f_2^k(X)$$

Then the fitness function which the genetic algorithm needs to maximize is set as:

$$C_{\text{ga}}(\alpha_1,\alpha_2) = \frac{1}{M}\sum_{i=1}^{M} T_{\hat{k}}(X_i), \text{ where } \hat{k} = \arg\max_{k=1}^{K} f_{\text{comb}}^k(X_i,\alpha_1,\alpha_2)$$

where M is the number of training data, K is the number of classes, and $T_k(X_i)$ is defined as in equation (10).

The genetic algorithm is used to search for the α_1 and α_2 which maximize $C_{\text{ga}}(\alpha_1,\alpha_2)$. Our settings of parameters is slightly different from the original paper. The population size is 20. The fractions of crossover and migration are 0.8 and 0.2, respectively. Because we used the default settings of the Matlab GA function, the reader can refer to the Matlab function `ga(fitnessfcn,nvars)` for the settings of other parameters. The reader may refer to the MathWorks' website `http://www.mathworks.com/` for an introduction to Matlab software.

The GA algorithm runs for 100 generations and picks the α_1 and α_2 which yield the largest value of $C_{\text{ga}}(\alpha_1,\alpha_2)$. To make the GA method comparable with other methods, the optimal weighting coefficient α is then found as $\frac{\alpha}{1-\alpha} = \frac{a_1}{a_2}$.

Table 1 shows the means $\overline{\alpha_{\text{opt}}}$ and variances $\sigma_{\alpha_{\text{opt}}}$ of these four methods using 200 bootstrap iterations ($N = 200$). We have used a range of A values for the proposed method and, similarly, a range of η values for the smoothed error rate estimation method. Those shown in table are the sub-ranges over which good estimates (i.e., low variances) were obtained.

Method		$\overline{\alpha_{\text{opt}}}$	$\sigma_{\alpha_{\text{opt}}}$
Empirical		0.2548	0.0545
Genetic Algorithm		0.2685	0.0547
Proposed	$A = 0.001$	0.2478	0.0419
	$A = 0.002$	0.2464	0.0421
	$A = 0.005$	0.2470	0.0425
	$A = 0.01$	0.2480	0.0400
	$A = 0.02$	0.2665	0.0386
	$A = 0.05$	0.2488	0.0358
	$A = 0.1$	0.2635	0.0308
	$A = 0.2$	0.3001	0.0355
Smoothed Error Rate	$\eta = 5$	0.2664	0.0552
	$\eta = 10$	0.3598	0.0388
	$\eta = 15$	0.3051	0.0373
	$\eta = 20$	0.2879	0.0388
	$\eta = 25$	0.2798	0.0411
	$\eta = 30$	0.2766	0.0424
	$\eta = 35$	0.2683	0.0475
	$\eta = 40$	0.2667	0.0486

Table 1
The means and variances of the four methods for estimating the optimal weighting parameter α_{opt} with the real speech and video data.

We can see from the table that the four methods provide similar means. The proposed method and the smoothed error rate method give generally smaller variances than the other two methods (although this is of course achieved with the advantage of an adjustable parameter). The proposed method appears to give a rather smaller variance than the smoothed error rate method, but this is uncertain.

7 Results with Simulated Data

Table 1 indicates that the proposed method performs slightly better in reducing the estimation variance, but it does not show that this method is also good at reducing the estimation bias, i.e., if $\overline{\alpha_{\text{opt}}}$ estimated by this method is close to the true value of the optimal weighting parameter. With the real data used in the previous section,

this question can not be answered because this true value is unknown.

In this section, we try to answer this question in some aspects. First, we construct simulated data with a known probability distribution, so the true value of the optimal weighting parameter can be exactly calculated. Finally, we use the bootstrap method with $N = 200$ to estimate the optimal weighting parameter, and see how close the estimated optimal weighting parameter is to the true value of that parameter. The idea of using simulated data to test classifier combination methods was proposed by Kittler and Alkoot (2003), and has become a benchmark approach. Here will will also use simulated data to test our method of choosing weighting parameters.

Consider a K-class problem. We need to construct the K scores of an input X which belongs to a specific class, say, $X \in \omega_k$, where $k \in \{1, 2, \ldots, K\}$. First, we generate K random numbers, each of which is uniformly distributed in the range $[0, 200]$. We use $n_1, n_2, \ldots, n_K$ to represent these K numbers. We choose n_k as the maximum of these K numbers ($n_k = \max\{n_1, n_2, \ldots, n_K\}$), since it is reasonable to assume that the highest score will be obtained for the correct class. Next we generate another K random numbers, $\sigma_1, \sigma_2, \ldots, \sigma_K$, each of which is uniformly distributed in the range $[0, \sigma_{\max}]$. Here $\sigma_{\max}$ is a controlling parameter. The scores $f^1(X), f^2(X), \ldots, f^K(X)$ are generated as follows.

For each $k \in \{1, 2, \ldots, K\}$, $f^k(X)$ is a random sample drawn from a normal distribution with mean n_k and variance σ_k. We construct two classifiers, denoted Classifier 1 and Classifier 2. For Classifier 1, we set its $\sigma_{\max}$ to 10; and for Classifier 2, we set its $\sigma_{\max}$ to 20. Thus, Classifier 1 is a strong classifier and Classifier 2 is a weak classifier. For both , we set K to 74, equal to the number of classes in the audio-visual speaker identification task. For each class, we generate 8 sets of scores from Classifier 1, and 24 sets of scores from Classifier 2, which is also the same as the audio-visual speaker identification task. Using the bootstrap method, we then obtain the means and variances of the four methods. Since the simulated data are generated from a known distribution, we can also accurately calculate the true optimal weighting parameter, α_{true}, by using all the parameters $n_1, n_2, \ldots, n_K$ and $\sigma_1, \sigma_2, \ldots, \sigma_K$. For simplicity, the details of how to calculate α_{true} are omitted here. We mention only that we can accurately calculate α_{true} since the score distributions are known.

Table 2 shows the estimated means and variances for normally-distributed simulated data using the four methods. Here, $\alpha_{\text{true}} = 0.650$. It can be observed that the means of the empirical, genetic algorithm and proposed methods are closer to α_{true} than the smoothed error rate method, but the proposed method gives much smaller variance than the empirical and genetic algorithm methods. However, we need to remember that the simulated data are generated with a Gaussian distribution, so conforming to the major assumption underlying our proposed method. Thus, we have also carried out performance comparisons with data with a rectangular

Method		$\overline{\alpha_{\text{opt}}}$	$\sigma_{\alpha_{\text{opt}}}$
Empirical		0.6509	0.0428
Genetic Algorithm		0.6510	0.0399
Proposed	$A = 0.001$	0.6494	0.0350
	$A = 0.002$	0.6499	0.0344
	$A = 0.005$	0.6494	0.0303
	$A = 0.01$	0.6520	0.0298
	$A = 0.02$	0.6520	0.0277
	$A = 0.05$	0.6491	0.0180
	$A = 0.1$	0.6185	0.0089
	$A = 0.2$	0.5719	0.0058
Smoothed Error Rate	$\eta = 5$	0.5225	0.0045
	$\eta = 10$	0.5750	0.0047
	$\eta = 15$	0.6070	0.0078
	$\eta = 20$	0.6269	0.0113
	$\eta = 25$	0.6358	0.0124
	$\eta = 30$	0.6436	0.0159
	$\eta = 35$	0.6436	0.0159
	$\eta = 40$	0.6436	0.0159

Table 2
The means and variances of four methods for estimating α_{opt} on simulated data generated to have a Gaussian distribution. Here $\alpha_{\text{opt}} = 0.650$.

(uniform) distribution. This should show our method at maximum disadvantage relative to the competitors.

As before, the sets of random numbers were generated from which we obtain n_k and σ_k. A uniform distribution in the range $(n_k - \sigma_k, n_k + \sigma_k)$ was then generated. The results in Table 3 for the data with uniform distribution show that the proposed method holds up well in the face of violation of the underlying assumption of normally-distributed data. The optimal weighting parameter is estimated with very low bias and low variance, certainly relative to the empirical and GA methods. Performance is slightly but noticeably better than the smoothed error rate method.

It is not suitable to use the empirical method directly to decide the optimal weighting parameter, because it gives very high variance. A better solution is to calculate the average of the optimal weighting parameters by using the bootstrap

Method		$\overline{\alpha_{\text{opt}}}$	$\sigma_{\alpha_{\text{opt}}}$
Empirical		0.7324	0.0492
Genetic Algorithm		0.7368	0.0529
Proposed	$A = 0.001$	0.7359	0.0419
	$A = 0.002$	0.7367	0.0413
	$A = 0.005$	0.7346	0.0385
	$A = 0.01$	0.7343	0.0306
	$A = 0.02$	0.7327	0.0220
	$A = 0.05$	0.7270	0.0176
	$A = 0.1$	0.7177	0.0224
	$A = 0.2$	0.7541	0.0519
Smoothed Error Rate	$\eta = 5$	0.6106	0.1486
	$\eta = 10$	0.7043	0.0525
	$\eta = 15$	0.7135	0.0314
	$\eta = 20$	0.7182	0.0255
	$\eta = 25$	0.7224	0.0235
	$\eta = 30$	0.7256	0.0234
	$\eta = 35$	0.7256	0.0234
	$\eta = 40$	0.7256	0.0234

Table 3
The means and variances of four methods for estimating α_{opt} on simulated data generated with a uniform distribution. Here $\alpha_{\text{opt}} = 0.727$.

method. In situations where the training data are sparse, so that it is difficult to use the bootstrap method, the proposed method is highly recommended.

The main drawback of the proposed method is that we have to choose a suitable value of A and it is not clear how this should be done. Of course, the smoothed error rate technique shares this kind of problem, in that we have to fix a suitable value of η.

8 Conclusions and Future Work

This paper provides a method to estimate the optimal weighting parameter for fusion of scores in audio-visual person identification. It is based on estimation

of probability density functions for the scores under a Gaussian assumption. By use of bootstrapping, the performance of this method can be strictly analysed and compared with other methods. Using simulated data, such that the pdf is known, results indicate that this method has advantages in reducing the bias and variance of the estimation. The method is shown to perform well even when the underlying Gaussian assumption is violated. The main problem is in choosing a suitable value of smoothing parameter A. It is not clear at present how this should best be done.

The validity of the proposed method is based on two assumptions. First, the bootstrapping method as discussed in Section 6 is based on the assumption that the performances of the audio classifier and the visual classifier are independent. Intuitively, such an assumption is true because we have little information to imagining a person's face when only listening to his/her voice, and vice versa. Our future work will investigate whether the bootstrapping method is valid when the two modalities are strongly correlated. Another assumption is that, as discussed in Section 4.3, the training and the test data are drawn independently from a fixed probability distribution; thus, the optimal weighting parameter remains unchanged. Although this assumption is very common in theoretical pattern recognition studies (Vapnik, 1998), it may not be valid in practice. Thus, adaptive methods for choosing weighting parameter(s) may be preferable in practical situations.

It should be noted that, although our method is developed for the identification task, it can be applied to verification. For verification, a similar approach can be taken for choosing the optimal weighting parameter based on minimising the equal error rate (EER), instead of maximising the correct identification rate for the identification case. One of our future works is to generalize this method to person verification.

References

Adjoudani, A. and Benoît, C. (1995). Audio-visual speech recognition compared across two architectures. In *Proceedings of the 4th European Conference on Speech Communication and Technology*, volume 2, pages 1563–1567, Madrid, Spain.

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

Bengio, S. (2003). Multimodal authentication using asynchronous HMMs. In *Proceedings of the 4th International Conference on Audio- and Video-based Biometric Person Authentication*, pages 770–777, Guildford, UK.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford, UK.

Brunelli, R. and Falavigna, D. (1995). Person identification using multiple cues.

IEEE Transactions on Pattern Analysis and Machine Intelligence, **17**(10), 955–966.

Chibelushi, C. C., Deravi, F., and Mason, J. S. (1993). Voice and facial image integration for speaker recognition. In *IEEE International Symposium on Multimedia Technologies and Future Applications*, pages 155–161, Southampton, UK.

Davis, S. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, **28**(4), 357–366.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge University Press, Cambridge, UK.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum-likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **39**(1), 1–38.

Duin, R. P. W. (2002). The combining classifier: To train or not to train? In *Proceedings of 16th International Conference on Pattern Recognition*, volume II, pages 765–770, Quebec City, Canada.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, **7**(1), 1–26.

Erzin, E., Yemez, Y., and Tekalp, A. M. (2005). Multimodal speaker identification using adaptive classifier cascade based on modality reliability. *IEEE Transactions on Multimedia*, **7**(5), 840–852.

Fu, T., Liu, X. X., Liang, L. H., Pi, X., and Nefian, A. V. (2003). Audio-visual speaker identification using coupled hidden Markov model. In *Proceedings of the IEEE International Conference on Image Processing, ICIP'03*, volume 3, pages 29–32, Barcelona, Spain.

Hu, R. and Damper, R. I. (2005). Fusion of two classifiers for speaker identification: Removing and not removing silence. In *Proceedings of 8th International Conference on Information Fusion*, volume 1, pages 429–436, Philadelphia, PA.

Hu, R. and Damper, R. I. (2006). A 'no panacea' theorem for multiple classifier combination. In *International Conference on Pattern Recognition, (ICPR 2006)*, Hong Kong, China. No pagination, Proceedings on CD-ROM.

Jain, A., Nandakumar, K., and Ross, A. (2005). Score normalization in multimodal biometric systems. *Pattern Recognition*, **38**(12), 2270–2285.

Kittler, J. and Alkoot, F. M. (2003). Sum versus vote fusion in multiple classifier systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **25**(1), 110–115.

Kittler, J., Hatef, M., Duin, R. P. W., and Matas, J. (1998). On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **20**(3), 226–239.

Lades, M., Vorbruggen, J., Buhmann, J., Lange, J., von der Malsburg, C., and Wurtz, R. (1993). Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers*, **42**(3), 300–311.

Lam, L. and Suen, C. Y. (1995). Optimal combination of pattern classifiers. *Pattern Recognition Letters*, **16**(9), 945–954.

Lucey, S., Chen, T., Sridharan, S., and Chandran, V. (2005). Integration strategies for audio-visual speech processing: Applied to text-dependent speaker recognition. *IEEE Transactions on Multimedia*, **7**(3), 495–506.

Luettin, J. (1997). *Visual Speech and Speaker Recognition*. PhD thesis, Department of Electronic and Electrical Engineering, University of Sheffield, Sheffield, UK.

Maison, B., Neti, C., and Senior, A. (1999). Audio-visual speaker recognition for video broadcast news: some fusion techniques. In *Proceedings of the IEEE Workshop on Multimedia Signal Processing*, pages 161–167, Copenhagen, Denmark.

Messer, K., Matas, J., Kittler, J., Luettin, J., and Maitre, G. (1999). XM2VTSDB: The extended M2VTS database. In *Proceedings of 2nd International Conference on Audio and Video-based Biometric Person Authentication, AVBPA'99*, pages 72–77, Washington, DC.

Ney, H. (1995). On the probabilistic interpretation of neural network classifiers and discriminative training criteria. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **17**(2), 107–119.

Rabiner, L. R. and Sambur, M. R. (1975). An algorithm for determining the endpoints of isolated utterances. *Bell Systems Technical Journal*, **54**(2), 297–315.

Reynolds, D. A. and Rose, R. C. (1995). Robust text-independent speaker identification using Gaussian mixture models. *IEEE Transactions on Speech and Audio Processing*, **3**(1), 72–83.

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

Toh, K.-A. and Yau, W.-Y. (2004). Combination of hyperbolic functions for multimodal biometrics data fusion. *IEEE Transactions on System, Man, and Cybernetics: Part B Cybernetics*, **34**(2), 1196–1209.

Vapnik, V. N. (1998). *Statistical Learning Theory*. Wiley, New York, NY.

Wark, T. (2000). *Multi-Modal Speech Processing for Automatic Speaker Recognition*. PhD thesis, School of Electrical and Electronic Systems Engineering, Queensland University of Technology, Brisbane, Australia.

Wark, T., Sridharan, S., and Chandran, V. (1999). Robust speaker verification via fusion of speech and lip modalities. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP'99*, volume 6, pages 3061–3064, Phoenix, AZ.

Wiskott, L., Fellous, J.-M., Krüger, N., and von der Malsburg, C. (1997). Face recognition by elastic bunch graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **19**(7), 775–779.

Xu, L., Krzyzak, A., and Suen, C. Y. (1992). Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Transactions on Systems, Man, and Cybernetics*, **22**(3), 418–435.