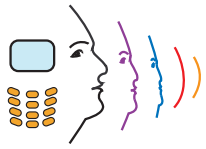


MOBIO

Mobile Biometry

http://www.mobioproject.org/

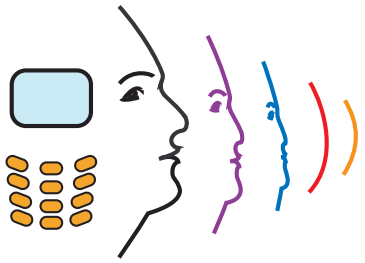
Funded under the 7th FP (Seventh Framework Programme)
Theme ICT-2007.1.4
[Secure, dependable and trusted Infrastructure]

D4.4: Description and evaluation of advanced algorithms for joint bi-modal authentication

Due date: 15/06/2010 **Submission date:** 17/05/2010
Project start date: 01/01/2008 **Duration:** 36 months
WP Manager: Norman Poh **Revision:** 1

Author(s): Anindya Roy and Sébastien Marcel

Project funded by the European Commission in the 7th Framework Programme (2008-2010)		
Dissemination Level		
PU	Public	Yes
RE	Restricted to a group specified by the consortium (includes Commission Services)	No
CO	Confidential, only for members of the consortium (includes Commission Services)	No



D4.4: Description and evaluation of advanced algorithms for joint bi-modal authentication

Abstract:

In this deliverable, we proposed two advanced fusion schemes for audio-visual person authentication: 1) a video-based score-level fusion scheme that uses a set of score distribution descriptors extracted from video, and 2) a feature-level fusion scheme that uses a novel concept called audio-visual slice. The first framework exploits the abundant score information present in video streams, while the second exploits person-specific information in the joint audio-visual space. Both of them were evaluated on standard audio-visual databases and were shown to perform reasonably well, with each of them showing particular advantages over other reference systems.

Contents

1 Introduction

In this deliverable, we investigate advanced fusion techniques for joint audio-visual person authentication. This work is divided into two distinct parts. In the first part, we propose a novel video-based framework for score-level fusion. In the second part, we propose a novel framework for feature-level fusion using a boosted ensemble of classifiers.

Video-based biometric systems are becoming feasible thanks to advancement in both algorithms [VJ04] and computation platforms. Such systems have many advantages: improved robustness to spoof attack, performance gain thanks to variance reduction [PB03], and increased resolution [WLT07]. In the first part of this work, we propose a framework for video-based score-level fusion which exploits statistics of scores extracted from multiple frames of video. Our framework enables an existing biometric system to further harness the availability of abundant scores derived from frames of video, using a set of distribution descriptors. Experimental results based on face and speech unimodal systems, as well as multimodal fusion, show that our proposal can improve over the standard fixed rule strategies by as much as 50%.

Feature-level fusion has a unique advantage over score-level fusion: it would be able to extract person-specific information which might be embedded *jointly* in both the modalities. In the second part of this work, we investigate feature-level fusion of audio and visual modalities. We first perform a preliminary study aimed at finding out if feature-level fusion would really benefit a biometric system. Next, we propose a method for feature-level fusion using a feature combination technique called "slice", a 2-dimensional projection of the joint audio-visual space. We use this concept in a boosting framework [FHT98] to create a computationally efficient bimodal fusion system. Experimental results suggest that the proposed feature-level fusion system compares well with a standard reference system based on score fusion and is particularly robust to high levels of noise in the audio modality.

In the following sections, we describe each of these two contributions in detail.

2 Video-based Score-level Fusion

Thanks to the advancement of sensing technology as well as advancement in highly efficient computational algorithms and hardware computation platforms, it is now possible to acquire and process video biometric data in real time. For instance, Viola and Jones' face detector [VJ04] is a vivid manifestation of such advancement. There are at least three potential advantages in exploiting the temporal information for biometric person recognition: (i) as an evidence of biometric liveness, for instance, using continuously acquired fingerprint images [AS09] or in audio-visual biometrics [BMWC06]; (ii) improved accuracy via variance reduction [PB03, PB05] and (iii) as a means to derive super-resolution biometric samples [WLT07].

Information fusion can, in principle, be carried out at the data-level for each frame of a video or at the score-level where a classifier is invoked. The former requires not only vast amount of computational resources, but also modification to the matching function in order

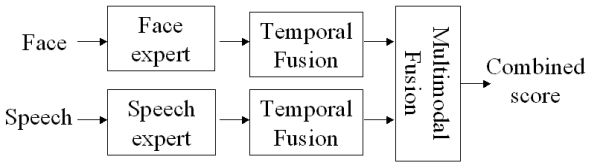


Figure 1: The architecture of our proposal

to compare the biometric samples as image sets [KKC07]. In comparison, by working at the score-level, the latter effectively conceals the complex information associated with the extracted features or raw biometric data. Furthermore, an existing classifier can be used without modification. As a matter of fact, most well established face recognition software uses proprietary algorithms. As a result, the internal functionalities of the system (feature representation and the matching function) are not accessible. For the above reasons, video-derived score-level fusion is a first resort before applying such image-set based approach as [KKC07].

In the literature, score-level fusion often involves scores derived from multiple sensors (e.g., 3D vs. 2D sensors), multiple modalities (body parts), multiple comparison (matching) algorithms (e.g., text-based vs minutiae-based fingerprint matching algorithms), multiple instances (e.g., left vs. right iris images) and multiple samples (e.g., face images observed from cameras positioned at different angles) [RP09]. Our fusion problem can thus be considered a special case of multi-sample fusion wherein the biometric samples come from a video stream. The distinguishing feature of video-based multi-sample fusion with the one referred to in the literature is that the former can potentially contain hundreds of observations.

We therefore extend the concept of multi-sample score-level fusion to include the notion of time, which we refer to as *short-term temporal fusion* or simply temporal fusion. The aim here is to investigate if the overall system accuracy can be better than the conventional approach using such simple rules as mean, maximum or minimum of the scores. Rather than using these simple statistics, for each stream of data, we propose to coarsely characterize the distribution of the score sets by a vector of distribution descriptors consisting of mean, standard deviation, skewness, median, the 25-th and 75-th percentiles, as well as minimum and maximum of the scores. The importance of these parameters with respect to the classification task is directly learned from the data via logistic regression. If there are two modalities, the temporal information fusion first takes place for each modality and the resultant scores are then combined at the multimodal level using the product rule (see Figure 1).

Our contributions are as follow: First, we contribute to the state of the art in information fusion by proposing a *video-based* multi-sample score-level fusion framework (Section 2.1). Second, we explore a novel fusion strategy utilizing a vector of descriptors of the score distributions derived from the video (also in Section 2.1). Third, we demonstrate the merit of our proposal on both the (talking) face and speech biometric modalities

(Section 2.2).

Experimental evidence obtained from the BANCA (talking) face and speech video database suggests that our proposal can improve the face expert by as much as 30% and the speech expert by 10%. Furthermore, the fusion of the two modality-dependent experts, after applying the temporal fusion, results in a relative improvement of 50% compared to the conventional fusion (using only simple statistics to combine video-based scores), i.e., from about 2% of equal error rate to 1%.

2.1 A video-based multi-sample score-level fusion framework

Let $\mathcal{Y} \in \{y_1, \ldots, y_N\}$ denote the set of matching scores $y_i \in \mathbb{R}$ derived from a video query consisting of N frames of processed and *valid* biometric samples. For instance, for the speech modality that we will use, N means the total number of Mel-scale Frequency Cepstral Coefficient (MFCC) features containing *voiced* speech (with the silence segments removed). For the face modality, N denotes the total number of images for which our face detector can confidently find a face and the face matching algorithm can produce a score.

The most conventional strategy to obtaining a single score y_{com} from the score set $\mathcal{Y}$ is to use a simple fixed fusion rule. For the speech expert whose output is a log-likelihood ratio of two hypotheses – one hypothesis supporting that the claimant utterance comes from the "target" speaker or enrolled client versus the alternative. The speech expert that we use is a modified state-of-the-art classifier based on Gaussian Mixture Model with Maximum a posteriori adaptation (MAP-GMM) [RQD00]. It combines the N scores by using the mean rule:

$$y_{com} = \frac{1}{N} \sum_{i=1}^{N} y_i$$

We employed two *parts-based* face experts (systems) which are very different in architecture. By parts-based we understand that a face image is represented by a set of fixed-size windows of much smaller size then the original image. The first expert uses a subset of coefficients of a Discrete Cosine Transform [CSB06], known as *DCTmod2*, to represent the texture information of each subwindow. The sequence of DCTmod2 features so-derived is then classified using the MAP-GMM approach similar to the speech expert. Subsequently, the N scores (from a video) are combined using the mean rule.

The second face expert represents each subwindow using non-uniform Local Binary Pattern (LBP) followed by a Fisher Linear Discriminant (FLD) projection [Cha08]. During query, a template feature in the LBP-FLD space is compared with that of a query feature using normalized correlation. The matching scores for the respective subwindows are then averaged to produce y_i for each frame i in the video. It was empirically found that the maximum rule works best to combine the scores from each video:

$$y_{com} = \max_{y \in \mathcal{Y}} y$$

In the sequel, we studied two possible types of architecture: one generative and another discriminative. The basic idea of both types of architecture departs significantly from the

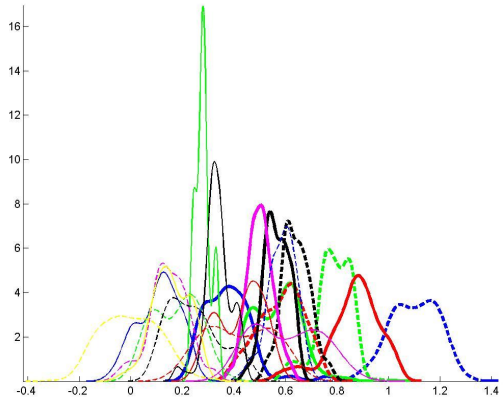


Figure 2: Each thick (resp. thin) curve denotes the *match* (resp. *non-match*) score distribution estimated from a video sequence. These scores are obtained from the UNIS face expert based on LBP.

simple fixed rules in two ways. First, the video-based scores are now treated as a *set* which follows a certain distribution. Second, the distributions of client and impostor video-based scores are distinctive and repeatable for different legitimate users (also known as "target" users or clients in the speaker recognition literature or "gallery subjects" in the face recognition literature). Some training data has to be necessarily made available in order to characterize the score distributions.

In both types of architecture, we shall estimate the density of $y \in \mathcal{Y}$. Let this estimated score density of a query video be $\hat{p}(y)$. We shall assume the availability of some training data in terms of video-based score sets of known labels. At this point, it is useful to distinguish the *query* video score set, $\mathcal{Y}$, from the *training* video score sets, $\mathcal{Y}'_{pq}$. In the latter case, p is the identity index assigned to a legitimate user (i.e., the *claimed identity*) and q is the identity of the actual subject from which a biometric is acquired (i.e., the *true identity*). A video score set is considered a *match* when $p = q$ and is considered is *non-match* when $p \neq q$. In the discussion that follows, the match event is denoted as ω_1 and the non-match event is denoted as ω_0.

Figure 2 plots the estimated density $\hat{(}y)$ under the match and the non-match events. Each curve in this figure represent a score distribution estimated from *one* video sequence. Two important observations can be made by looking at this figure. First, the expected value of a *match* video score set is larger than the *non-match* score set counterpart. The reason for this is that the scores are similarity or likelihood scores, hence, higher values imply that a query video is a match. Second, the variance of a video-based match score set is larger when the query video is a match than when it is a non-match. In the experiments (Section 2.2), we verified that the two observations above are *consistent* for both the face and the speech modalities.

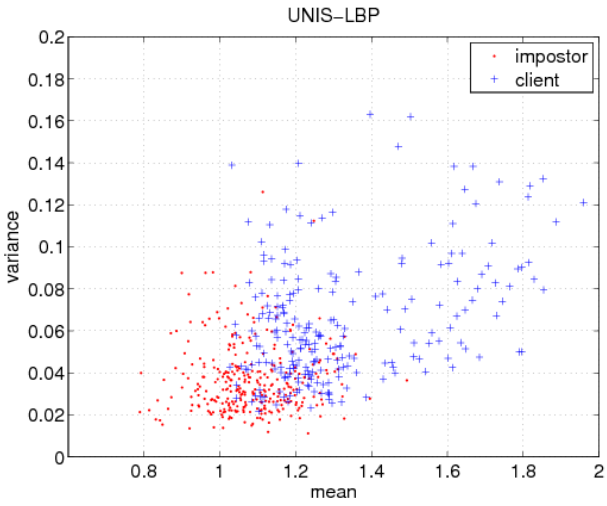


Figure 3: Scatter plot of mean versus standard deviation of the distribution parameters conditioned on match and non-match video queries.

The training score sets, arising from the non-match event, is denoted by $\mathcal{Y}_0 \equiv \cup_p \mathcal{Y}'_{p,q}$ for all $p \neq q$. The corresponding score sets, for the match event, is denoted by $\mathcal{Y}_1 \equiv \cup_p \mathcal{Y}'_{p,p}$ for all legitimate users p. $\mathcal{Y}_0$ therefore contains *non-match* (impostor) scores pooled from many non-match video score sets. Similarly, $\mathcal{Y}_1$ contains *match* (genuine) scores pooled from many match video score sets.

Let $\hat{p}(y|\mathcal{Y}_k)$ denotes the distribution estimated from $\mathcal{Y}_k$ for $k \in \{0, 1\}$. The generative approach we considered consists of computing the following distance:

$$y_{com} = dist(\hat{p}(y|\mathcal{Y}), \hat{p}(y|\mathcal{Y}_1) - dist(\hat{p}(y|\mathcal{Y}), \hat{p}(y|\mathcal{Y}_0) \quad (1)$$

where *dist* is a distance metric between two distributions. The most common choice of this metric, without making any assumption about the form of the distribution is the relative entropy (or Kullback Leibler divergence), Bhattacharyya distance, Chi-square or histogram intersection [CS02].

There are several difficulties when using the above generative approach. First, one has to estimate the shape of distributions from the query data as well as the training data. This requires choosing the right form of distribution or else resorting to using a non-parametric approach such as the kernel density approach (Parzen window) [Bis99, Chap 2]. Second, one needs to choose a distance metric between two distributions. Finally, (1) is but only one possible way of comparing the merits of two distance metrics, each supporting its own hypothesis that the query video is a match or a non-match attempt. Due to the generative nature of this technique, many intermediate approximation steps are required, preventing us to directly minimizing the classification error. For this reason, we explored a discriminative solution which aims precisely at minimizing this error criterion. The idea consists of approximating the distribution of a video score $\hat{p}(y)$ for $y \in \mathcal{Y}''$ (which could be a query (test) video score set $\mathcal{Y}$ or the training score set $\mathcal{Y}'_{pq}$ for different claimed identity

p and true identities q) using simple non-parametric statistics such as mean, standard deviation, skewness, the 25-th, 50-th (medean) and 75-th percentiles, as well as minimum and maximum of the scores, and the number of samples:

$$\boldsymbol{\theta}(\mathcal{Y}'') = [\mu, \sigma, \gamma, Q_1, Q_2, Q_3, \min(y), \max(y), N]'. \tag{2}$$

In essence, the above parameters summarize the entire video score set in a very coarse way. While in principal one can use many more points at different percentiles, the number of dimensions can be high, making the problem unnecessarily difficult, i.e., the curse of dimensionality [Bis99].

Having derived the parameters $\boldsymbol{\theta}(\mathcal{Y}'')$, the next step consists of training a classifier using the parameters as input, estimating the posterior probability of being a genuine user, $P(\omega_1|\boldsymbol{\theta}(\mathcal{Y})$. We used logistic regression for this purpose:

$$P(\omega_1|\boldsymbol{\theta}(\mathcal{Y})) = \frac{1}{1 + \exp(-g(\boldsymbol{\theta})} \tag{3}$$

where

$$g(\boldsymbol{\theta}) = \sum_r \theta_r w_r + w_0$$

is a linear combination of the elements, θ_r, of the vector $\boldsymbol{\theta}$, defined by weights $w_r|\forall_r$.

In order to train the logistic regression, we generated a set of positive samples, $\{\boldsymbol{\theta}(\mathcal{Y}_{pp})\}$, where p are the identities of the legitimate users. The negative samples are obtained from $\{\boldsymbol{\theta}(\mathcal{Y}_{pq})|p \neq q\}$.

Because the classifier is linear in the θ space, the complexity of the classifier is directly related to the number of dimensions in this space. This implies that one way to increase complexity of the logistic regression is to increase the dimension of $\boldsymbol{\theta}$. This can be done, for instance, by increasing the number of percentiles describing a video score distribution. A pre-test shows that the set of parameters used in (2) is adequate, and adding more parameters with additional percentile samples did not show any significant improvement nor degradation in performance. For this reason, in the experiments to be reported in Section 2.2, only those parameters are used.

The discussion so far has been limited to a single expert system. In order to combine the scores of two (or more) experts, we shall introduce $\mathcal{Y}^m$ to denote the video score set of modality m. Assuming independence among the modalities m, one can employ the product rule taking each of the m-th posterior probability of (3). So, following the Naive Bayes principal, the final output is computed as:

$$y_{final} = \sum_m \log\left\{\frac{P(\omega_1|\boldsymbol{\theta}(\mathcal{Y}^m))}{1 - P(\omega_1|\boldsymbol{\theta}(\mathcal{Y}^m))}\right\}$$

In order to make the accept/reject decision at the multimodal level, one simply compares y_{final} with a decision threshold. However, since our methodology also works for unimodal biometric systems, the accept/reject decision can be made by comparing $P(\omega_1|\boldsymbol{\theta}(\mathcal{Y}^m))$ with a decision threshold, for each m modality independently.

2.2 Experiments

2.2.1 Expert Systems

In principle, any classifier that process a video frame-by-frame can be used in our framework. For this reason, we shall present only the systems we used here briefly.

The face and speaker verification baseline systems (experts) are Bayesian classifiers whose class-conditional densities are approximated using Gaussian Mixture Models (GMMs) with the Maximum *a posteriori* adaptation [RQD00]. This is a well established state-of-the-art classifier for the speaker verification, but since then, has also been successfully used for the face verification problem. The GMM-based face expert system that we use is reported in [CSB06], with the source codes available at http://torch3vision.idiap.ch.

Another face expert that we used is thoroughly discussed in [Cha08]. This expert processes each frame in a video where a face can be detected. It is worth noting that in the last Multiple Biometric Grand Challenge (MBGC) evaluation organized by NIST, an slightly more advanced version of this classifier was ranked second in the controlled evaluation setting and third in the uncontrolled setting.

The speech expert we used here differs slightly from the standard one [RQD00] in that the speech variability across sessions is removed by factor analysis [MSFB07]. This technique is applied to all training and test data prior to building a (client-specific) GMM-MAP adapted model.

2.2.2 Database

The BANCA database contains recording of face and speech biometric modalities using a camcorder, registering 52 people reading text-prompted sentences as well as answering short questions. The BANCA English subset is used in our experiments.

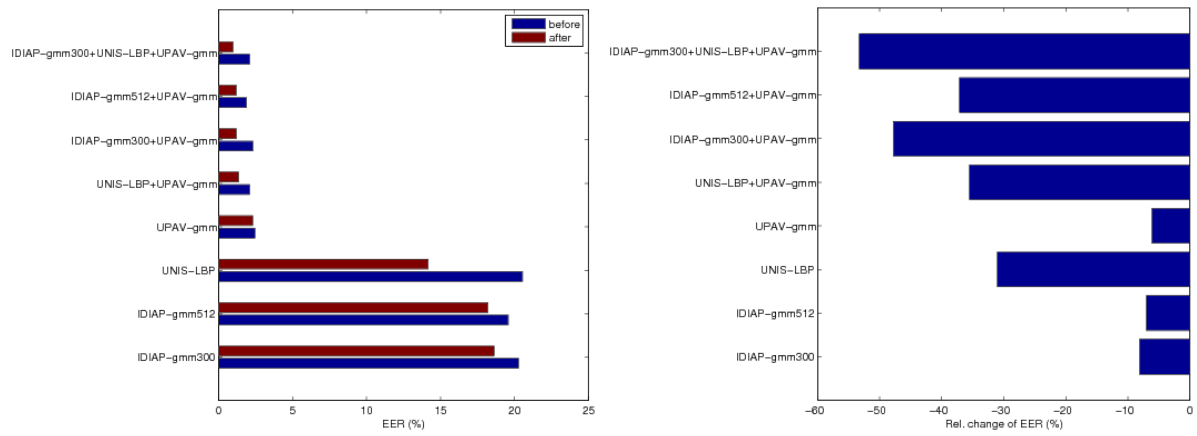
A consequence of this BANCA database setting is that the face verification problem becomes extremely challenging, compared to the speaker verification problem. This is because in both the adverse and degraded conditions, the noise due to the environmental conditions affecting the speech modality, which are all indoor recordings, is still relatively unimportant in comparison with the face modality.

A novel aspect concerning the usage of this database, unlike precedent efforts, is that *video sequences* are actually used here, rather than *still images* extracted from the video sequence.

2.2.3 Results

We conducted our experiments in two stages: (i) unimodal experiments and (ii) multimodal experiments.

In the first set of experiments, we assessed the performance of video-based frame level fusion and compared it with the standard frame-level fusion techniques. For the speech modality the so-called standard fusion is the sum rule [RQD00]. We also confirm that this



(a) EER (b) Rel. change of EER

Figure 4: (a) EER and (b) relative change of EER of unimodal and multimodal systems before and after applying our video-based score-level fusion obtained by logistic regression.

is the best strategy among all the known fixed rules. For the face modality, the max rule turns out to be the best strategy [Cha08].

The objective of the second set of experiments is to assess the extent of improvement possible when the video-based score-level fusion is applied to the underlying expert outputs. As a control, the baseline fusion system is one whose underlying expert outputs are scores obtained on the standard fusion strategy. Therefore, the *multimodal fusion module* is the same (i.e., summing the expert outputs) for both systems.

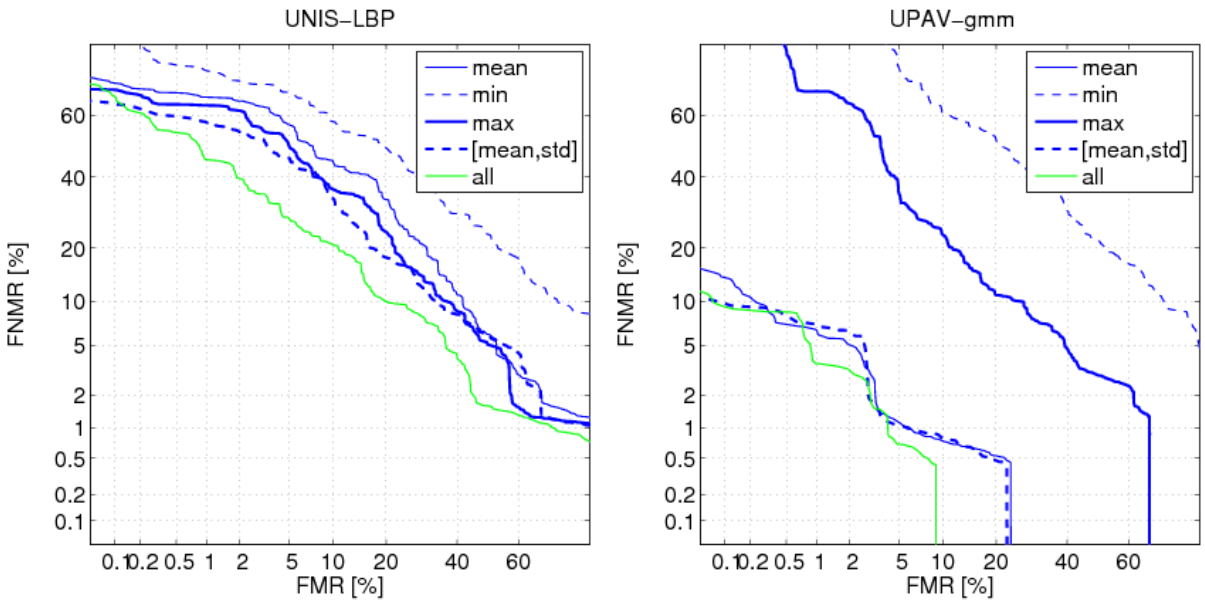
The results of these two sets of experiments are summarized in Figure 4. Figure (a) reports absolute performance for each unimodal and multimodal systems in terms of Equal Error Rate (EER). This is the point at which the probability of a false accept is equal to the probability of a false reject.

Figure (b) reports the *relative change* of EER, which is defined as

$$\text{rel. change of EER} = \frac{\text{EER}_a - \text{EER}_b}{\text{EER}_b}$$

where EER_b is the EER *before* applying the proposed video-based score-level fusion (i.e, the standard fixed rule strategy) where EER_a is the EER *after* applying the proposed technique. Therefore, negative change of EER implies improvement. As can be observed, although the proposed method improves only marginally the unimodal systems, the benefit is greater at the fusion level.

Since EER is not the only point of interest, we also examined the entire DET curve of each system. The DET curves of two unimodal systems are shown in Figure 5. In these figures, we also used a reduced set of features for logistic regression, namely the mean and standard deviation. As can be observed, maximum performance gain is obtained when all the distribution descriptors are used.



(a) LBP-based face expert (b) GMM-based speech expert

Figure 5:

We further examined the weights of logistic regression after training for each of the unimodal systems. They are plotted in Figure 6. Note that the weights shown here are based on the distribution descriptors whose values have been normalized to zero mean and unit variance. Therefore, all the weights are comparable. We observe that the most important distribution descriptor is the standard deviation, followed by the mean statistics. Furthermore, this observation is *consistent* across all the systems we tested.

2.3 Video-based Score-level Fusion - Concluding Remarks

Video-based biometric systems have several advantages over its static image-based counterpart: improved robustness to spoof attack, improved accuracy via variance reduction, and possibility of construction of data of higher resolution/quality. This paper explores a score-level fusion framework in which an existing biometric system can produce a matching score for each valid frame. We proposed score-level fusion strategy that relies on a set of distribution descriptors. The experimental results confirm our conjecture that the abundant scores made available by video-based biometric data can outperform the standard score-level fusion strategy. When applied to a multimodal system, as much as 50% relative gain in performance was observed (i.e., halving the EER). This result is an evidence of the merit of our proposal.

A possible extension of this work is to examine the correlation of scores between the two modalities. However, at present, this is not yet possible since the face and speech scores are not necessarily aligned. Hence, before correlation can be exploited, the alignment problem

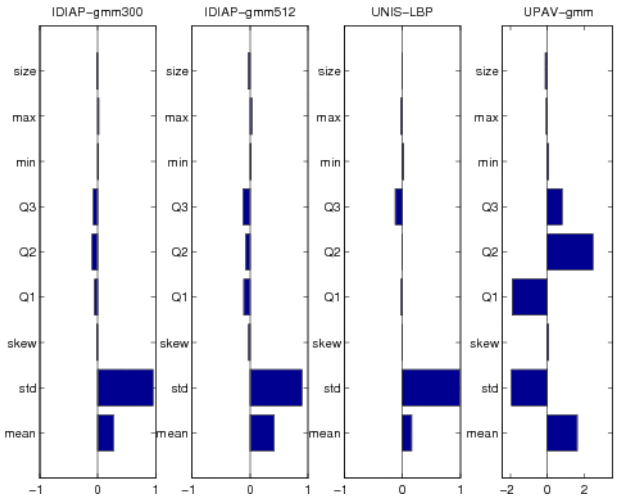


Figure 6: The weights of logistic regression after training.

must be solved. This constitutes a possible future research direction.

3 Feature-level fusion using Audio-Visual Slice Classifiers

Multimodal fusion techniques involve either fusion at the *feature-level* or at the *score-level* [RNJ06][San02]. Biometric systems involving feature-level fusion include the AHMM system proposed by Bengio et al. [Ben03] and feature fusion scheme using face and hand biometrics proposed by Ross et al. [RG05]. In general, feature-level fusion is reported less in the literature compared to score-level fusion, especially for audio-visual biometrics. This is mainly due to the *curse of dimensionality* [Bis99] and its associated computational complexity. However, feature-level fusion has an advantage: it does not assume statistical independence between the modalities as score fusion often does. It has been shown that such an assumption is not always true [RM10] and it could lead to a degradation in performance of score-level fusion systems [NRJ09]. Thus it is important to investigate feature-level fusion systems too, at the same time trying to overcome their inherent problem of dimensionality.

In this part of the work, we first investigate a basic question : would feature-level fusion be indeed helpful in a person identification task? A preliminary study suggests that the answer is yes. Motivated by this, we propose such a system based on a novel concept called "slice", coupled with a boosting framework for classifier selection [FHT98]. Experimental results on a standard audio-visual database validate the robustness of our framework in noisy acoustic environments and show that it compares well with score-level fusion. Furthermore, our framework is computationally efficient, which makes it suitable for embedding in mobile devices which have comparatively limited computational capabilities.

3.1 Can Feature-level Fusion improve the performance of a biometric system? - A Preliminary Study

In this section, we investigate if feature-level fusion can indeed improve the performance of a biometric system. Instead of approaching the problem directly, we propose a related task: person identification in a cross-modal scenario, i.e., matching the speaker in an audio recording to the same speaker in a video (visual-only) recording, where the two recordings have been made during different sessions, using speaker specific information which is common to both the audio and visual modalities [RM10]. We hypothesize that such a related though much more difficult task can be solved by a pattern recognition system with a performance statistically significantly higher than chance *if and only if* crossmodal identity information is indeed available and shared between the audio and visual modalities. Hence, it provides a *sufficient* though not necessary condition for the hypothesis that feature-level fusion might indeed be beneficial for a biometric system. This is because such a fusion system can exploit such crossmodal identity information, if it exists, while a system based on score fusion cannot. In the following paragraphs, we describe in more detail about the crossmodal task.

We often create a mental image of a person whose voice is familiar (from telephone conversations, for example) but whom we have never seen. We often also create a mental "voice model" from visual information (either static or dynamic) of persons we have never heard. Recent studies have investigated these phenomena scientifically [KHLVB03] [LP04] [KFM02], asking human observers to match an audio recording of an unknown voice X to two video recordings of two unknown speakers, A and B, one of which is X, and vice versa. It was found that humans performed in this task with an accuracy significantly above chance. Let us define this crossmodal matching task, termed as the XAB task [KHLVB03], as follows.

The XAB task has two stages : (1) the learning stage and (2) the matching stage. In the learning stage, joint audio and visual information is available in the form of synchronized audio and video (dynamic facial appearance) recordings of persons speaking. The purpose of this stage is to extract or store knowledge required to map speaker identities between audio and visual modalities. In the matching stage, there are two cases, the Audio-to-Visual (a-v) matching task and the Visual-to-Audio (v-a) matching task. In the a-v task, an audio recording of a person X speaking, and two video recordings showing two different persons speaking, A and B, are provided. Given that exactly one out of A and B is X, the task is to decide which one it is. For all the speakers in the matching stage, it is critical that no joint (synchronized) audio and visual information be available. We term this the Audio-Visual Mismatch criterion. This causes the XAB task to be distinct from a simple audio-to-visual synchronization task where both modalities capture the same event in time [Kea09]. To ensure this, the audio and video recordings in the matching stage should be temporally non-overlapping, i.e., they should be made during different sessions, and speakers in the matching stage should be all distinct from speakers in the learning stage. The converse v-a task is exactly the same as the a-v task with the roles of the modalities reversed.

There are several studies with human observers performing the XAB task.[1] Lachs et al. [LP04] and Kamachi et al. [KHLVB03] reported human observers correctly matching X to A or B around 65% of the times. Krauss et al. have shown similar matching performance using static instead of dynamic visual information [KFM02]. Campanella et al. [CB07] provide additional insights on cross-modal information transfer in humans.

3.1.1 Two frameworks for crossmodal speaker matching

We explored a possible solution to the XAB task by creating modality independent speaker models which can be used equally on both audio and visual data. We studied two approaches for this, the K-means clustering (KMC) approach and the K-nearest neighbour (KNN) approach.

Before we explain these two approaches, we briefly describe the audio and visual features we used. For the audio modality, the audio data sampled at 8kHz was blocked into frames equal in duration to the video frames (corresponding to 320 samples per frame) and 16 Mel-Frequency Cepstral Coefficients (MFCC) [PNLM04] were extracted from each block, out of which 1^{st} to 8^{th} were retained to form the audio feature vectors. For each audio sequence, Cepstral Mean Subtraction [PNLM04] was performed. For the visual modality, we concentrated on lip appearance features since they have been shown to be robust and efficient [PNLM04]. The video frame rate was 25fps. From each video frame, a 16×16 Region-Of-Interest (ROI) around the lips was extracted using available annotation, followed by geometric normalization and inter-frame alignment. Next, 2D-DCT features [PNLM04] were extracted and 3^{rd} to 10^{th} highest energy coefficients were retained to form the visual feature vectors.[2] Mean normalization was performed for each visual feature sequence [PNLM04]. It is to be noted that only voiced frames were used, both for audio and visual modalities.

Let $\mathbf{R}^a$ and $\mathbf{R}^v$ denote the audio and visual feature spaces. For the learning stage, synchronized audio and visual data is available. Let $\mathbf{S}^a$ and $\mathbf{S}^v$ denote the sets of audio and visual feature vectors extracted from this data, i.e. $\mathbf{S}^a \subset \mathbf{R}^a$, $\mathbf{S}^v \subset \mathbf{R}^v$. These sets, termed the audio and visual learning sets, are ordered such that the i-th element $\mathbf{x}_i^a \in \mathbf{S}^a$ is synchronous to the i-th element, $\mathbf{x}_i^v \in \mathbf{S}^v$. For the matching stage, let X, A and B also denote the respective recordings as well as the persons X, A and B. Let $\mathbf{S}_{\text{X}}^m, \mathbf{S}_{\text{A}}^m, \mathbf{S}_{\text{B}}^m$ denote the feature vectors extracted from X, A and B, where m can indicate either the audio (a) or the visual (v) modality depending on whether it is an (a-v) or (v-a) task. Let $|\cdot|$ denote the size of a countable set, and $\mathbf{1}_{\mathbf{S}}(\mathbf{x})$, the indicator function of any set $\mathbf{S}$, i.e. $\mathbf{1}_{\mathbf{S}}(\mathbf{x}) = 1$ if $\mathbf{x} \in \mathbf{S}$ and is zero otherwise.

1) K-means Clustering (KMC) Approach

In the learning stage, the learning sets $\mathbf{S}^a$ and $\mathbf{S}^v$ are independently clustered into K clusters, $\{\mathbf{S}_k^a\}_{k=1}^K$ and $\{\mathbf{S}_k^v\}_{k=1}^K$, using K-means algorithm [DHS00] with squared-Euclidean

[1] For humans, the learning stage comprises of all speech-related joint audio-visual stimuli received as part of normal day-to-day activities prior to the experiments.

[2] For both the audio and visual cases, the coefficients retained have been selected by trial-and-error to give best performance.

distance. Let $\{\mathbf{R}_k^a\}_{k=1}^K$ and $\{\mathbf{R}_k^v\}_{k=1}^K$ denote the corresponding Voronoi cells formed by segmenting the spaces $\mathbf{R}^a$ and $\mathbf{R}^v$ according to these clusters, i.e, $\mathbf{S}_k^a \subset \mathbf{R}_k^a$, $\mathbf{S}_k^v \subset \mathbf{R}_k^v$ for $1 \leq k \leq K$. Let $\mathbf{H}^{va}$ denote the $K \times K$ Hebbian projection matrix [Coe06], each of whose elements $\mathbf{H}^{va}(k_a, k_v)$ estimates the probability that an audio vector $\mathbf{x}^a$ belongs to a particular cell $\mathbf{R}_{k_a}^a$ in the audio feature space, given that its synchronous visual vector $\mathbf{x}^v$ belongs to the cell $\mathbf{R}_{k_v}^v$ in the visual feature space, i.e. $\mathbf{H}^{va}(k_a, k_v) = \Pr(\mathbf{x}^a \in \mathbf{R}_{k_a}^a | \mathbf{x}^v \in \mathbf{R}_{k_v}^v)$. It is estimated as

$$\mathbf{H}^{va}(k_a, k_v) = \frac{1}{|\mathbf{S}_{k_v}^v|} \sum_{\mathbf{x}^v \in \mathbf{S}_{k_v}^v} \mathbf{1}_{\mathbf{S}_{k_a}^a}(\mathbf{x}^a) \tag{4}$$

where $1 \leq k_a, k_v \leq K$, $\mathbf{x}^a$ is the audio vector synchronous with visual vector $\mathbf{x}^v$ and $|\cdot|$ denotes the size of a countable set. The inverse Hebbian projection, $\mathbf{H}^{av}$ can be calculated as in Eqn. 4 by interchanging the audio and visual modalities. The matrices $\mathbf{H}^{av}$ and $\mathbf{H}^{va}$ are the outputs of the learning stage.

For the matching stage, let us consider the (a-v) task. Let $\mathbf{p}_{\mathrm{X}}^a$,$\mathbf{p}_{\mathrm{A}}^v$ and $\mathbf{p}_{\mathrm{B}}^v$ be the probability mass functions (PMF) of the feature vectors extracted from X, A and B, i.e. $\mathbf{S}_{\mathrm{X}}^a, \mathbf{S}_{\mathrm{A}}^v$ and $\mathbf{S}_{\mathrm{B}}^v$ respectively, based on the K clusters formed in the learning stage. Thus, $\mathbf{p}_{\mathrm{X}}^a(k) = \Pr(\mathbf{x}^a \in \mathbf{R}_k^a | \mathbf{x}^a \in \mathbf{S}_{\mathrm{X}}^a)$, $\mathbf{p}_{\mathrm{A}}^v(k) = \Pr(\mathbf{x}^v \in \mathbf{R}_k^v | \mathbf{x}^v \in \mathbf{S}_{\mathrm{A}}^v)$ and $\mathbf{p}_{\mathrm{B}}^v(k) = \Pr(\mathbf{x}^v \in \mathbf{R}_k^v | \mathbf{x}^v \in \mathbf{S}_{\mathrm{B}}^v)$. These PMFs are estimated as,

$$\mathbf{p}_{\mathrm{X}}^a(k) = \frac{1}{|\mathbf{S}_{\mathrm{X}}^a|} \sum_{\mathbf{x}^a \in \mathbf{S}_{\mathrm{X}}^a} \mathbf{1}_{\mathbf{R}_k^a}(\mathbf{x}^a) \tag{5}$$

$$\mathbf{p}_{\mathrm{A}}^v(k) = \frac{1}{|\mathbf{S}_{\mathrm{A}}^v|} \sum_{\mathbf{x}^v \in \mathbf{S}_{\mathrm{A}}^v} \mathbf{1}_{\mathbf{R}_k^v}(\mathbf{x}^v) \tag{6}$$

$$\mathbf{p}_{\mathrm{B}}^v(k) = \frac{1}{|\mathbf{S}_{\mathrm{B}}^v|} \sum_{\mathbf{x}^v \in \mathbf{S}_{\mathrm{B}}^v} \mathbf{1}_{\mathbf{R}_k^v}(\mathbf{x}^v) \tag{7}$$

where $1 \leq k \leq K$. Next, we use the Hebbian projection matrix, $\mathbf{H}^{va}$ to project the two PMFs in the visual space, $\mathbf{p}_{\mathrm{A}}^v, \mathbf{p}_{\mathrm{B}}^v$ to the audio space, as follows,

$$\tilde{\mathbf{p}}_{\mathrm{A}}^a = \mathbf{H}^{va} \mathbf{p}_{\mathrm{A}}^v \tag{8}$$

$$\tilde{\mathbf{p}}_{\mathrm{B}}^a = \mathbf{H}^{va} \mathbf{p}_{\mathrm{B}}^v \tag{9}$$

These two PMFs (which we term as pseudo-PMFs) are used to approximate the true PMFs of the unavailable audio feature vectors corresponding to the visual-only recordings A and B [Coe06]. For the matching task, we consider these PMFs as speaker specific models and decide,

$$\mathrm{X} \equiv \begin{cases} \mathrm{A} & \text{if } \rho_{\mathrm{B}}(\mathbf{p}_{\mathrm{X}}^a, \tilde{\mathbf{p}}_{\mathrm{A}}^a) \geq \rho_{\mathrm{B}}(\mathbf{p}_{\mathrm{X}}^a, \tilde{\mathbf{p}}_{\mathrm{B}}^a), \\ \mathrm{B} & \text{if } \rho_{\mathrm{B}}(\mathbf{p}_{\mathrm{X}}^a, \tilde{\mathbf{p}}_{\mathrm{A}}^a) < \rho_{\mathrm{B}}(\mathbf{p}_{\mathrm{X}}^a, \tilde{\mathbf{p}}_{\mathrm{B}}^a) \end{cases} \tag{10}$$

where ρ_B denotes the Bhattacharyya coefficient [DHS00] between two PMFs $\mathbf{p}_1, \mathbf{p}_2$ and is calculated as, $\rho_B(\mathbf{p}_1, \mathbf{p}_2) = \sum_{\forall k} \mathbf{p}_1(k)^{\frac{1}{2}} \mathbf{p}_2(k)^{\frac{1}{2}}$. For the (v-a) task, a similar procedure was followed, interchanging the roles of the audio and visual modalities.

2) K-Nearest Neighbours (KNN) Approach

There is no separate learning stage in this approach. Information in the audio and visual learning sets $\mathbf{S}^a, \mathbf{S}^v$ (ref. Sec.3.1.1) is directly used in the matching stage. For the matching stage, let us again consider the (a-v) task. For each audio vector $\mathbf{x}^a_{\mathrm{X},i} \in \mathbf{S}^a_{\mathrm{X}}$ extracted from X, we form the set $\mathbf{\Psi}_{X,i}$ of the indices of K_a-nearest neighbours [DHS00] of $\mathbf{x}^a_{\mathrm{X},i}$ in $\mathbf{S}^a$, the audio learning set. Similarly, we form sets of indices of K_v-nearest neighbours $\{\mathbf{\Psi}_{\mathrm{A},i}\}, \{\mathbf{\Psi}_{\mathrm{B},i}\}$ for each vector in $\mathbf{S}^v_{\mathrm{A}}, \mathbf{S}^v_{\mathrm{B}}$, the visual vectors extracted from A and B respectively, from $\mathbf{S}^v$, the visual learning set. These nearest neighbour sets are independent of modalities since each element in $\mathbf{S}^v$ has a corresponding element in $\mathbf{S}^a$ (ref. Sec.3.1.1). This forms the basis of the cross-modal mapping in this approach. To match X to A or B, we use the sum of the sizes of intersections $\mathbf{s}_I$ between the nearest neighbour sets of X and those of A,B, as follows,

$$\mathrm{X} \equiv \begin{cases} \mathrm{A} & \text{if } \mathbf{s}_I(\mathrm{X},\mathrm{A}) \geq \mathbf{s}_I(\mathrm{X},\mathrm{B}), \\ \mathrm{B} & \mathbf{s}_I(\mathrm{X},\mathrm{A}) < \mathbf{s}_I(\mathrm{X},\mathrm{B}) \end{cases} \tag{11}$$

where $\mathbf{s}_I(\mathrm{X},\mathrm{A}), \mathbf{s}_I(\mathrm{X},\mathrm{B})$ are defined as follows,

$$\mathbf{s}_I(\mathrm{X},\mathrm{A}) = \frac{1}{|\mathbf{S}^a_{\mathrm{X}}||\mathbf{S}^v_{\mathrm{A}}|} \sum_{\mathbf{x}^a_{\mathrm{X},i} \in \mathbf{S}^a_{\mathrm{X}}} \sum_{\mathbf{x}^v_{\mathrm{A},j} \in \mathbf{S}^v_{\mathrm{A}}} |\mathbf{\Psi}_{\mathrm{X},i} \cap \mathbf{\Psi}_{\mathrm{A},j}| \tag{12}$$

$$\mathbf{s}_I(\mathrm{X},\mathrm{B}) = \frac{1}{|\mathbf{S}^a_{\mathrm{X}}||\mathbf{S}^v_{\mathrm{B}}|} \sum_{\mathbf{x}^a_{\mathrm{X},i} \in \mathbf{S}^a_{\mathrm{X}}} \sum_{\mathbf{x}^v_{\mathrm{B},j} \in \mathbf{S}^v_{\mathrm{B}}} |\mathbf{\Psi}_{\mathrm{X},i} \cap \mathbf{\Psi}_{\mathrm{B},j}| \tag{13}$$

For the (v-a) task, a similar procedure was followed, interchanging the role of the audio and visual modalities. It can be shown that the sums $\mathbf{s}_I(\mathrm{X},\mathrm{A}), \mathbf{s}_I(\mathrm{X},\mathrm{B})$ can be equivalently expressed as approximations to the $\mathbf{L}^2$-inner product of the PMFs corresponding to the audio and visual data. However, compared to Sec.3.1.1, the feature space is now subdivided much more minutely, each vector in the learning sets $\mathbf{S}^a, \mathbf{S}^v$ forming its own cell. This amounts to exploiting maximally the information available for cross-modal matching. Our proposed matching criterion based on comparing the $\mathbf{s}_I$ values is motivated by the use of the $\mathbf{L}^2$ inner product kernel in state-of-the-art speaker verification systems [CSR06].

3.1.2 Experiments

All experiments were performed on the M2VTS audio-visual database [M2V] with 24 male and 10 female speakers. Synchronized audio and visual data was recorded in a controlled environment across multiple sessions separated by one week intervals. Lip annotations were obtained from `http://www.ee.surrey.ac.uk/Projects/M2VTS/experiments/lip_tracking/`. We tested our approach on two conditions : (1) lexically matched and (2)

lexically mismatched. For condition (1), speech content in X, A and B were lexically matched. Recordings from the database were used as it is : in each recording, the speaker counted from '0' to '9' in their native language. For the second (more difficult) condition, the recordings were rearranged so that segments used for X were lexically mismatched with A and B : if X contained '0' to '4', A and B contained '5' to '9' and vice-versa. Of course, the Audio-Visual Mismatch criterion (ref. Sec.3.1) was always maintained in both conditions. X, A and B consisted of around 4.5 seconds of data each. Separate experiments were performed on only male (M), only female (F) and both male and female (F+M) speakers. For each XAB task, two speakers were separated from the complete set, these two were used in the matching stage, while all the remaining speakers were used in the learning stage. For one complete experiment, the XAB task was repeated for all possible pairs of speakers in the matching stage. Considering all possible combinations, the total number of times the XAB task (a-v and v-a each) was independently evaluated is 2208 for the M case, 360 for the F case and 4488 for the F+M case. The match score for each experiment is calculated as,

$$\text{Match score} = \frac{\text{No. of succesful matches}}{\text{Total no. of XAB tasks}} \times 100\% \tag{14}$$

Since each task has two alternatives only one out of which is correct, the expected score for a random classifier would be 50%. Each experiment was repeated for different values of K, the number of clusters, and K_a, K_v, the number of nearest neighbours, for the KMC and KNN approaches respectively. Optimal value of K was 64, while for K_a, K_v it varied from 2 to 256 according to the conditions tested. Table 1 gives the results of our experiments in terms of the match scores obtained, using the optimal parameter values.

3.1.3 Discussions and conclusions of the preliminary study

For the lexically matched case, both the KMC and KNN approaches give match scores around 65%. This is statistically significant, given the total number of times the XAB task was evaluated (ref. Sec.3.1.2). For the lexically mismatched case, the performance of KNN drops by 10% but is still significant; KMC is unable to perform at more than chance level. This shows the relative robustness of the KNN approach. Our method compares well with results reported by studies using human observers on the XAB task [KHLVB03] [LP04] as shown in Table 2, although it is to be noted that these studies used different databases. It is to be noted that human performance fell drastically for time-reversed stimuli [KHLVB03] [LP04]; our method is unaffected by this, being based on static feature vectors only. Furthermore, human observers had information from the entire face available to them, while our method uses information exclusively from the lip region.

Since both the two approaches have shown performance signicantly better than chance, this satisfies the sufficient condition for our hypothesis about feature-level fusion to be true. *Now we have reasonable justification to implement feature-level fusion in multimodal biometric systems with an aim to improve verification performance.* This will be explored

Proposed Approach	XAB task type		Lex. matched	Lex. mismatched
KMC	a-v	M	66.6	*
		F	79.4	*
		F+M	66.4	*
	v-a	M	65.1	*
		F	60.0	*
		F+M	64.9	*
KNN	a-v	M	68.9	56.0
		F	64.2	57.8
		F+M	66.4	56.6
	v-a	M	66.0	55.6
		F	61.9	60.6
		F+M	63.4	56.1

Table 1: Match scores (%) for the XAB task using the proposed approaches. An asterisk (*) denotes that a match score better than random chance (50%) could not be obtained.

	XAB task type	Lex. matched	Lex. mismatched
Kamachi et al. [KHLVB03]	a-v	69.0	59.0
	v-a	66.0	60.0
Lachs et al. [LP04]	a-v	60.7	n.a.
	v-a	65.1	n.a.

Table 2: Match scores (%) for the XAB task performed by human observers.

in the subsequent sections.[3]

3.2 Boosted Slice Classifiers (BSC) - the Proposed Framework

Motivated by the positive results from our preliminary study reported above, we propose in this section a novel framework for feature-level fusion of audio and visual modalities.

We first assume that the raw audio and visual streams have been synchronized and

[3]As a side note, this method for crossmodal matching could be developed further using this preliminary study as a basis, and the match scores could be improved so that it can be used in practical applications, such as (1) a cross-modal surveillance scenario where prior speech data (but no visual data, for example via telephone conversations) about a person X has been collected and presently it is required to identify this person out of a group which is under video surveillance (but no audio data is currently available, for example due to distance from group or noisy environment), and (2) a multimodal biometric system which uses cross-modalities (a-v, v-a) to augment the normal audio and visual modalities and make it more reliable. However, this is beyond the scope of the current work.

processed to give a sequence of audio and visual feature vectors. It is to be noted that the audio stream must be framed at a rate equal to the visual stream, so as to give audio and visual feature vectors with a one-to-one synchrony. A detailed description of the different audio and visual feature spaces investigated is given in Sec.3.3.2. For now, let us denote the audio and visual feature spaces by $\mathbf{R}^a$ and $\mathbf{R}^v$ respectively. Let N_a and N_v be the sizes of $\mathbf{R}^a$ and $\mathbf{R}^v$ respectively.

3.2.1 The concept of Slice

The audio and visual feature spaces $\mathbf{R}^a$, $\mathbf{R}^v$ can be combined to form the joint audio-visual feature space, $\mathbf{R}^{av} = \mathbf{R}^a \times \mathbf{R}^v$ of size $N_{av} = N_a + N_v$. Assuming synchronous extraction of audio and visual features, an audio-visual event can be represented as a set of points in $\mathbf{R}^{av}$. Due to the problem of dimensionality, modelling these points in the high-dimensional $\mathbf{R}^{av}$ space is difficult [Bis99, Sec.8.6]. To solve this issue, let us define a slice $\mathbf{L}$ as a two-dimensional subspace of $\mathbf{R}^{av}$. It is necessary for feature-level bimodal fusion that $\mathbf{L}$ has at least one audio component $\mathbf{L}^a$ extracted from $\mathbf{R}^a$ and at least one visual component $\mathbf{L}^v$ extracted from $\mathbf{R}^v$. Since there are N_a different audio components in $\mathbf{R}^a$ and N_v different visual components in $\mathbf{R}^v$, the total number of all possible slices are $N_{\mathbf{L}} = N_a \times N_v$. Let $\Lambda = \{\mathbf{L}_i\}_{i=1}^{N_{\mathbf{L}}}$ denote the complete set of all possible slices.

3.2.2 Slice Classifiers

Each slice $\mathbf{L}_i \in \Lambda$ is associated with a slice classifier h_i. Given a classification task (for example, audio-visual speaker authentication) in the high-dimensional audio-visual space $\mathbf{R}^{av}$, the classifier h_i is trained and tested on projections of data exclusively on $\mathbf{L}_i$. Let $H = \{h_i\}_{i=1}^{N_{\mathbf{L}}}$ denote the complete set of all slice classifiers. In this work, we have selected the classifier to be a quadratic discriminant function [DHS00], assuming a normal distribution of data. Although other classifiers are possible, experiments have shown that it serves its purpose sufficiently well, without being too complex at the same time. In practice, the normal assumption is rarely valid; however, this does not affect the performance of the system.

For a speaker authentication task, with client and impostor classes denoted by '1' and '0' respectively, a slice classifier can be expressed as a function $h_i : \mathbf{L}_i \rightarrow \{0, 1\}$. Given a point $\mathbf{x} \in \mathbf{L}_i$,

$$h_i(\mathbf{x}) = \begin{cases} 1 & \text{if} - (\mathbf{x} - \mu_{1,i})^T \Sigma_{1,i}^{-1} (\mathbf{x} - \mu_{1,i}) + (\mathbf{x} - \mu_{0,i})^T \Sigma_{0,i}^{-1} (\mathbf{x} - \mu_{0,i}) \geq \theta_i \\ 0 & \text{otherwise.} \end{cases} \tag{15}$$

where $\mu_{1,i}, \mu_{0,i}$ are the estimated means of classes '1' and '0' projected on $\mathbf{L}_i$, and $\Sigma_{1,i}, \Sigma_{0,i}$ are their estimated covariance matrices. The threshold θ_i is chosen to minimize misclassification error on the training set.

It is to be noted that a single slice classifier by itself is unlikely to perform sufficiently well in its task. However, it is hypothesized that there will be at least some optimal slice classifiers which will perform comparatively better than others. Such optimal classifiers

will be associated with slices which contain maximally discriminative joint audio-visual information. Such optimal classifiers could be combined in a suitable way to obtain a final classifier which is strong enough to perform the task sufficiently well.

3.2.3 Slice Classifier Selection and Combination by Boosting

Out of the complete set of slice classifiers H, a certain number of classifiers are iteratively selected *for each client* according to their discriminative ability with respect to that client. This selection is based on the Discrete Adaboost algorithm [FHT98] with weighted sampling, which is widely used for selection tasks [Rod06] and is known for its robust performance [FHT98]. The algorithm, which is to be run once for each client, is as follows:

Algorithm: Slice Classifier Selection by Discrete Adaboost

Inputs: N_{tr} training vectors $\{\mathbf{x}_j\}_{j=1}^{N_{tr}}$, corresponding class labels, $y_j \in \{0,1\}$ (0:*impostor*, 1:*client*), N_h, the number of classifiers to be selected, N_{tr}^*, the number of training vectors to be randomly sampled at each iteration ($N_{tr}^* < N_{tr}$).

- Initialize the weights $\{w_{1,j}\} \leftarrow \frac{1}{2N_{tr}^{(0)}}, \frac{1}{2N_{tr}^{(1)}}$ for $y_j = 0, 1$ respectively, where $N_{tr}^{(0)}$ and $N_{tr}^{(1)}$ are the number of impostor and client training vectors respectively.
- Repeat for $n = 1, 2, \cdots N_h$:
  - Normalize weights, $w_{n,j} \leftarrow \frac{w_{n,j}}{\sum_{j'=1}^{N_{tr}} w_{n,j'}}$
  - Randomly sample N_{tr}^* training vectors, according to the distribution $\{w_{n,j}\}$
  - For each h_i in H, choose θ_i to minimize misclassification error, $\epsilon_i = \frac{1}{N_{tr}^*} \sum_{j=1}^{N_{tr}^*} \mathbf{1}_{\{h_i(\mathbf{x}_j^{(i)}) \neq y_j\}}$ over the sampled set.
  - Select the next best classifier, $h_n^* = h_{i^*}$ where $i^* = \arg\min_i \epsilon_i$
  - Set $\beta_n \leftarrow \frac{\epsilon_{i^*}}{1-\epsilon_{i^*}}$
  - Update the weights, $w_{n+1,j} \leftarrow w_{n,j} \beta_n^{\mathbf{1}_{\{h_n^*(\mathbf{x}_j^{(n)}) = y_j\}}}$

Output: The sequence of selected best slice classifiers, $\{h_n^*\}_{n=1}^{N_h}$.

For the database and framing parameters used, N_{tr} was around 8,000, and $N_{tr}^{(1)}$, which varies for each client, was around 150. N_{tr}^* was set to 1000 and N_h to 30. For each client, the selected slice classifiers are combined linearly to give a strong classifier F [FHT98]:

$$\mathbf{h}(\mathbf{x}) = \sum_{n=1}^{N_h} \alpha_n h_n(\mathbf{x}). \tag{16}$$

The weights $\{\alpha_n\}$ are calculated to minimize the exponential loss [FHT98] and normalized to sum to unity for each client, $\alpha_n = \frac{log(\beta_n)}{\sum_{n'=1}^{N_h} log(\beta_{n'})}$. Since a decision is only required at

the utterance level and not at the frame level, the responses $\mathbf{h}(\mathbf{x})$ of each frame $\mathbf{x}$ in an utterance are added and normalized by the number of frames, to obtain the final score S for the utterance. This is compared with a preset threshold to decide if the utterance was made by a client or an impostor. This preset threshold Θ is calculated by minimizing the Equal Error Rate [Bim04] on a separate development set.

In brief, this section has outlined a solution to the problem of dimensionality in multimodal fusion by proposing the concept of "slice". Furthermore, it describes how the boosting framework can be used to select only those pairs of audio and visual features which are maximally discriminative for a particular client. In the next section, we describe several experiments evaluating the proposed framework as well as several reference systems.

3.3 Experiments

3.3.1 Database and Protocol

As in Sec.3.1.2, all experiments in this section were performed on the M2VTS database [M2V] using lip annotations from `http://www.ee.surrey.ac.uk/Projects/M2VTS/experiments/lip_tracking/`. We followed the speaker verification protocol for this database as outlined in [Ben03].[4] This protocol involves a 4-fold cross-validation procedure described as follows.

The clients were firstly divided into 4 disjoint sets, with 8 clients in each set. For each fold, one particular set out of the four was set as the *evaluation set*, while the remaining 3 sets formed the *development set* for that fold. The experiment was conducted in three phases: *training, development and evaluation*, repeated individually for each fold.

In the *training* phase, only the first 2 recordings of each client were used to create client-specific models. For all systems using the Universal Background Model-Gaussian Mixture model (UBM-GMM) framework (ref. Sec.3.3.2), all clients in the development set of a particular fold were used to train the *world model* for that fold. Next, this world model was adapted for *all* clients in the database to give client-specific models for that fold.

For systems using the proposed Boosted Slice Classifier framework, there were two cases. For each client in the evaluation set of a particular fold, all clients in the corresponding development set contributed to the negative samples for the boosting process of that client, while for each client in the development set of that fold, all clients in the development set *other than this one* contributed to the negative samples. Thus, for all clients in the evaluation set, no data from another client also in the same evaluation set was used as negative samples while training its model.

In the *development* phase, speaker verification is performed on the development set of each fold using the third and fourth recordings of each client. For each fold, system parameters (for e.g., the number of classifiers N_h to be boosted for the Slice Classifier framework and the decision threshold Θ) are optimized based on their performance on this task. No data from the evaluation set is used.

[4] A few recordings had to be discarded due to absence of lip annotation for those recordings. This caused a minor modification in the protocol.

In the *evaluation* phase, speaker verification is performed on the evaluation set of each fold using the third and fourth recordings of each client and using the optimal parameter values obtained from the development phase for that fold. The verification performance (in terms of the mean HTER %) is averaged over all 4 folds and reported. Since no parameters were calculated using the evaluation data, this can be considered an unbiased estimate of the system performance in a real scenario [Ben03].

Furthermore, for the *evaluation* phase, two different conditions were evaluated, a) *Matched-clean*: The original clean data (third and fourth recordings of the evaluation phase) was used as it is. This represents a controlled scenario, where the evaluation data and the training and development data are matched. b) *Mismatched-noisy*: In this condition, two types of noise, namely, white noise and babble noise, from the standard Noisex-92 database [VSTJ92] were added at 3 different SNR levels (10dB, 5dB and 0dB) to the original clean speech of the third and fourth recording before testing. This represents a more difficult realistic scenario where the evaluation data is noisy and hence mismatched with the training and development data [Ben03]. We report results for both these conditions.

3.3.2 Systems implemented

Two groups of speaker verification systems were implemented. The first group involves the Boosted Slice Classifier framework described in this work. The second group includes certain reference systems which are conventionally used for audio-visual speaker verification with score fusion. The performance of the two groups are compared. **1) Boosted Slice Classifier (BSC) Systems** These systems used the Boosted Slice Classifier framework as described in Sec.3.2 for the task of audio-visual speaker verification. Boosted Slice classifiers are associated with slices derived from an audio visual feature space pair. To form this pair, different audio and visual feature spaces were investigated as described below. For each feature space, its code name (by which it is indicated in subsequent sections) is provided in parentheses. The number in the code is the dimensionality of the space.

Audio feature spaces Apart from the conventional cepstral representation of speech using 16 Mel Frequency Cepstral Coefficients (MFCC) [Bim04] (MC16), we also investigated spectral representations which have shown promising performance in a similar boosting framework for speaker verification [RMDM10]. In particular, Mel spectra calculated using 24, 32 and 40 Mel filters (MS24, MS32 and MS40) and Fourier spectra calculated using 256-point and 128-point Discrete Fourier Transform (FS128, FS64) were investigated. It is to be noted that the magnitude spectra were used, hence only one half of the spectrum was retained, since they are symmetric.

Visual feature spaces Firstly, a Region-of-Interest (ROI) around the lips was extracted using available annotation. Next, either a 2D-DCT was performed on it and the 15 highest energy coefficients were retained to form the features (DCT15) [PNLM04] or the gray-scale values in the extracted region were directly used as features. For the latter case, two ROI sizes were considered, a 16×16 ROI and an 8×8 ROI (GS256 and GS64 respectively). **2) Reference Systems** The following reference systems were implemented:

Audio modality A standard speaker verification system [Bim04] using 16 MFCC, 16 Δ-MFCC and Δ-energy modelled by the UBM-GMM framework was implemented. Cepstral Mean Subtraction was performed and silence removal was by a bi-Gaussian [Bim04]. We refer this system as MC-GMM in subsequent sections.

Visual modality A standard face verification system using block-based features modelled by the UBM-GMM framework [CSM03, LC04] was implemented. From each block, 18 DCTmod2 features [SP02] were extracted. For a detailed description of the system, please refer to Appendix A of Deliverable 4.2. We refer to this system as F-GMM in subsequent sections.

Audio-visual score fusion Score fusion using the Normalization-based approach (ref. Sec. 2.3.2, Deliverable 4.2) was implemented. In this approach, audio and visual modalities are assumed to be independent and hence the Naive-Bayes principle can be used. The fusion score S_{fusion} is calculated as a simple sum of the scores from each modality as follows,

$$S_{fusion} = \sum_{i=1}^{M} s_i$$

where $\{s_i\}_{i=1}^{M}$ denote the individual log-likelihood scores calculated from each modality. Here, the number of modalities, $M = 2$. It is to be noted that the two modalities are modelled by UBM-GMM systems which directly give log-likelihood scores as their output.

3.3.3 Results

In Tables 3-9, we show the verification performance of the Boosted Slice Classifier framework, using different combinations of audio-visual space pairs. In Table 3, we show the Matched-clean condition (ref. Sec. 3.3.1). In Tables 4-9, we show the 6 different cases for the Mismatched-noisy conditions (2 noise types $\times$ 3 SNR levels). Finally, in Table 10, we compare the performance of the reference systems with the some of the consistently better performing Boosted Slice Classifier systems.

		Audio feature sets					
		MS40	MS32	MS24	FS128	FS64	MC16
Visual feature sets	GS256	**6.5848**	9.1518	8.2589	**6.5848**	8.2589	10.0446
	GS64	9.2634	**6.5848**	12.9464	8.7054	**5.9152**	13.7277
	DCT15	**6.4732**	10.2679	11.8304	8.1473	8.9286	14.0625

Table 3: Verification performance (HTER %) of the Boosted Slice Classifier system using various combinations of audio and visual feature sets, under **Matched-clean condition**. Lowest HTERs are marked in bold.

		Audio feature sets					
		MS40	MS32	MS24	FS128	FS64	MC16
Visual feature sets	GS256	8.8170	8.7054	**8.2589**	**8.5938**	10.1562	9.3750
	GS64	10.0446	9.8214	12.3884	9.7098	**8.3705**	13.3929
	DCT15	14.8438	14.0625	14.3973	16.6295	19.5312	12.0536

Table 4: Verification performance (HTER %) of the Boosted Slice Classifier system using various combinations of audio and visual feature sets, under **Mismatched-noisy condition**. Noise type: **white noise**, SNR: **10dB**. Lowest HTERs are marked in bold.

		Audio feature sets					
		MS40	MS32	MS24	FS128	FS64	MC16
Visual feature sets	GS256	**8.9286**	11.1607	**8.1473**	10.3795	10.7143	9.2634
	GS64	10.7143	10.4911	13.5045	12.0536	**9.7098**	13.3929
	DCT15	25.5580	21.0938	26.5625	23.5491	25.4464	12.6116

Table 5: Verification performance (HTER %) of the Boosted Slice Classifier system using various combinations of audio and visual feature sets, under **Mismatched-noisy condition**. Noise type: **white noise**, SNR: **5dB**. Lowest HTERs are marked in bold.

3.4 Discussions

3.4.1 Speaker Verification Performance

Firstly, we discuss the performance of only the Boosted Slice Classifier systems, the feature-level fusion framework proposed in this work. From Tables 3-9, it is evident that several pairs out of the 18 audio-visual feature space pairs investigated have performed well on the speaker verification task. Apart from reasonable performance in the Matched-clean condition, the systems have shown significant robustness to the two types of noise at medium to high noise levels in the Mismatched-noisy condition. This is a significant advantage of the proposed framework. As in some recently proposed speaker (audio-only) verification systems using a similar framework to boost classifiers each involving only a small part of the entire feature space [RMDM10], this noise robustness may be due to the fact that the noise might be affecting some of the slices but not *all* the slices *at the same time*. Since the effect on one slice is restricted only to that slice, the final output (linear sum of the slice classifier outputs) is affected less than for a conventional UBM-GMM based system in a similar noisy scenario.

Among the BSC systems, it is to be noted that systems GS64-FS64 and GS256-MS24

		Audio feature sets					
		MS40	MS32	MS24	FS128	FS64	MC16
Visual feature sets	GS256	11.9420	16.6295	**11.8304**	12.0536	13.2812	**8.9286**
	GS64	13.2812	15.9598	16.6295	12.7232	**10.4911**	13.0580
	DCT15	28.5714	29.5759	29.0179	34.3750	30.0223	18.5268

Table 6: Verification performance (HTER %) of the Boosted Slice Classifier system using various combinations of audio and visual feature sets, under **Mismatched-noisy condition**. Noise type: **white noise**, SNR: **0dB**. Lowest HTERs are marked in bold.

		Audio feature sets					
		MS40	MS32	MS24	FS128	FS64	MC16
Visual feature sets	GS256	**7.9241**	9.3750	**7.0312**	8.7054	10.7143	10.7143
	GS64	9.7098	**7.5893**	14.5089	10.2679	8.3705	12.8348
	DCT15	15.9598	14.1741	14.3973	15.9598	19.0848	12.2768

Table 7: Verification performance (HTER %) of the Boosted Slice Classifier system using various combinations of audio and visual feature sets, under **Mismatched-noisy condition**. Noise type: **babble noise**, SNR: **10dB**. Lowest HTERs are marked in bold.

are two of the consistently better performing combinations (ref. Tables 3-9) although others perform almost as well. The first system, GS64-FS64, uses the 8×8 lip-ROI grayscale values as its visual features and the Fourier Spectra (with 128-point DFT) as its audio features. The second system, GS256-MS24, uses the 16×16 lip-ROI grayscale values as its visual features and the Mel Spectra (with 24 Mel filters) as its audio features.

Next, we discuss the comparison of performance of the Boosted Slice Classifier systems with the reference systems (ref. Table 10). For the Matched-clean condition, it is evident that the score fusion of the reference audio and visual systems (MC-GMM and F-GMM) have performed the best compared to the Boosted Slice Classifier systems. However, for the more realistic Mismatched-noisy condition, the proposed Boosted Slice Classifier system have outperformed the reference score fusion system significantly in most of the cases, for different noise types and noise levels. This shows that the proposed feature-level fusion framework does have a significant benefit in such scenarios and it is a promising approach to multimodal biometric systems, comparable to score-level fusion approaches.[5]

[5]For completeness, we also implemented another feature-level fusion approach that is completely different from ours, the Asynchronous Hidden Markov Model (AHMM) [Ben03]. It is one of the very rare approaches proposed for feature-level fusion. Our experimental results have shown that the proposed

		Audio feature sets					
		MS40	MS32	MS24	FS128	FS64	MC16
Visual feature sets	GS256	**9.4866**	10.9375	**7.9241**	11.8304	**9.4866**	**9.4866**
	GS64	10.2679	11.7188	15.9598	12.3884	10.3795	12.8348
	DCT15	28.0134	26.1161	25.5580	25.4464	29.2411	16.9643

Table 8: Verification performance (HTER %) of the Boosted Slice Classifier system using various combinations of audio and visual feature sets, under **Mismatched-noisy condition**. Noise type: **babble noise**, SNR: **5dB**. Lowest HTERs are marked in bold.

		Audio feature sets					
		MS40	MS32	MS24	FS128	FS64	MC16
Visual feature sets	GS256	**14.3973**	17.5223	14.8438	16.8527	14.7321	**11.9420**
	GS64	17.0759	17.5223	18.4152	16.7411	17.9688	**14.0625**
	DCT15	37.1652	32.3661	36.1607	38.9509	33.1473	28.9062

Table 9: Verification performance (HTER %) of the Boosted Slice Classifier system using various combinations of audio and visual feature sets, under **Mismatched-noisy condition**. Noise type: **babble noise**, SNR: **0dB**. Lowest HTERs are marked in bold.

It is to be noted that score fusion performance could be improved by using more sophisticated techniques [SP04]. However, this is beyond the scope of the current work. Furthermore, such approaches will increase the computational complexity of the system.

3.4.2 Computational Complexity

In addition to showing a reasonably robust verification performance, the Boosted Slice Classifier systems are computational much faster than the conventional systems. This is partly due to the simple nature of the individual slice classifiers which are implemented as quadratic discriminant classifiers in 2-dimensional space. Restricting the slices to only 2 dimensions solves the "curse of dimensionality" problem. Each slice classifier can be evaluated using very few floating point operations. Furthermore, the mean number of boosted features N_h as selected in the development stage (ref. Sec. 3.2.3) was found to vary between 10 to 20; hence, the final strong classifier can be evaluated as a simple linear sum of small number of boosted slice classifier outputs.

Boosted Slice Classifier framework compares well with the AHMM system in all the experimental conditions.

		Matched clean	Mismatched-noisy					
			white noise			babble noise		
			10dB	5dB	0dB	10dB	5dB	0dB
Reference systems	MC-GMM (audio)	4.13	31.92	39.73	45.76	16.63	42.97	46.88
	F-GMM (visual)	5.24	5.24	5.24	5.24	5.24	5.24	5.24
	MC-GMM + F-GMM (score fusion)	2.79	8.26	15.18	28.13	2.57	10.16	25.01
Boosted Slice Classifier systems	GS64-FS64	5.92	8.37	9.71	10.49	8.37	10.38	17.97
	GS256-MS24	8.26	8.26	8.15	11.83	7.03	7.92	14.84

Table 10: Comparison of verification performance (HTER %) of the Boosted Slice Classifier system using the consistently better performing combinations of audio and visual feature sets with the reference systems under various conditions.

In comparison, both the audio and visual reference systems (MC-GMM and F-GMM) use the UBM-GMM framework. Evaluating each individual Gaussian involves many more floating point operations than a single slice classifier, since they are calculated on the full audio (33-dimensional) or visual (18-dimensional) feature space. These include complex mathematical operations like exponentiation and logarithm extraction, neither of which are required by the BSC systems. Furthermore, the GMMs for the audio system use 32 Gaussians, while those for the visual system use 256 Gaussians, leading to many more floating point operations in total than the corresponding strong classifier of the proposed system.

3.5 Feature-level Fusion using Boosted Slice Classifiers - Concluding Remarks

In this part of the deliverable, we proposed an advanced fusion strategy involving feature-level fusion of audio and visual modalities for the task of bimodal person verification. Firstly, we reported a preliminary study which investigates if feature-level fusion can indeed be useful for biometric systems. This study has shown positive results.

Based on these results, we proposed a feature combination technique called “slice” and used this in a boosting framework to create a fast and reasonably reliable bimodal verification system. This system has shown robustness under mismatched conditions involving two kinds of noise at medium to high SNRs in the audio modality.

Our experiments suggest that feature-level fusion approaches have promise compared to conventional score fusion and should be investigated further. One direction is to include dynamic information from the audio and visual frames in addition to the static feature vectors used in this work. Furthermore, the dimensionality of the slices could be increased

to more than two. This could extract more joint audio-visual information about a person, although at the cost of increased computational complexity. [6]

4 Final conclusion

In this deliverable, we investigated advanced fusion schemes for audio-visual person authentication from two distinct perspectives: 1) video-based score-level fusion and 2) feature-level fusion.

In the first part, we proposed a score-level fusion strategy that relies on a set of score distribution descriptors extracted from video. Experiments conducted on a standard database showed that improved results can be obtained by exploiting the abundant score information made available by video-based biometric data, for both unimodal systems and bimodal systems.

In the second part, we proposed a feature-level fusion strategy using a novel concept called audio-visual slice, which performed reasonably well on both matched (clean) and mismatched (noisy) data, and is computationally efficient at the same time.

Acknowledgments

The authors thank Dr.Chris McCool for the face expert and Dr.Driss Matrouf for the speech expert in Sec.2.

References

[AS09] A. Abhyankar and S. Schuckers. Integrating a wavelet based perspiration liveness check with fingerprint recognition. *Pattern Recogn.*, 42(3):452–464, 2009.

[Ben03] S. Bengio. Multimodal Authentication using Asynchronous HMMs. In *Proc. of 4th Intl. Conf. on Audio- and Video- based Biometric Person Authentication*, volume 4. Springer, 2003.

[Bim04] F. Bimbot et al. A Tutorial on Text-Independent Speaker Verification. *EURASIP Journal on Applied Signal Processing*, (4):431–451, 2004.

[Bis99] C. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1999.

[BMWC06] H. Bredin, A. Miguel, I.H. Witten, and G. Chollet. Detecting replay attacks in audiovisual identity verification. In *Acoustics, Speech and Signal Processing,*

[6] As a side note, it would also be interesting to investigate the individual "slices" selected by the boosting procedure for each client, in order to analyze the joint person-specific information contained in them.

2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on, volume 1, pages I–I, May 2006.

[CB07] S. Campanella and P. Bellin. Integrating face and voice in person perception. *Trends in Cognitive Science*, 11(12), 2007.

[Cha08] C. Chan. *Multi-scale Local Binary Pattern Histogram for Face Recognition.* PhD thesis, University of Surrey, 2008.

[Coe06] M.H. Coen. Multimodal Dynamics : Self-Supervised Learning in Perceptual and Motor Systems. Phd thesis, Massachusetts Institute of Technology, 2006.

[CS02] Sung-Hyuk Cha and S. N. Srihari. On measuring the distance between histograms. *Pattern Recognition*, 35(6):1355 – 1370, 2002.

[CSB06] F. Cardinaux, C. Sanderson, and S. Bengio. User Authentication via Adapted Statistical Models of Face Images. *IEEE Trans. on Signal Processing*, 54(1):361–373, January 2006.

[CSM03] F. Cardinaux, C. Sanderson, and S. Marcel. Comparison of MLP and GMM classifiers for face verification on XM2VTS. In *Proc. International Conference on Audio- and Video-based Biometric Person Authentication*, pages 1058–1059, 2003.

[CSR06] W.M. Campbell, D.E. Sturim, and D.A. Reynolds. Support Vector Machines using GMM Supervectors for Speaker Verification. *IEEE Signal Processing Letters*, 13(5), 2006.

[DHS00] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification.* John Wiley and Sons, 2000.

[FHT98] J. Friedman, T. Hastie, and R. Tibshirani. Additive Logistic Regression: a Statistical View of Boosting. *Annals of Statistics*, 28:2000, 1998.

[Kea09] K. Kumar and et al. Audio-Visual Speech Synchronization Detection Using a Bimodal Linear Prediction Model. In *CVPR*, 2009.

[KFM02] R.M. Krauss, R. Freyberg, and E. Morsella. Inferring speakers' physical attributes from their voices. *Journal of Experimental Social Psychology*, 38:618–625, 2002.

[KHLVB03] M. Kamachi, H. Hill, K. Lander, and E. Vatikiotis-Bateson. 'Putting the Face to the Voice': Matching Identity across Modality. *Current Biology*, 13:1709–1714, 2003.

[KKC07] Tae-Kyun Kim, J. Kittler, and R. Cipolla. Discriminative learning and recognition of image set classes using canonical correlations. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(6):1005–1018, June 2007.

[LC04] S. Lucey and T. Chen. A GMM parts based face representation for improved verification through relevance adaptation. In *Proc. International Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 855–861, 2004.

[LP04] L. Lachs and P.B. Pisoni. Crossmodal source identification in speech perception. *Ecological Psychology*, 16(3):159–187, 2004.

[M2V] M2VTS Multimodal Face Database, Release 1.00. `http://www.tele.ucl.ac.be/PROJECTS/M2VTS/m2fdb.html`.

[MSFB07] D. Matrouf, N. Scheffer, B. Fauve, and J-F. Bonastre. A straightforward and efficient implementation of the factor analysis model for speaker verification. In *INTERSPEECH Conference, Antewerp, Belgium*, 2007.

[NRJ09] K. Nandakumar, A. Ross, and A.K. Jain. Biometric fusion: Does modelling correlation really matter? In *Proc. BTAS*, 2009.

[PB03] N. Poh and S. Bengio. Non-Linear Variance Reduction Techniques in Biometric Authentication. In *Workshop on Multimodal User Authentication (MMUA 2003)*, pages 123–130, Santa Barbara, 2003.

[PB05] N. Poh and S. Bengio. How Do Correlation and Variance of Base Classifiers Affect Fusion in Biometric Authentication Tasks? *IEEE Trans. Signal Processing*, 53(11):4384–4396, 2005.

[PNLM04] Gerasimos Potamianos, Chalapathy Neti, Juergen Luettin, and Iain Matthews. Audio-visual automatic speech recognition: An overview. In *Issues in Visual and Audio-visual Speech Processing*. MIT Press, 2004.

[RG05] A. Ross and R. Govindarajan. Feature level fusion using hand and face biometrics. In *Proc. SPIE Conf. on Biometric Technologies for Human Identification II*, 2005.

[RM10] A. Roy and S. Marcel. Crossmodal matching of speakers using lip and voice features in temporally non-overlapping audio and video streams. In *20th International Conference on Pattern Recognition*, 2010.

[RMDM10] A. Roy, M. Magimai-Doss, and S. Marcel. Boosted binary features for noise-robust speaker verification. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010.

[RNJ06] A. Ross, K. Nandakumar, and A.K. Jain. *Handbook of Multibiometrics*. Springer Verlag, 2006.

[Rod06] Y. Rodriguez. Face Detection and Verification using Local Binary Patterns. PhD Thesis 3681, Ecole Polytechnique Federale de Lausanne, 2006.

[RP09] A. Ross and N. Poh. *Fusion in Biometrics: An Overview of Multibiometric Systems*, chapter 8, pages 273–292. Springer London, 2009.

[RQD00] D. A. Reynolds, T. Quatieri, and R. Dunn. Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing*, 10(1–3):19–41, 2000.

[San02] C. Sanderson. *Automatic Person Verification Using Speech and Face Information*. PhD thesis, Griffith University, Queensland, Australia, 2002.

[SP02] C. Sanderson and K. K. Paliwal. Fast feature extraction method for robust face verification. *Electronic Letters*, 38(25):1648–1650, 2002.

[SP04] C. Sanderson and K.K. Paliwal. On the use of speech and face information for identity verification. Research Report 04-10, Idiap Research Institute, 2004.

[VJ04] P. Viola and M. Jones. Robust real-time object detection. *International Journal of Computer Vision*, 57(2), 2004.

[VSTJ92] A.P. Varga, H.J.M Steeneken, M. Tomlinson, and D. Jones. The NOISEX-92 study on the effect of additive noise on automatic speech recognition. Technical report, DRA Speech Research Unit, 1992.

[WLT07] F. W. Wheeler, X. Liu, and P.H. Tu. Multi-frame super-resolution for face recognition. In *Biometrics: Theory, Applications, and Systems, 2007. BTAS 2007. First IEEE International Conference on*, pages 1–6, Sept. 2007.