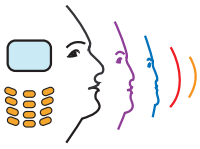


MOBIO

Mobile Biometry

http://www.mobioproject.org/

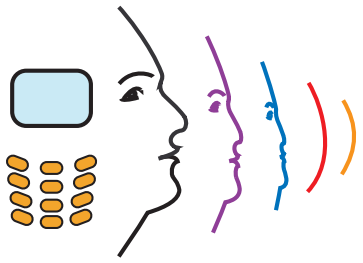
Funded under the 7th FP (Seventh Framework Programme)
Theme ICT-2007.1.4
[Secure, dependable and trusted Infrastructure]

D4.2: Report on the description and the evaluation of baseline algorithms for bi-modal authentication

Due date: 31/06/2009 **Submission date:** 15/05/2009
Project start date: 01/01/2008 **Duration:** 36 months
WP Manager: Norman Poh **Revision:** 2

Author(s): Norman Poh and Josef Kittler

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



D4.2: Report on the description and the evaluation of baseline algorithms for bi-modal authentication

Abstract:

This deliverable describes state-of-the-art fusion approaches for an audio-visual based biometric person authentication system. Two very different fusion approaches have been investigated, namely, joint-score based and normalization-based approaches. Both approaches were implemented using logistic regression. They were validated using both state-of-the-art face and speaker verification classifiers on the publicly available BANCA bimodal database. A preliminary analysis on the expert outputs show that the speaker verification expert is several orders better than the face verification expert. Nevertheless, both fusion approaches can still provide a marginal improvement over the best single expert. Although the normalization-based approach has already been investigated elsewhere in the literature, a novelty in this report is that the link between this approach with the Naive Bayes principle is clarified from a Bayesian perspective.

Contents

1 Introduction

Portable electronic devices such as mobile phones and PDAs are becoming important means to provide wireless access to the Internet and other telecommunication networks anytime, anywhere. Very often, such access requires the verification of the user's identity in order to ensure that the person is really who he/she claims to be. While knowledge-based authentication such as PINs or passwords can be used, they can be forgotten, or easily compromised when shared, copied or stolen. In comparison, biometrics is a more effective alternative because it is by far a more natural, reliable and friendly means of authentication. Thanks to the availability of cameras and microphones in today's mobile devices, audio- and visual-based biometrics such as face and speech can be readily used for this purpose.

In this context, we aim to develop and evaluate new mobile services that are secured by bi-modal speech and face biometrics. We shall call this problem "mobile biometry" (Mobio). Due to the device mobility, the problem of biometric authentication is much more challenging for at least two reasons: First, it has to deal with changing and often uncontrolled environments, e.g., external noise and varying illumination conditions. Under such conditions, a biometric query data can appear very differently from the one acquired during enrollment (template). As a consequence, the device performance can degrade drastically. Second, mobile devices have limited memory and CPU resources. This provides a natural constraint on the size of biometric template/model[1] and the type of processing algorithms (which favor those of low computation).

One possible way to improve the authentication performance is by using more than one biometrics, also known as *multimodal biometrics* in the literature [RNJ06] (and references herein).

2 Multimodal Biometric Fusion

Combining several systems is a well studied subject in pattern recognition [TH94] in general; in applications related to audio-visual speech processing [Luc02] [Lue97, CR98]; in speech recognition – examples of methods are multi-band [Cer99], multi-stream [Dup00, Hag01], front-end multi-feature [Shi01] approaches and the union model [MS03]; in the form of ensemble [Bro03]; in audio-visual person authentication [San02]; and, in multi-biometrics [Nan05, Poh06, Kry07, Ric07, RNJ06] (and references herein), among others. In fact, one of the earliest work addressing multimodal biometric fusion was reported in 1978 [Fej78]. Therefore, biometric fusion has a history of more than 30 years.

[1] We use the term "template" when referring to the stored features representing a person's biometric trait whereas the term "model" as a more general concept in order to refer to the parameters of a discriminative classifier or a statistical model, or that of an intermediate feature extraction process such as the Eigenspace analysis.

2.1 Fusion Techniques

In the literature, there are several methods to combine multimodal information. These methods are known as *fusion techniques*. Common fusion techniques include fusion at the *feature level* (extracted or internal representation of the data stream) or *score level* (output of a single system). Between the two, the latter is more commonly used in the literature.

Some studies further categorize three levels of score level fusion [BF95], namely, fusion using the scores directly, using a *set of most probable* category labels (called abstract level) or using the *single most probable* categorical label (called decision level). We will focus on the score level for two reasons: the last two cases can be derived from the score and more importantly, by using only labels instead of scores, precious information is lost, thus resulting in inferior performance [LLJ+05].

2.2 Feature Level versus Score Level Fusion

Although information fusion at the feature level is certainly much richer, exploiting such information by concatenation, for instance, may result in the *curse of dimensionality* [Bis99, Sec. 8.6]. In brief, it states that combined information (feature vector) may have a too high dimension that the problem cannot be solved easily by a given classifier. Furthermore, not all feature types are *compatible* at this level, i.e., of the same dimension, type and sampling rate. The feature level fusion certainly merits a thorough investigation but will not be addressed here. This issue will be addressed in the deliverables D4.3 and D4.4.

On the other hand, working at the score level conceals both the problems of curse of dimensionality and feature compatibility. Furthermore, the algorithms developed at the score level can be independent of any biometric system. Being aware that the only information retained is score, any additional information desired to be tapped must be fed externally. It should be noted that the feature level fusion converges to the score level fusion by assuming independence among the biometric feature sets. This assumption is perfectly acceptable in the context of multimodal biometric fusion but does not hold when the feature sets are derived from the same biometric sample. In this situation, the dependency at the feature level will certainly occur at the score level. Consequently, such dependency can still be handled at the score level. For the audio-visual based fusion problem treated here, although there are two biometric modalities, both of them are obtained from the same video recordings. As a result, a certain degree of dependency is expected.

2.3 Fusion Techniques

While many fusion classifiers have been used in the literature, as summarized in [RNJ06], as baseline fusion techniques, we have opted for a linear discriminative classifier, and in particular, logistic regression. It is, in essence, a weighted sum fusion with the non-linear sigmoid (or logistic) function. To our best knowledge, non-linear fusion classifiers have not been reported to outperform linear fusion classifiers [PB03]. A plausible explanation of this is the fusion problem in the expert output space can be solved using a linear decision

boundary. In fact, a direct examination of the data, after appropriately processing the expert outputs, does suggest such a conjecture (see Figure 3, for instance).

2.3.1 Joint-score Based Fusion

Let y_1 be the face expert (verification classifier) output whereas y_2 be that of the speech expert. It is convenient to stack the expert outputs together to form a vector $\mathbf{y} = [y_1, y_2, \ldots, M]'$ for combining M expert outputs.

Note that the discussion here is intended to be very general in two ways. First, the fusion approaches to be discussed are applicable to any number of expert outputs, although we only deal with $M = 2$ here. Second, as long as posterior probability is concerned, any *non-linear* classifier giving probabilistic output can be used. For the purpose of illustration, logistic regression is used. This is not a limitation because, for instance, non-linearity can be introduced into logistic regression by expanding the observation space using a polynomial function, e.g., [KPF+07].

The logistic regression classifier gives the following output:

$$y_{com}^{LR} \equiv P(\mathtt{C}|\mathbf{y}) = \frac{1}{1 + \exp(-g(\mathbf{y}))}, \tag{1}$$

where

$$g(\mathbf{y}) = \sum_{i=1}^{M} \beta_i y_i + \beta_0, \tag{2}$$

and the weight parameters β_i are optimized using gradient ascent to maximize the likelihood of the training data given the LR model [Dob90].

We shall refer to (1) as the *joint-score* based fusion, as opposed to the normalization-based approach [JNR05]. The *normalization-based approach* is based on the premise that the expert outputs are independent and hence can be processed separately. In the probabilistic sense, this implies the use of the Naive Bayes principle.

2.3.2 Normalization-based Fusion

In order to show how the normalization-based approach works, we shall use the logistic regression classifier already presented but applied to each expert output individually. Hence, for (1), we shall use the term y_i instead of $\mathbf{y}$, thus estimating $P(\mathtt{C}|y_i)$, for each i. In this context, for (2), one only needs to estimate two parameters (i.e., β_1 and β_0) for each expert i.

If one applies the logit transform to the posterior probability for each expert output i, one obtains the following form

$$g(y_i) = \log\left\{\frac{P(\mathtt{C}|y_i)}{1 - P(\mathtt{C}|y_i)}\right\} = \log\left\{\frac{P(\mathtt{C}|y_i)}{P(\mathtt{I}|y_i)}\right\}, \tag{3}$$

which is, in essence, a log-likelihood ratio test.

From the Bayes rule, it can be recognized that the posterior probability can, alternatively, be estimated via the following relationship:

$$P(\mathtt{C}|y_i) = \frac{P(\mathtt{C})p(y_i|\mathtt{C})}{\sum_{k=\mathtt{C},\mathtt{I}} P(k)p(y_i|k)} \propto P(\mathtt{C})p(y_i|\mathtt{C}), \tag{4}$$

where $P(k)$ for $k \in \{\mathtt{C}, \mathtt{I}\}$ is the prior class probability according to the training data and $p(y_i|k)$ is the density of expert output i, which can be estimated using any density estimator, e.g., Gaussian Mixture Models (GMM), Parzen windows, k-nearest neighbour, etc [Bis99]. Using the same approach, it can also be deduced that

$$P(\mathtt{I}|y_i) \propto P(\mathtt{I})p(y_i|\mathtt{I}). \tag{5}$$

Writing (3) using (4) and (5), we obtain:

$$g(y_i) = \log\left\{\frac{p(y_i|\mathtt{C})P(\mathtt{C})}{p(y_i|\mathtt{I})P(\mathtt{I})}\right\}. \tag{6}$$

Therefore, we observe that the logit transform simply maps the posterior probability to an output which could have been otherwise achieved using a Bayes classifier (hence requiring the estimation of $p(y_i|k)$ for both classes k) implementing a standard log-likelihood ratio test.

The normalization-based fusion to be used in this deliverable will actually take the sum of all $g(y_i)$'s, i.e.,

$$y_{com}^{sum} = \sum_{i=1}^{M} g(y_i) \tag{7}$$

where $M = 2$ in our case (bimodal fusion). To show the significance of (7) as a realization of the Naive Bayes principle (to an ignorable constant), we shall again consider the case if the post-processed expert output (in the log-likelihood ratio domain) was actually estimated using a Bayesian classifier, via (6). Then, (7) can be rewritten as:

$$\begin{aligned} y_{com}^{sum} &= \sum_{i=1}^{M} \log\left\{\frac{p(y_i|\mathtt{C})P(\mathtt{C})}{p(y_i|\mathtt{I})P(\mathtt{I})}\right\} \\ &= \log\left\{\frac{\prod_{i=1}^{M} p(y_i|\mathtt{C})}{\prod_{i=1}^{M} p(y_i|\mathtt{I})}\right\} + \underbrace{M \log\left\{\frac{P(\mathtt{C})}{P(\mathtt{I})}\right\}}. \end{aligned} \tag{8}$$

The output y_{com}^{sum} is, however, different from the usual Naive Bayes classifier which is:

$$y_{com}^{naive} = \log\left\{\frac{\prod_{i=1}^{M} p(y_i|\mathtt{C})}{\prod_{i=1}^{M} p(y_i|\mathtt{I})}\right\} + \underbrace{\log\left\{\frac{P(\mathtt{C})}{P(\mathtt{I})}\right\}} \tag{9}$$

Comparing (9) to (8), it is immediate apparent that y_{com}^{sum} is more than y_{com}^{naive} by a constant term of $(M-1)\log\left\{\frac{P(\mathtt{C})}{P(\mathtt{I})}\right\}$. This term is constant because it is independent of a test data

point, but dependent only on the training data, noting that the log prior ratio term in (9) corresponds to that of the training data set. Therefore, strictly speaking, in order to apply the Naive Bayes principle using the sum of logits, as in (7), one should compute the log ratio as follows:

$$y_{com}^{naive} = \sum_{i=1}^{M} g(y_i) - (M-1)\log\left\{\frac{P(\mathtt{C})}{P(\mathtt{I})}\right\}$$

Fortunately, in all cases, the underbraced constants in both (9) and (8) are immaterial since for any combined score y_{com}, the following decision is used:

$$\text{decision}(y_{com}) = \begin{cases} accept & \text{if } y_{com} > \Delta \\ reject & \text{otherwise,} \end{cases} \tag{10}$$

where the optimal decision threshold Δ is *optimized separately* depending on the cost of false acceptance and false rejection (to be discussed in Section 3.3). As can be observed, the adjustment for the threshold Δ is to compensate for the different application-dependent costs (of false acceptance and false rejection), which will affect the log prior ratio during testing. Therefore, the log prior ratio represents an external prior knowledge which, in any case, is likely to be different from that of the training data. An important implication as well as advantage of this is that, even though the class priors have changed, one does not need to retrain the fusion classifier. This justifies a separate optimization procedure for setting the global decision threshold.

3 Experiments

3.1 Database

In order to be consistent with the previous deliverables (D3.1 and D3.2), we shall use the same database, baseline systems and experimental protocols. The database used here is the BANCA database [MKS+04a]. This is a bimodal database recording from a camcorder, registering 52 people reading text-prompted sentences as well answering short questions. The sample images, for all three conditions are shown in Figure 1.

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.

A novel aspect concerning the usage of this database, unlike precedent efforts in [MKS+04a] or [MKS+04b], is that *video sequences* are actually used here, rather than *still images* extracted from the video sequence.

(a) Controlled (b) Adverse (c) Degraded

Figure 1: Automatically cropped face images from the three conditions in the BANCA databases, namely controlled, adverse and degraded conditions. Following the "P-protocol", the face models were trained only on the data set captured under the controlled condition, but tested on all three conditions.

3.2 Baseline Systems

The face and speaker verification baseline systems are Bayesian classifiers whose class-conditional densities are approximated using Gaussian Mixture Models (GMMs) with the Maximum *a posteriori* adaptation [RQD00]. This is a long-standing state-of-the-art classifier for the speaker verification, but since then, has also been successfully used for the face verification problem [CSM03]. The face verification problem can benefit from this approach mainly thanks to parts-based local feature descriptors, as illustrated in Figure 2. The parts-based approach first divides an image into overlapping or non-overlapping blocks of image. For each block of image, its texture is described using a *local feature descriptor*. The local feature descriptors used here are based on a post-processed subset of Discrete Cosine Transform features called "DCTMod2" [SP02].

Let $\mathbf{X} \equiv \{\mathbf{x}_i | i = 1, \ldots, N\}$ be a sequence of N feature frames and each feature frame is denoted by $\mathbf{x}_i$ (for the i-th frame). For the face modality, a feature frame is a vector containing the DCT coefficients of a block of image. For the speech modality, a feature frame contains Mel-scale Cepstral Coefficients [RJ93]. These features are a short-term representation of spectral envelopes filtered by a set of filters motivated by the human auditory system.

Let $p(\mathbf{x}|\omega_o)$ be the likelihood function of the world or background model and $p(\mathbf{x}|\omega_j)$ be the model for the claimed identity $j \in \{1, \ldots, J\}$[2]. In parts-based face or speaker

[2]Note that we use a different notation here, i.e., ω_j, to denote the classes, as compared to Section 2, where $k \in \{\mathtt{C}, \mathtt{I}\}$ was used. The reason is that in the fusion process, one considers only binary classification, i.e., a person is either a genuine person (client), claiming to be the reference identity, or an impostor. In essence, there is only a single fusion classifier for all the enrollees in the database. For the baseline expert, on the other hand, the system designer needs to design an expert *for each enrollee*. As a result, it is necessary to distinguish models of different enrollees using ω_j.

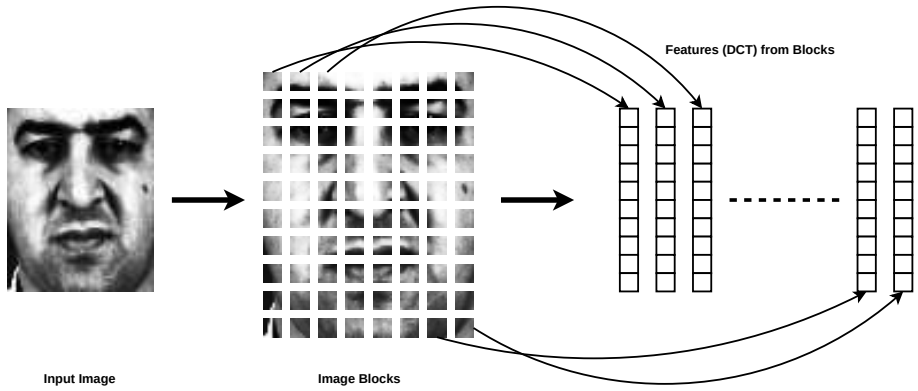


Figure 2: A flow chart of describing the extraction of feature vectors from the face image for the parts-based approach.

verification, both $p(\mathbf{x}|\omega_o)$ and $p(\mathbf{x}|\omega_j)$, for any j, are estimated using a Gaussian Mixture Model (GMM) [Bis99]. The world model is first obtained from a large pool of sequences $\{\mathbf{X}\}$ contributed by a large and possibly separate population of users (possibly from an external database than the one used for enrollment/testing). Each client-specific model is then obtained by adapting the world model upon the presentation of the enrollment data of a specfici user/client.

The GMM-based Bayesian classifier applies the log-likelihood ratio test, which is optimal in the Neyman-Pearson sense [DHS01]:

$$y = \frac{1}{N}\sum_i \log\left\{\frac{p(\mathbf{x}_i|\omega_j)}{p(\mathbf{x}_i|\omega_o)}\right\} \tag{11}$$

An important assumption here is that all feature frames are independently and identically distributed. An interpretation of this is that, thanks to the part-based approach, the relative location of an eye to the noise, or any two salient facial features are unimportant. Because of this, a practical advantage offered by this approach is that it is fairly robust to imperfectly found centers of the eye coordinates (needed for cropping a face from the background) given by a face detector.

If the score y is greater than a pre-specified threshold, one declares that the query data $\mathbf{X}$ belongs to the model j. Hence, this will result in an acceptance decision. Otherwise, one rejects the hypothesis and hence rejects the identity claim. The details of the face verification system can be found in Section A.

The speaker verification classifier used here differs from the face one in the following ways. First, the variability across sessions are removed thanks to now a standard technique called factor analysis [KBD05, VBS05, MSFB07]. This technique is applied to all training and test data prior to building a (client-specific) GMM model.

Second, rather than using the log-likelihood ratio test as in (11), a client-specific SVM is used instead. The SVM is designed to classify features not at the sequence level (in the

space $\{\mathbf{X}\}$) but at the so-called "GMM supervector" space. If a GMM has C components, the observation in this space consists of the mean vectors of all the C Gaussian components concatenated together to form a *supervector*. A supervector is thus a vector of fixed-size that is *independent* of the length of the speech utterance, hence allowing the speaker verification problem to be solved using discriminative approaches (which are well suited for classification problems with fixed-size observations). During training as well as testing, the GMM supervectors are submitted to the channel compensation technique via factor analysis. The SVM is thus trained to distinguish the supervector of one user versus all other users, with the impact of channel variability significantly reduced. As a result, the output of the speaker verification system used here is in terms of margin, i.e., how far a supervector (in the implicitly embedded space defined by a given kernel) is from the optimal decision boundary separating the claimed user identity from the rest of the users (clients) in the database. The details of the speaker verification system can be found in Section B.

3.3 Evaluation Metrics

We use two types of curves in order to compare the performance: the Detection Error Trade-off (DET) curve [MDK$^+$97] and the Expected Performance Curve (EPC) [BM04]. A DET curve is actually a Receiver Operator Curve (ROC) curve plotted on a scale defined by the inverse of a cumulative Gaussian density function, but otherwise similar in all aspects. We have opted to use EPC because it has been pointed out in [BM04] that two DET curves resulting from two systems are not comparable. This is because such comparison does not take into account how the decision thresholds are selected. EPC turns out to be able to make such comparison possible. Furthermore, the performance across different data sets, resulting in several EPCs, can be merged into a single EPC [PB05]. Although reporting performance in EPC is more meaningful than DET as far as performance comparison is concerned, it is relatively new and has not gained a widespread acceptance in the biometric community. As such, we shall also report performance in DET curves, but using only a subset of operating points.

The EPC curve, however, is less convenient to use because it requires two sets of match scores, one used for tuning the threshold (for a given operating cost), and the other used for assessing the performance. In our context, with the two-fold cross-validation defined on the database (as determined by g1 and g2), these two match scores can be conveniently used.

According to [BM04], one possible, and often used criterion is the weighted error rate (WER), defined by:

$$\mathrm{WER}(\beta, \Delta) = \beta\, \mathrm{FAR}(\Delta) + (1 - \beta)\, \mathrm{FRR}(\Delta), \tag{12}$$

where FAR is the false acceptance rate, FRR is the false rejection rate at a given threshold Δ and $\beta \in [0, 1]$ is a user-specified coefficient which balances FAR and FRR. The WER criterion generalizes the criterion used in the annual NIST's speaker evaluation [MPC05]

as well as the three operating points used in the past face verification competitions on the BANCA database [MKS+04b, MKS+04a], which used only three coefficients of β:

$$\beta = \frac{1}{1+R} \text{ for } R = \{0.1, 1, 10\}$$

which yields approximately $\beta = \{0.9, 0.5, 0.1\}$, respectively. Rather than just using the above specific β values, in this study, we use $\beta \in \{0.1, 0.2, \ldots, 0.8, 0.9\}$.

The procedure to calculate an EPC is as follows: Use g1 to generate the development match scores; and g2, the evaluation counterpart. For each chosen β, the development score set is used to minimize (12) in order to obtain an operational threshold. This threshold is then applied to the evaluation set in order to obtain the final pair of false acceptance rate (FAR) and false rejection rate (FRR). The EPC curve simply plots half total error rate (HTER) versus β, where HTER is the average of FAR and FRR. Alternatively, the generalization performance can also be reported in WER (as done in the previous BANCA face competitions). To plot the corresponding DET curve, we use the pair of FAR and FRR of all the operating points, as determined by β. Note that this DET curve is a *subset* (in fact discrete version) of a conventional continuous DET curve because the latter is plotted from continuous empirical functions of FAR and FRR. By plotting the discrete version of the DET curve, we establish a *direct correspondence* between EPC and DET, satisfying both camps of biometric practitioners, while retaining the advantage of EPC which makes performance comparison between systems less biased.

3.4 Analysis

At the score-level fusion, the bimodal face and speech fusion problem is greatly simplified to solving a two dimensional problem defined by the expert output. The data for the g1 data set is shown in Figure 3.

Since both the face and speaker verification experts rely on the same input signal, i.e., from the same video, we expect that there is a certain degree of correlation. On the other hand, a counter argument is that since both modalities are affected by different noise sources, it is reasonable to expect both modalities to be independent. This motivated us to perform a correlation-based analysis.

The result of correlation, conditioned on each class (subjecting to being genuine or impostor matching) are shown in Figures 4(a) and (b). For the client class, unfortunately, we observed that the marginal distribution is not Gaussian. As a result, the measured correlation may not reflect well the true underlying dependency (the latter being the real criterion of interest).

As an exploratory study, we applied the Gaussian Copula model [CLV04] to better uncover the underlying correlation structure. This consists of first transforming each expert output scores to a uniform distribution, and then reprojecting back the scores to a Gaussian distribution. After this transformation, the marginal of the transformed outputs will be a standard normal distribution. The pairwise correlation between the two expert outputs

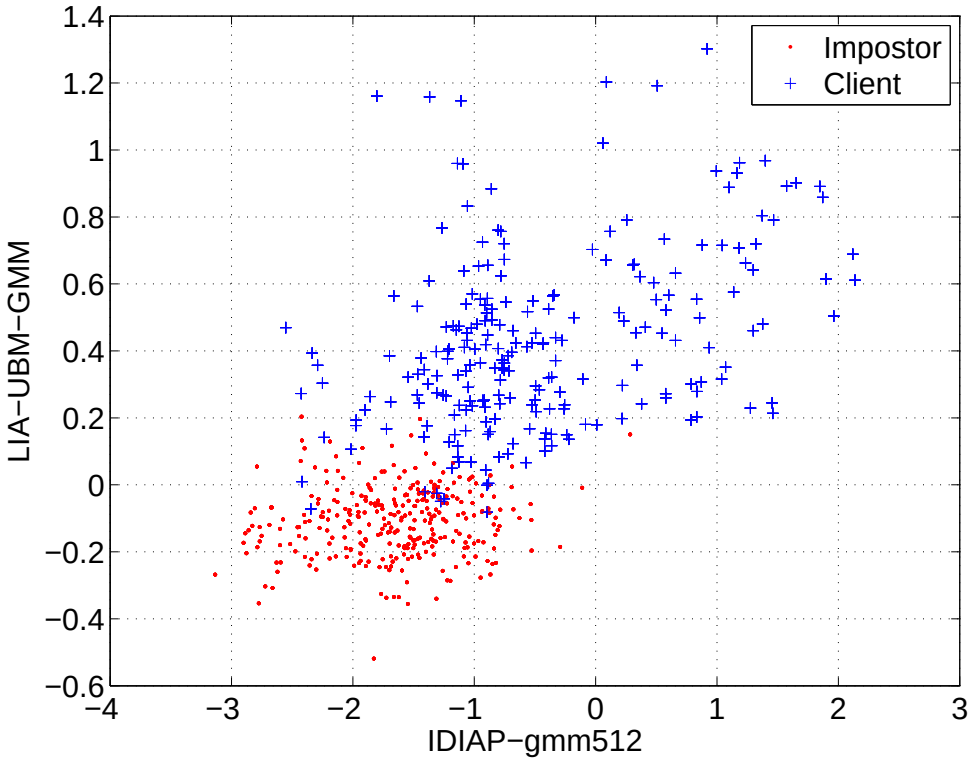


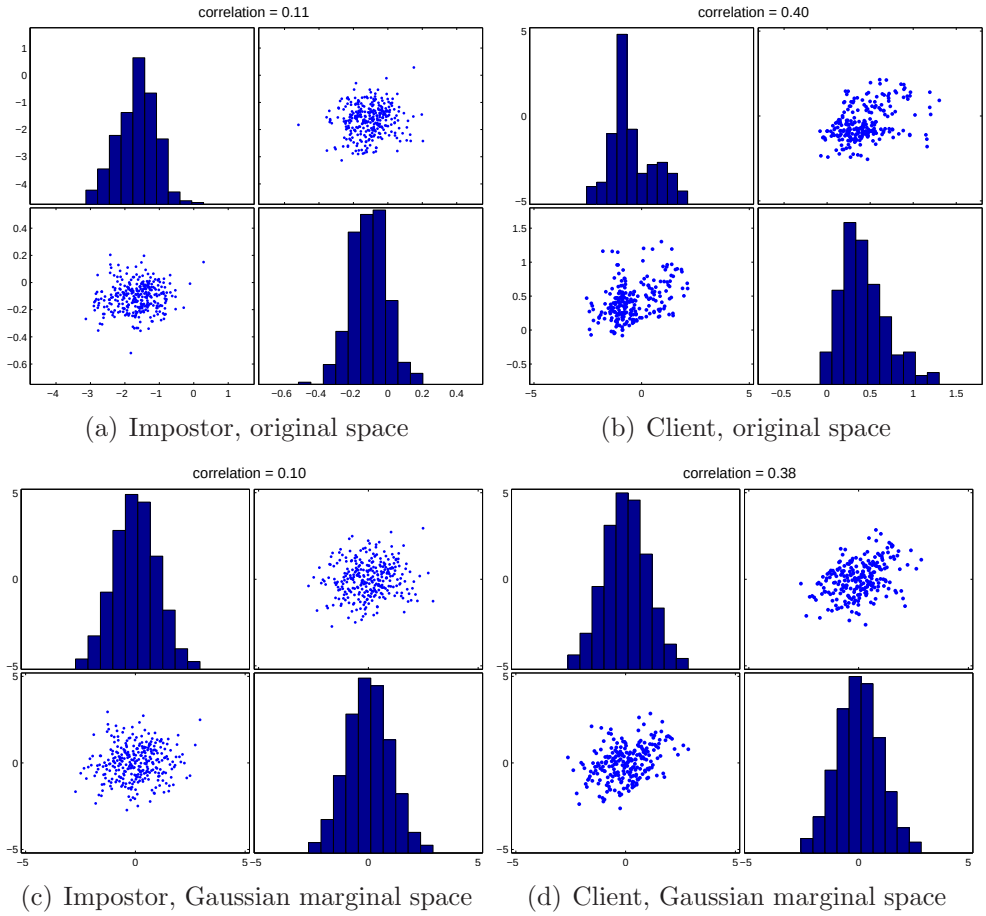
Figure 3: The face and speech bimodal fusion problem. The x-axis is the output of the face verification expert whereas the y-axis is that of the speaker one. Blue plus signs denote client accesses where red dot denote impostor accesses.

can then be measured in this space. The pairwise scatter plots of the transformed outputs conditioned for each class are shown in Figures 4(c) and (d).

As can be observed, the correlation for the client match scores are relatively high, i.e., 0.38 according to the Gaussian Copula model (or 0.40 in the original space), but are relative low for the impostor scores, i.e., 0.10 (resp. 0.11).

The correlation of the client scores is considered *weak*, but never the less exists (a strong correlation would achieve 0.95 or higher, for instance). One could also conclude that there exists no correlation between the experts for the impostor scores. This is perfectly acceptable. An interpretation of this is that the face of person A and the speech of person B are not related to each other under zero-effort impersonation. This, however, may not hold under active/informed impostor attack. Unfortunately, no data is available to validate this conjecture[3]

[3]The MOBIO database could be used for this purpose.



(a) Impostor, original space (b) Client, original space

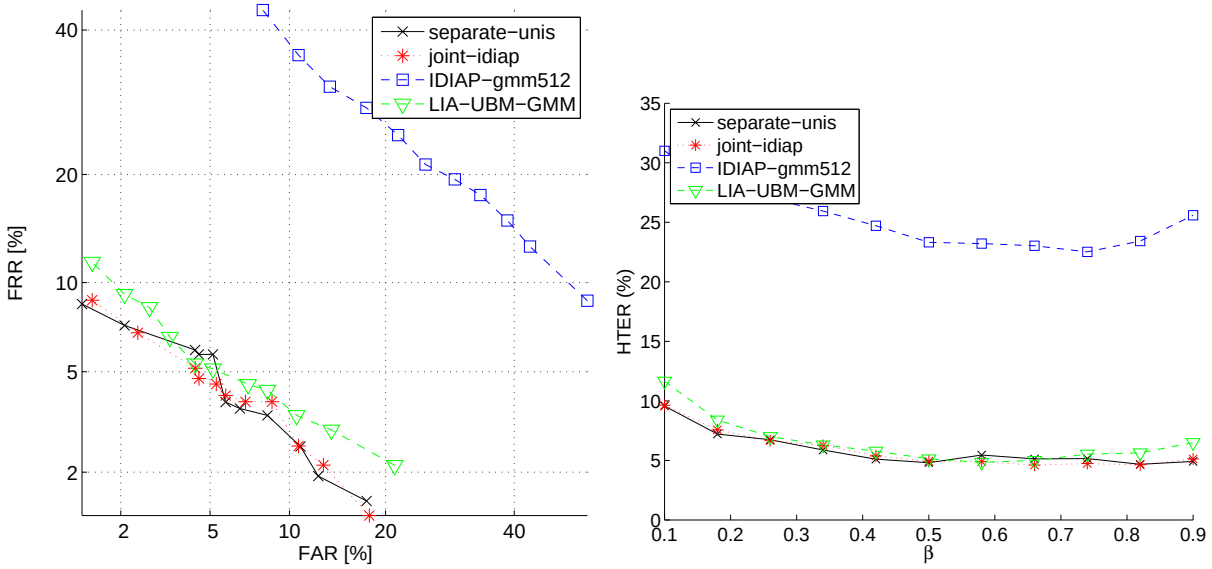
(c) Impostor, Gaussian marginal space (d) Client, Gaussian marginal space

Figure 4: A pair-wise scatter plot of the bimodal expert outputs subjecting to impostor or genuine (client) matching in the original space as well as the Gaussian marginal space obtained after applying the Gaussian Copula transformation.

3.5 Fusion Results

This section reports the performance of the joint-score based fusion (using (1)) versus the normalization-based fusion (using (7)). The results are shown in Figure 5 using both DET and EPC curves. As can be observed, the face verification expert is many more times worse than the speaker verification expert. This is because the scenarios defined for the BANCA database is such that the degradation in the face modality is much more drastic than the speech modality. However, as can be observed, on both cases, the fusion classifier was able to benefit from both modalities in order to further improve over the speech modality, albeit insignificantly.

Both the two fusion classifiers have similar performance, although, according to the correlation study reported in Section 3.4, one would expect that the joint-score approach to outperform the normalization-based approach. This result is not particularly surprising



(a) DET (b) EPC

Figure 5: The pooled performance of the baseline systems as well as fusion systems, assessed on both the g1 and g2 data set, using (a) DET and (b) EPC curves. The experiments were carried out using the P-protocol. IDIAP-gmm512 refers to the face verification expert; LIA UBM-GMM, a speaker verification expert; joint-idiap, the joint-score based fusion; and, separate-unis, the normalization-based fusion.

as various independent researchers have previously shown the effectiveness of the Naive Bayes principle even under dependent features.

4 Conclusions

This deliverable provides a very first baseline of bimodal fusion of a video-based biometric authentication/verification system. This was validated using both state-of-the-art face and speaker verification classifiers on the publicly available BANCA bimodal database. We have implemented two fusion classifiers using two very different approaches, namely, joint-score based and normalization-based fusion strategies, implemented using logistic regression. Although the normalization-based approach has already been investigated elsewhere [JNR05], the link between this approach with the Naive Bayes principle has not been clarified. We formally show this link using a Bayesian framework.

Our experiments suggest that the performance of a multimodal biometric fusion system is better than that of any single best expert. In our case, this performance gain is marginal because the speech expert is an order of magnitude better than the face expert. This is consistent with the findings in the literature, e.g., [KHDM98, RNJ06].

Acknowledgments

We would like to thank the following partners for their contributions to this deliverable:

- Sébastien Marcel (IDIAP) for providing the face verification system and for commenting on the deliverable
- Christopher McCool (IDIAP) for providing the IDIAP joint-score fusion system
- Driss Matrouf (LIA) for providing the speaker verification system

References

[BBF+04] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska, and D. A. Reynolds. A tutorial on text-independent speaker verification. *EURASIP Journal on Applied Signal Processing, Special issue on biometric signal processing*, 2004.

[BF95] R. Brunelli and D. Falavigna. Personal Identification Using Multiple Cues. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 17(10):955–966, 1995.

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

[BM04] S. Bengio and J. Marithoz. The Expected Performance Curve: a New Assessment Measure for Person Authentication. In *The Speaker and Language Recognition Workshop (Odyssey)*, pages 279–284, Toledo, 2004.

[Bro03] G. Brown. *Diversity in Neural Network Ensembles*. PhD thesis, School of Computer Science, Uni. of Birmingham, 2003.

[Cer99] C. Cerisara. *Contribution de l'Approache Multi-Bande à la Reconnaissance Automatic de la Parole*. PhD thesis, Institute Nationale Polytéchnique de Lorraine, Nancy, France, 1999.

[CL01] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[CLV04] U. Cherubini, E. Luciano, and W. Vecchiato. *Copula Methods in Finance*. Wiley, 2004.

[CR98] T. Chen and R. Rao. Audio-Visual Integration in Multimodal Communications. *Proc. IEEE*, 86(5):837–852, 1998.

[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.

[DHS01] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification and Scene Analysis*. John Wiley and Sons, New York, 2001.

[Dob90] A. J. Dobson. *An Introduction to Generalized Linear Models*. CRC Press, 1990.

[DPMR00] G. Doddington, M. Przybocki, A. Martin, and D. Reynolds. The NIST speaker recognition evaluation — overview, methodology, systems, results, perspective. *Speech Communication*, 31(2-3):225–254, 2000.

[Dup00] S. Dupont. *Étude et Dévelopment de Nouveaux Paradigmes pour la Reconnaissance Robuste de la Parole*. PhD thesis, Laboratoire TCTS, Universite de Mons, Belgium, 2000.

[Fej78] A. Fejfar. Combining Techniques to Improve Security in Automated Entry Control. In *Carnahan Conf. On Crime Countermeasures*, 1978. Mitre Corp. MTP-191.

[Hag01] Astrid Hagen. *Robust Speech Recognition Based on Multi-Stream Processing*. PhD thesis, Ecole Polytechnique Federale de Lausanne, Switzerland, 2001.

[JNR05] A. Jain, K. Nandakumar, and A. Ross. Score Normalisation in Multimodal Biometric Systems. *Pattern Recognition*, 38(12):2270–2285, 2005.

[KBD05] P. Kenny, G. Boulianne, and P. Dumouchel. Eigenvoice Modeling With Sparse Training Data. *IEEE Transactions on Speech and Audio Processing*, 13(3):345, 2005.

[KHDM98] J. Kittler, M. Hatef, R. P.W. Duin, and J. Matas. On Combining Classifiers. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998.

[KPF+07] J. Kittler, N. Poh, O. Fatukasi, K. Messer, K. Kryszczuk, J. Richiardi, and A. Drygajlo. Quality Dependent Fusion of Intramodal and Multimodal Biometric Experts. In *Proc. of SPIE Defense and Security Symposium, Workshop on Biometric Technology for Human Identification*, volume 6539, 2007.

[Kry07] K. Kryszczuk. *Classification with Class-independent Quality Information for Biometric Verification*. PhD thesis, Swiss Federal Institute of Technology in Lausanne (Ecole Polytechnique Fédérale de Lausanne), 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.

[LG96] C. Lee and J. Gauvain. Bayesian adaptive learning and map estimation of hmm. In C.-H. Lee, F. Soong, and K. Paliwal, editors, *Automatic speech and speaker recognition : Advanced topics*, pages 83–107. Kluwer Academic Publishers, Boston, Massachusetts, USA, 1996.

[LLJ+05] Y. Lee, K. Lee, H. Jee, Y. Gil, W. Choi, D. Ahn, and S. Pan. Fusion for Multimodal Biometric Identification. In *LNCS 3546, 5th Int'l. Conf. Audio- and Video-Based Biometric Person Authentication (AVBPA 2005)*, pages 1071–1079, New York, 2005.

[Luc02] S. Lucey. *Audio Visual Speech Processing*. PhD thesis, Queensland University of Technology, 2002.

[Lue97] J. Luettin. *Visual Speech and Speaker Recognition*. PhD thesis, Department of Computer Science, University of Sheffield, 1997.

[MDK+97] A. Martin, G. Doddington, T. Kamm, M. Ordowsk, and M. Przybocki. The DET Curve in Assessment of Detection Task Performance. In *Proc. Eurospeech'97*, pages 1895–1898, Rhodes, 1997.

[MKS+04a] K. Messer, J. Kittler, M. Sadeghi, M. Hamouz, A. Kostin, F. Cardinaux, S. Marcel, S. Bengio, C. Sanderson, N. Poh, Y. Rodriguez, J. Czyz, L. Vandendorpe, C. McCool, S. Lowther, S. Sridharan, V. Chandran, R. P. Palacios, E. Vidal, L. Bai, L-L. Shen, Y. Wang, Chiang Yueh-Hsuan, H-C. Liu, Y-P. Hung, A. Heinrichs, M. Muller, A. Tewes, C. vd Malsburg, R. Wurtz, Zg. Wang, Feng Xue, Yong Ma, Qiong Yang, Chi Fang, Xq. Ding, S. Lucey, R. Goss, , and H. Schneiderman. Face authentication test on the banca database. In *Int'l Conf. Pattern Recognition (ICPR)*, volume 4, pages 523–532, 2004.

[MKS+04b] K. Messer, J. Kittler, M. Sadeghi, M. Hamouz, A. Kostyn, S. Marcel, S. Bengio, F. Cardinaux, C. Sanderson, N. Poh, Y. Rodriguez, K. Kryszczuk, J. Czyz, L. Vandendorpe, J. Ng, H. Cheung, and B. Tang. Face authentication competition on the banca database. In *Intl. Conf. Biometric Authentication*, pages 8–15, 2004.

[MPC05] A. Martin, M. Przybocki, and J. P. Campbell. *The NIST Speaker Recognition Evaluation Program*, chapter 8. Springer, 2005.

[MS03] Ji Ming and F. Jack Smith. Speech Recognition with Unknown Partial Feature Corruption - a Review of the Union Model. *Computer Speech and Language*, 17:287–305, 2003.

[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.

[Nan05] K. Nandakumar. Integration of M]ultiple Cues in Biometric Systems. Master's thesis, Michigan State University, 2005.

[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. Database, Protocol and Tools for Evaluating Score-Level Fusion Algorithms in Biometric Authentication. *Pattern Recognition*, 39(2):223–233, February 2005.

[PM93] W. B. Pennebaker and J. L. Mitchell. *JPEG still image data compression standard*. New York: Van Nostrand Reinhold, 1993.

[Poh06] N. Poh. *Multi-system Biometric Authentication: Optimal Fusion and User-Specific Information*. PhD thesis, Swiss Federal Institute of Technology in Lausanne (Ecole Polytechnique Fédérale de Lausanne), 2006.

[Rey97] D. A. Reynolds. Comparison of background normalization methods for text-independent speaker verification. In *Proc. European Conference on Speech Communication and Technology (EuroSpeech)*, pages 963–966, Rhodes, Greece, September 1997.

[Ric07] J. Richiardi. *Probabilistic Models for Multi-Classifier Biometric Authentication Using Quality Measures*. PhD thesis, Swiss Federal Institute of Technology in Lausanne (Ecole Polytechnique Fédérale de Lausanne), 2007.

[RJ93] L. Rabiner and B-H Juang. *Fundamentals of Speech Recognition*. Oxford University Press, 1993.

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

[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.

[Shi01] M. L. Shire. *Discriminant Training of Front-End and Acoustic Modeling Stages to Heterogeneous Acoustic Environments for Multi-Stream Automatic Speech Recognition*. PhD thesis, University of California, Berkeley, USA, 2001.

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

[TH94] S.N. Srihari T.K. Ho, J.J. Hull. Decision Combination in Multiple Classifier Systems. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 16(1):66–75, January 1994.

[VBS05] R. Vogt, B. Baker, and S. Sridharan. Modelling Session Variability in Text-Independent Speaker Verification. In *Proc. European Conference on Speech Communication and Technology (EuroSpeech)*, 2005.

A Parts-Based Gaussian Mixture Model (PB-GMM) for Face Verification

The first face verification baseline model implementation presented in this report combines part-based approaches and GMM modeling. Parts-based approaches divide the face into blocks, or parts, and treats each block as a separate observation of the same underlying signal (the face). According to this technique, a feature vector is obtained from each block by applying the Discrete Cosine Transform (DCT) and the distribution of these feature vectors is then modelled using GMMs. Several advances have been made upon this technique, for instance, Cardinaux *et al.* [CSM03] proposed the use of background model adaptation while Lucey and Chen [LC04] examined a method to retain part of the structure of the face utilising the parts-based framework as well as proposing a relevance based adaptation.

Feature Extraction

The feature extraction algorithm is described by the following steps. The face is normalised, registered and cropped. This cropped and normalised face is divided into blocks (parts) and from each block (part) a feature vector is obtained. Each feature vector is treated as a separate observation of the same underlying signal (in this case the face) and the distribution of the feature vectors is modelled using GMMs. This process is illustrated in Figure 2.

The feature vectors from each block are obtained by applying the DCT. Even advanced feature extraction methods such as the DCTmod2 method [SP02] use the DCT as their basis feature vector; the DCTmod2 feature vectors incorporate spatial information within the feature vector by using the deltas from neighbouring blocks. The advantage of using only DCT feature vectors is that each DCT coefficient can be considered to be a frequency response from the image (or block). This property is exploited by the JPEG standard [PM93] where the coefficients are ranked in ascending order of their frequency.

Feature Distribution Modelling

Feature distribution modelling is achieved by performing background model adaptation of GMMs [CSM03, LC04]. The use of background model adaptation is not new to the field of biometric authentication; in fact, it is commonly used in the field of speaker verification [DPMR00]. Background model adaptation first trains a world (background) model Ω_{world} from a set of faces and then derives the client model for the i^{th} client Ω^i_{client} by adapting the world model to match the observations of the client.

Two common methods of performing adaptation are mean only adaptation [Rey97] and full adaptation [LG96]. Mean only adaptation is often used when there are few observations available because adapting the means of each mixture component requires fewer

observations to derive a useful approximation. Full adaptation is used where there are sufficient observations to adapt all the parameters of each mode. Mean only adaptation is the method chosen for this work as it requires fewer observations to perform adaptation, this is the same adaptation method employed by Cardinaux *et al.* [CSM03].

Verification

To verify an observation, $\mathbf{x}$, it is scored against both the client (Ω^i_{client}) and world (Ω_{model}) model, this is true even for methods that do not perform background models adaptation [SP02]. The two models, Ω^i_{client} and Ω_{world}, produce a log-likelihood score which is then combined using the log-likelihood ratio (LLR),

$$h(\mathbf{x}) = \ln(p(\mathbf{x} \mid \Omega^i_{client})) - \ln(p(\mathbf{x} \mid \Omega_{world})), \tag{13}$$

to produce a single score. This score is used to assign the observation to the world class of faces (not the client) or the client class of faces (it is the client) and consequently a threshold τ has to be applied to the score $h(\mathbf{x})$ to declare (verify) that $\mathbf{x}$ matches to the i^{th} client model Ω^i_{client}, i.e if $h(\mathbf{x}) \geq \tau$.

B Gaussian Mixture Model-Support Vector Machine Based Speaker Verification

The use of GMM in a GMM-UBM framework has been a standard approach in the speaker verification [BBF+04]. In addition to this framework, the Latent Factor Analysis (LFA) is systematically applied for all systems in training and testing [KBD05, VBS05, MSFB07]. From the resulting session compensated model it is possible to extract supervectors by concatenating Gaussian means. These supervectors can be used directly in a SVM classifier. This association between the factor analysis and SVM allows to benefit from the FA decomposition power and SVM classification power. The implemented baseline system uses Z-T-norm for score normalization.

Feature extraction

The signal is characterized by 50 coefficients including 19 linear frequency cepstral coefficients (LFCC), their first derivative, their first 11 coefficients of second derivatives and the delta-energy. They are obtained as follows: 24 filter bank coefficients are first computed over 20ms Hamming windowed frames at a 10ms frame rate. Bandwidth is limited to the 300-3400Hz range.

Here, the energy coefficients are first normalized using a mean removal and variance normalization in order to fit a 0-mean and 1-variance distribution. The energy component is then used to train a three component GMM, which aims at selecting informative frames. The most energized frames are selected through the GMM. Once the speech segments of a signal are selected, a final process is applied in order to refine the speech segmentation:
1- overlapped speech segments between both the sides of a conversation are removed,
2- morphological rules are applied on speech segments to avoid too short ones, adding or removing some speech frames.

Finally, the parameter vectors are normalized to fit a 0-mean and 1-variance distribution. The mean and variance estimators used for the normalization are computed file by file on all the frames kept after applying the frame removal processing.

World models

Two GMM world models are used, one for males and one for females. The two GMM are trained using Fisher English Training Speech Part 1 (LDC:LDC2004S13), and consists of about 10 million speech frames each for males and females.

Resulting world models are 512 gender dependent GMM's with diagonal covariance matrices. For a better separation of initial classes, frames are randomly selected among the entire learning signal via a probability followed by an iteration of the EM algorithm, to

estimate the GMM parameters. During the estimation of the world model parameters, instead of using all the learning signals in their temporal order, 10% of frames is selected randomly at each new iteration. For the two last iterations, the entire signal is classically used in its temporal order. During all the process, a variance flooring is applied so that no variance value is less than 0.5.

Client, test and impostor models with Factor Analysis

A speaker model can be decomposed into three different components: world, a speaker dependent and session dependent components [KBD05, VBS05, MSFB07]. A GMM mean super-vector is defined as the concatenation of the GMM component means. In the following, (h, s) will indicate the session h of the speaker s. The latent factor analysis model, can be written as:

$$\mathbf{m}_{(h,s)} = \mathbf{m} + \mathbf{D}\mathbf{y}_s + \mathbf{U}\mathbf{x}_{(h,s)}, \tag{14}$$

where $\mathbf{m}_{(h,s)}$ is the session-speaker dependent super-vector mean, $\mathbf{D}$ is $S \times S$ diagonal matrix (S is the dimension of the supervector), $\mathbf{y}_s$ the speaker vector (its size equal S), $\mathbf{U}$ is the session variability matrix of low rank R (a $S \times R$ matrix) and $\mathbf{x}_{(h,s)}$ are the session factors, a R vector. Both $\mathbf{y}_s$ and $\mathbf{x}_{(h,s)}$ are normally distributed among $\mathcal{N}(0, I)$. $\mathbf{D}$ satisfies the following equation $\mathbf{I} = \tau\mathbf{D}^t\mathbf{\Sigma}^{-1}\mathbf{D}$ where τ is the *relevance factor* required in the standard MAP adaptation.

The client model is obtained by performing the decomposition of equation 14 and by retaining only the speaker dependent components:

$$\mathbf{m}_s = \mathbf{m} + \mathbf{D}\mathbf{y}_s, \tag{15}$$

The success of the factor analysis model relies on a good estimation of the $\mathbf{U}$ matrix, thanks to a sufficiently high amount of data, where a high number of different recordings per speaker is available. In these experiments the U matrix is trained by using about 240 speakers (120 males and 120 females) coming from NIST'04. For each speaker about 20 sessions are considered.

Kernel based scoring and SVM modeling

By using (15), the factor analysis model estimates supervectors containing only speaker information, normalized with respect to the session variability. A probabilistic distance kernel that computes a distance between GMM's, well suited for a SVM classifier. Let $\mathcal{X}_{\mathbf{s}}$ and $\mathcal{X}_{\mathbf{s}'}$ be two sequences of speech data corresponding to speakers $\mathbf{s}$ and $\mathbf{s}'$, the kernel formulation is given below.

$$K(\mathcal{X}_{\mathbf{s}}, \mathcal{X}_{\mathbf{s}'}) = \sum_{g=1}^{M} \left(\sqrt{\alpha_g}\mathbf{\Sigma}_g^{-\frac{1}{2}}\mathbf{m}_s^g\right)^t \left(\sqrt{\alpha_g}\mathbf{\Sigma}_g^{-\frac{1}{2}}\mathbf{m}_{s'}^g\right). \tag{16}$$

This kernel is valid when only means of GMM models are varying (weights and covariance are taken from the world model). $\mathbf{m_s}$ is taken here from the model in eq. 15, *i.e.* $\mathbf{m}_s = \mathbf{m} + \mathbf{D}\mathbf{y}_s$.

The LIA_SpkDet toolkit benefits from the LIBSVM [CL01] library to induce SVM and to classify instances. SVM models are trained with an infinite (very large in practice) C parameter thus avoiding classification error on the training data (hard margin behavior). The negative labeled examples are speakers from the normalization cohort.