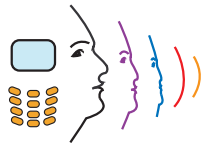


MOBIO

Mobile Biometry

http://www.mobioproject.org/

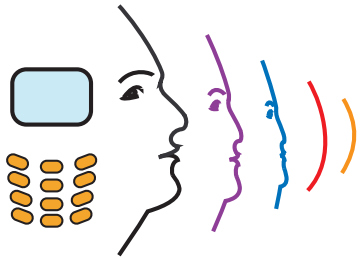
Funded under the 7th FP (Seventh Framework Programme)
Theme ICT-2007.1.4
[Secure, dependable and trusted Infrastructure]

D4.6: Description and Evaluation of Baseline Algorithms for Model Adaptation

Due date: 31/08/2009 **Submission date:** 30/07/2009
Project start date: 01/01/2008 **Duration:** 36 months
WP Manager: Norman Poh **Revision:** 1

Author(s): Norman Poh

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



D4.6: Description and Evaluation of Baseline Algorithms for Model Adaptation

Abstract:

This deliverable examines two solutions to guard against the change of the quality of biometric samples, in particular, as a result of changing acquisition environment as well as that of acquisition devices (i.e., matching between enrollment and query samples collected using different devices). The two solutions are model-level and score-level adaptation. The model-level adaptation attempts to update the parameters of the expert systems whereas the score-level adaptation merely post-processes the output of the baseline system. Although the potential of model-level adaptation is immense, e.g., reducing the error rate by as much as half, it requires additional labeled training data. In this study, manually labeled training data samples are considered, leading to a *supervised adaptation* strategy. This provides an upper bound of the achievable performance (with respect to an unsupervised strategy). The score-level adaptation we implemented is based on logistic regression. In comparison to the model-level adaptation strategy, the score-level adaptation one does not require labeled training data samples. However, the latter does require additional operational data reflecting the actual operational scenarios. With this procedure, our experiments based on the BANCA bimodal database reports that the improvement achievable (in terms of reduction in Equal Error Rate with repect to the baseline non-adaptive system) is 27% for the face expert and 42% for the speech expert.

Contents

1 Introduction

1.1 Motivations

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 whom 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 (referred to as reference model or 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).

1.2 Handling Variations in Biometric Samples

In order to guard against the degradation due to changing acquisition environment and/or devices, there are two somewhat opposing approaches being used in practice:

- by compensation, i.e., reducing the effect of a degrading factor as well as working on invariant feature representation
- by tracking the change, i.e., explicitly considering the change at all levels of the architecture

The first strategy seeks to rectify the change as much as possible. For instance, for face verification, one applies illumination normalization to guard against changes in illumination, or pose correction to rectify a given pose to a frontal one.

In the second strategy, one admits that a degrading factor cannot be completely removed. As a result, one attempts to track the change and update the system parameters as and when necessary. Several questions can arise when adopting this strategy:

[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 eigenspace analysis.

- Should the parameters of the existing model be overwritten?
- When several models are available, how can they be used jointly during inference?
- What criterion should be used to determine if a model needs to be updated?

The above issues are partially addressed in [PWKR09]. In summary, the first issue depends very much on the quality or an aspect of the degrading factor. For instance, in the case of face recognition, one would maintain a set of parameters for a given (discretized) head pose. Unfortunately, in practice, the noise affecting a face image sample is a composite factor, consisting of a given pose, illumination conditions, a given facial expression and a particular camera type (with a particular camera setting). As a result, ideally, one should maintain a model for the *same* composite factor (consisting of the same pose, the same illumination condition, the same facial expression) and the same camera type/configuration. As can be observed, the space of variation is possibly infinitesimal but can be finitely quantified via clustering, i.e., clustering the quality measures from a very large database of face images [PWKR09]. Then, ideally, one should maintain a set of parameters for each cluster of quality measures.

The second issue can be handled using a Bayesian framework. In essence, this consists of finding how probable that a face image belongs to a cluster of image quality (as found by clustering the quality measures) and the correct model is used for inference (i.e., testing the hypothesis that the model belongs to a particular claimed identity).

The third issue is addressed using a semi-supervised learning technique. In the case of bimodal authentication involving face and speech, one can employ the "co-training" algorithm [BM98]. The algorithm attempts to label a test data point using either a face or a speech expert. The labeled data point can then be used to train the expert systems. In this way, a test data point can then be incorporated as part of the training, thus capturing the needed variation not observed during training (enrollment). An obstacle to the wide deployment of the co-training strategy is that the labeling process may be erroneous. As a result, an impostor may be mislabeled as a legitimate client, hence hampering the discriminative power of the updated model in recognizing its true claimed identity [FMJ+00].

1.3 Objectives

The objective of this study is to examine the effect of *supervised adaptation*. This provides the most optimistic scenario where all test data points are known *a priori* and only the data points of the true claimed identity are used to adapt the client model. This is opposed to *unsupervised adaptation* where the identity of the biometric sample is not known. We thus expect that the performance of supervised adaptation to be much better than the baseline non-adaptive approach and that the performance of the unsupervised adaptation to be somewhere between the two. Thus, the supervised adaptation provides the *upper bound* of the achievable performance, which is the primary objective of this study.

In general, adaptation can be performed at the model level or at the score level. At the model level, one simply updates the model parameters. At the score level, one post-

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

Figure 1: The three scenarios of the BANCA database.

processes an expert (classifier) output so that one only needs a common decision threshold despite changes in signal quality. The secondary objective of this study is to examine the effect of score level adaptation.

We validated our experiments using the existing bimodal face and speech BANCA database [BBBB+03]. This database contains three acquisition conditions, namely controlled, adverse and degraded conditions. With respect to the controlled conditions, the adverse ones are due to acquisition in a noisy environment whereas the degraded ones are due to the use of a different acquisition device. The impact of these three conditions are clearly visible in Figure 1.

1.4 Organization

This report is organized as follows: Section 2 presents the model-level and the score-level adaptation strategies. Section 3 provides some experimental evidence of our approach on the BANCA database. Finally, Section 4 concludes the report.

2 Methodology

We shall structure the discussions here into two parts: model-level adaptation and score-level adaptation.

2.1 Model-level Adaptation

The model-level adaptation will update an existing model with the new incoming data:

$$\text{update: } model, data \rightarrow \text{new model}$$

The update operation can easily be motivated by the maximum a posteriori (MAP) principle. Let θ be the parameter of a model, and $p(\theta|x)$ be the likelihood of the distribution

model. Then, the MAP principle can be summarized as maximizing the posterior probability of the model parameter (θ) given the data:

$$p(\theta^{new}|x) \propto p(\mathbf{x}|\theta^{old})p(\theta^{old}). \tag{1}$$

A peaky distribution of $p(\theta^{new}|x)$ is important in order to ensure that the parameter value can be estimated with sufficient confidence. As an example, if the data is normally distributed, $p(\mathbf{x}|\theta^{old})$ is the likelihood of a Gaussian distribution and θ consists of mean and covariance defined for the variable x. The old parameter, $\theta^{old} = \arg\max_\theta p(\theta)$, corresponds to the maximum likelihood estimate of the (mean and covariance) parameters up to the last observed sample. The updated parameters, θ^{new}, is then computed using (1). In order to see that this is a recursive formulation, let us define a sequence of ordered samples $\{x_1, x_2, \ldots, x_T\}$. Assuming that the samples are independently and identically distributed (i.i.d.), the maximum likelihood estimate of θ given samples up to T can be estimated from $p(\theta|x_1 : x_T)$, which can be calculated as follows:

$$\begin{aligned}
p(\theta|x_1 : x_T) &\propto \prod_{i=1}^{T} p(x_i|\theta)p(\theta) \\
&\propto \prod_{i=2}^{T} p(x_i|\theta)p(\theta|x_1) \\
&\propto \prod_{i=3}^{T} p(x_i|\theta)p(\theta|x_1, x_2) \\
&\propto \vdots \\
&\propto p(x_T|\theta)p(\theta|x_1 : x_{T-1})
\end{aligned} \tag{2}$$

where $p(\theta|x_1) \propto p(x_1|\theta)p(\theta)$.

Comparing (1) and (2), we observe that

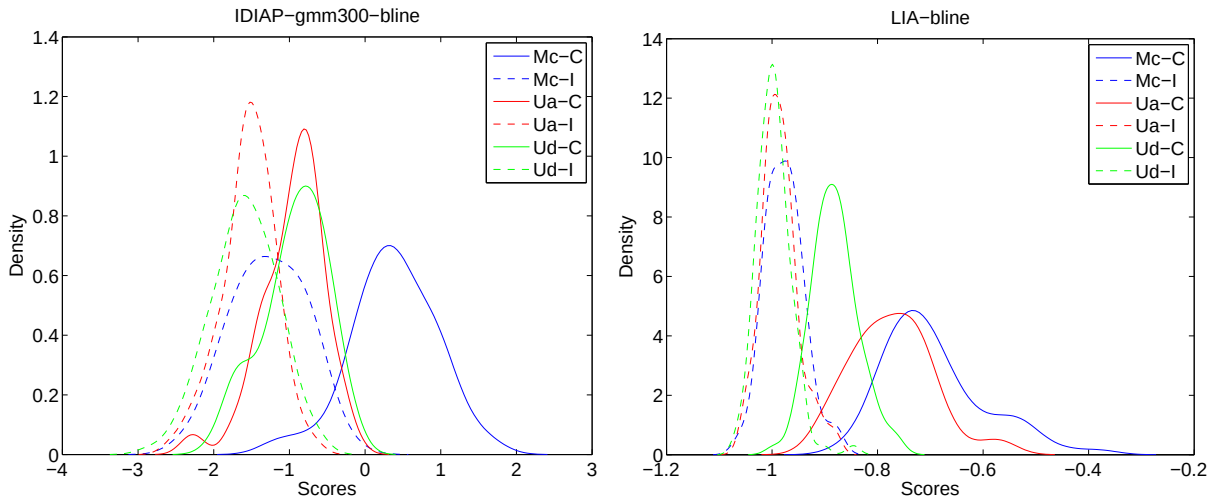
$$p(\theta^{new}|x) = p(\theta|x_1 : x_T)$$

and

$$p(\theta^{old}|x) = p(\theta|x_1 : x_{T-1}).$$

In other words, the MAP principle enables one to compute the new parameters from the old ones, given a new sample observation, under the i.i.d. assumption.

More details regarding the MAP principle applied to many parametric family of distributions can be found in classical references such as [DHS01, Bis07]. Apart from the above distributions, the mixture of Gaussian distribution, or Gaussian Mixture Model (GMM), and in particular, the GMM with MAP adaptation [RQD00] which serves as the face and speech expert that we will use (see also Section 3.1), is a realization of the MAP principle.



(a) Face (b) Speech

Figure 2: The fitted score distributions $p(y|k,Q)$ for the three scenarios $Q = \{Mc, Ua, Ud\}$ and $k \in \{\mathtt{G}, \mathtt{I}\}$, for (a) the face modality and (b) the speech modality.

2.2 Score-level Adaptation

Under changing signal quality, it is natural to ask how a particular type of acquisition conditions can affect the score. In order to do so, let the condition type be Q. Thus, for the BANCA database, Q will denote either controlled, adverse or degraded scenario (to be further described in Section 3.2), hence 3 discrete states. For other databases, the number of states in Q will have to found by using a clustering algorithm.

It is then instructive to examine the impact of Q on the class-conditional score distributions, which we denote as $p(y|k,Q)$ where k denotes the class of match scores, which can be either genuine or impostor matching, i.e., $k \in \{\mathtt{G}, \mathtt{I}\}$. The fitted distributions using Parzen windows are shown in Figures 2(a) and (b) for the face and speech modalities, respectively.

As can be observed, in both cases, the genuine and impostor score distributions are much more separated for the controlled scenario than the remaining two scenarios. This corresponds well to our expectation that the performance under the controlled scenario is better than the remaining two other scenarios.

In order to design a score normalization procedure using $p(y|k,Q)$, one can use a generative approach or a discriminative approach. In the first case, one needs to compute the following posterior probability:

$$P(\mathtt{G}|y,Q) = \frac{p(y|\mathtt{G},Q)P(\mathtt{G})}{\sum_k p(y|k,Q)P(k)} \tag{3}$$

where $P(k)$ is a prior class probability that needs to be defined.

If the state of Q is not known, it can be estimated from a set of *quality measures*, which quantify the signal quality such as contrast, brightness, reliability of face detection,

etc [PBK09, PBK10]. Let us denote the vector of quality measures by q. Then, one has to estimate the posterior $P(Q|q)$, which can be learned using a supervised or an unsupervised (clustering-based) approach. Once this quantity is available, instead of calculating (3), one can still calculate the posterior probability of a genuine class in the following way:

$$P(\mathtt{G}|y,q) = \frac{P(\mathtt{G})\sum_Q p(y|\mathtt{G},Q)P(Q|q)}{\sum_k P(k)\sum_Q p(y|k,Q)P(Q|q)} \tag{4}$$

The sum over all the states in Q is necessary because Q is not observed. This is a realization of the sum rule and is known as variable marginalization in the literature of Bayesian network [DHS01] or graphical models [Bis07]. It can be observed that by setting $P(Q = Q_*|q)$ to 1 (taking a particular realization Q_*) and $P(Q|q) = 0$ for $Q \neq Q_*$, (4) becomes (3).

Rather than using the generative approach as described in this section up to this point, we shall introduce a *discriminative* approach, avoiding the need to estimate $p(y|k,Q)$ altogether. Recalling that the objective is to estimate the posterior probability as shown in the left hand side of (3). This can be done more directly using logistic regression, i.e.,

$$P(\mathtt{G}|y,Q) = \frac{1}{1+\exp\left(-(g_Q(y))\right)} \tag{5}$$

where

$$g_Q(y) = w_1^{(Q)}y + w_0^{(Q)}$$

and $w_1^{(Q)}$ is a scaling factor and $w_0^{(Q)}$ is known as bias of logistic regression. These two parameters can be estimated using the maximum likelihood principle, i.e., maximizing (5) with respect to the two parameters, given a set of labeled training data. The realization of this principle for logistic regression is called gradient ascent (increasing the likelihood of the model given the data) [HTF01]. The logistic regression can also be seen as a single layer neuron, hence, whose parameters can be estimated using gradient descent, thus giving raise to an optimization algorithm called iterative re-weighted least square (minimizing the neuron output with respect to its target value) [Bis07]. Our implementation is based on Matlab using the gradient ascent approach.

In the case of unknown Q, just as in the case of the generative approach, i.e., (4), we can still infer Q from the quality measures q, i.e.,:

$$P(\mathtt{G}|y,q) = \sum_Q p(\mathtt{G}|y,Q)P(Q|q) \tag{6}$$

Note that (6) converges to (5) when the state of Q (i.e, the condition) is known.

2.3 Summary

In this section, we have discussed two types of adaptation strategies, i.e., model-level and score-level adaptation. The model-level adaptation consists of changing the model parameters directly, whereas the score-level adaptation aims to calibrate the scores of different

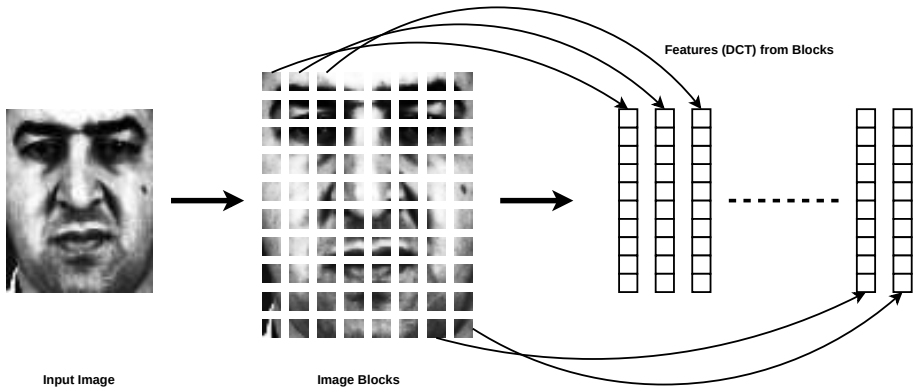


Figure 3: A flow chart of describing the extraction of feature vectors from the face image for the parts-based approach.

conditions (Q) to a common one, in terms of posterior probability of being a genuine user (client). Such a calibration is necessary when the different acquisition conditions/scenarios give raise to different genuine and impostor score distributions, as shown in Figure 2. We have described a generative as well as a discriminative approach to realize the score-level adaptation. The advantage of using a discriminative approach, as realized using logistic regression, is that one no longer needs to estimate the condition-dependent class-conditional score distributions. This reduces significantly the number of parameters that are needed to be estimated. For each version of the approaches, we also elaborate on the possibility of using quality measures to identify the conditions (Q). However, in this report, such an elaborate version is not used because the conditions are assumed to be known. In the experimental section, (5) is used as a representative method of the score-level adaptation strategy.

3 Experiments

3.1 Baseline Systems

The face and speaker verification baseline systems (also referred to as 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 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 3. 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 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 specific user/client.

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

$$y = \frac{1}{N} \sum_i \log \left\{ \frac{p(\mathbf{x}_i|\omega_j)}{p(\mathbf{x}_i|\omega_o)} \right\} \tag{7}$$

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 nose, 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 (7), 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,

[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.

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.

3.2 Database

In order to be consistent with the previous deliverables (D3.1 and D3.2), we shall use the same database, experts and similar experimental protocols[3]. 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 as 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 in comparison with the face modality.

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.

3.3 Experimental Protocols

In order to compare adaptive versus non-adaptive (baseline) systems, we had to modify the P and G experimental protocols. For each subject, there are 4 sessions of recordings per scenario, and there are 3 scenarios, i.e., controlled, adverse and degraded; these sessions are labeled 1–4, 5–8 and 9–12, respectively. The P protocol treats session 1 as the enrollment session and the remaining 2–12 as the query sessions (i.e, the test data). The G protocol, on the other hand, uses sessions 1, 5 and 9 for enrollment whereas sessions 2–4, 6–8 and 10–12 as the test data. The G protocol represents the ideal scenario where the enrollment data contains all the possible variations observable during testing (query). In comparison, in the P protocol, the experts only have the enrollment data for the controlled scenario,

[3]Modification to the protocol is imperative.

hence, is lacking in the observable variability during testing under the adverse and degraded scenarios.

The modification introduced here are as follows:

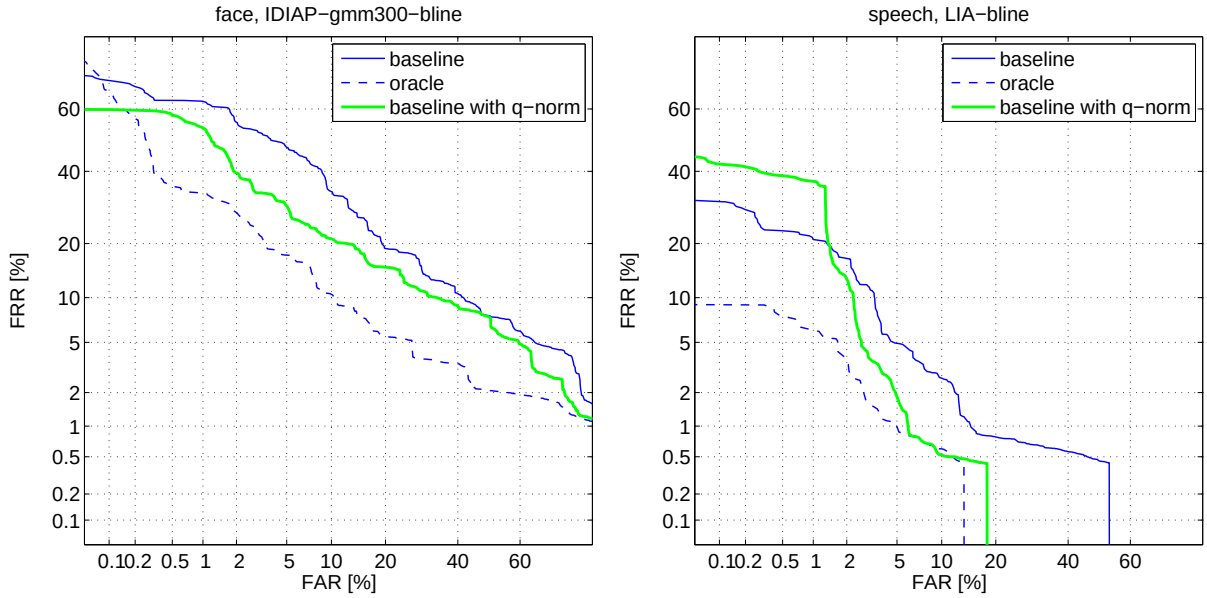
- Baseline non-adaptive system: The system is trained on session 1 but then is tested only on sessions 2–4, 6–8 and 10–12 (corresponding to the three scenarios).
- *Supervised* model-level adaptation: The system is trained on sessions 1 and is then tested on sessions 2–4, 6–8 and 10–12.
- Score-level adaptation: The system is trained on session 1 but then is tested only on sessions 2–4, 6–8 and 10–12. To realize this protocol, the output of the baseline non-adaptive systems are used (hence, its model parameters are not modified). In order to train the logistic regression, taking the output of the baseline system as input, we rely on the two-fold cross validated *score* data defined on the protocols: when the g1 (resp. g2) data set is used for testing, g2 (resp. g1) is used for training.

In essence, the above modified protocols uses the same test data sets as that of the original G protocol, but differ slightly in the training data set.

3.4 Results

This section presents the results of supervised adaptation as a representative method of the modal-level adaptation strategy as well as the logistic regression of (5) as a representative method of the score-level adaptation strategy. The results are shown in Figure 4. As can be observed, the supervised adaptation (labeled here as the oracle) can *significantly* outperform the baseline systems without any adaptation. For instance, for the face expert, the reduction is from 19.37% to 9.69% of EER, hence a reduction of 99.9%. For the speech expert, the reduction is from 4.80% to 2.29%, or reduction of 110%. In both cases, the error rate is roughly halved and the amount of training data between the baseline and the supervised adaptation is at a ratio of $1:3$.

Interestingly, the performance of the score-level adaptation (labeled here as the baseline system with quality normalization) can also improve the performance of the baseline systems, especially in regions around the Equal Error Rate (EER) where False Acceptance Rate and False Rejection Rate are roughly equal. The relative reduction in this case is 26.5% for the face expert and 42.0% for the speech expert. In comparison with the model-level supervised adaptation, the score-level adaptation is unsupervised, i.e., the subject identity of the training sessions in 5 and 9 are unknown. However, what is known here is the condition in training (which can be controlled, adverse or degraded). Although the condition in test is not required, for instance using (6), the availability of this information simplifies the procedure to (5). Between these two variations, according to [PBK09], (5) gives slightly better generalization performance. This means that rather than having to guess the condition, when this knowledge is available, it should always be used to its advantage.



(a) Face (b) Speech

Figure 4: The Performance of baseline, oracle (supervised adaptation) and score normalized system for (a) the face and (b) the speech system. The EER of these 3 systems are 19.37%, 9.69% and 15.31%, respectively, for the face modality; and are 4.80%, 2.29% and 3.38% for the speech modality. The relative improvement from the baseline to the score-normalized system is 26.5% for the face modality and 42.0% for the speech modality.

3.5 Discussions

It is worth examining why the score-level adaptation may work. Such an adaptation can be seen as a score normalization procedure:

$$\text{decision}(y) = \begin{cases} accept & \text{if } \Phi(y) > \Delta \\ reject & \text{otherwise,} \end{cases} \tag{8}$$

where Δ is the accept/reject decision threshold and $\Phi(y)$ is a score normalization procedure, which, in our case is $P(\mathtt{G}|y,Q)$. Rather than using the posterior probability, one could have used the logit transform of $P(\mathtt{G}|y,Q)$, which corresponds to $\log \frac{P(\mathtt{G}|y,Q)}{1-P(\mathtt{G}|y,Q)}$. This is a monotonic transform, hence will not have any impact on the performance, i.e., making decisions based on $P(\mathtt{G}|y,Q)$ or the log odd is exactly equivalent. However, the log odd has a nice interpretation (which is the very reason why logistic regression is very popular). We shall let $\Phi(y)$ be the log-odd:

$$\Phi(y) = \log \frac{P(\mathtt{G}|y,Q)}{1-P(\mathtt{G}|y,Q)} = w_1^{(Q)} y + w_0^{(Q)}$$

Hence, we observe that $\Phi(y)$ is a linear function of the score, controlled by Q. The decision function (8) can therefore be alternatively written as:

$$\text{decision}(y) = \begin{cases} accept & \text{if } y > \frac{\Delta - w_0^{(Q)}}{w_1^{(Q)}} \\ reject & \text{otherwise,} \end{cases} \tag{9}$$

Hence, we observe that the score-level adaptation has the implicit effect of refining the global decision threshold based on the condition Q. This implies that adaptively changing the decision threshold based on quality is beneficial.

In practice, by calculating the posterior probability, as realized using logistic regression, the parameters $w_i^{(Q)}$ for $i \in \{0, 1\}$ are completely determined by the training data and are dependent on the conditions Q. Hence, our approach is completely data driven, i.e., as long as there are additional training data points, the parameters can be optimally determined.

It remains the question of how many state of conditions Q, there are. One solution is to collect as many operational data samples as possible and then find the number of clusters of quality states by clustering each sample based on a set of designed quality measures, q. If a GMM is used as a clustering algorithm, it will approximate the following density:

$$p(q) = \sum_{Q} P(Q)p(q|Q)$$

where $p(q|Q)$ is likelihood of a Gaussian distribution and $P(Q)$ is the component prior probability. The number of components in Q can be found via cross-validation. This provides a data-driven approach to estimate the number of states (or clusters) in Q so that for each state Q one can train $P(\mathtt{G}|y, Q)$.

4 Conclusions

This deliverable examines the merit of model-level and score-level adaptation, in particular, using the supervised model-level adaptation and quality-based score normalization based on logistic regression for each case. Our experimental results show that supervised adaptation can reduce the generalization error by half whilst it demands two times more the amount of enrollment data (according to our experimental setting). On the other hand, the score-level adaptation does not require additional enrollment data (for the specific client) but requires a vast amount of the additional operational data samples that are representative of the operational scenarios. Despite the reliance of score-level adaptation on the baseline non-adaptive expert systems, the mere adaptation of score can still mitigate the effect of variation of conditions, causing significant changes in the class-conditional score distributions (as is evident in Figure 2) from one operational conditions to another. We show that this score-level adaptation procedure produces the same effect as adaptively changing the decision threshold according to the condition in an implicit manner. Although explicitly doing so is possible, our implicit approach has several advantages. Firstly, by outputting posterior probability, our approach is easier to interpret, especially when presented to a

human operator for further actions. Second, by outputting a real number (rather than a binary decision), the output can further be processed, for instance, for the purpose of bimodal fusion. Third, when the operational condition is unknown, our approach can still be used (via (6)).

Acknowledgments

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

- Josef Kittler (UNIS)
- Sébastien Marcel, Chris McCool and Niklas Johansson (IDIAP)
- Driss Matrouf (LIA)

References

[BBBB+03] E. Bailly-Baillière, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Marithoz, J. Matas, K. Messer, V. Popovici, F. Porée, B. Ruiz, and J.-P. Thiran. The BANCA Database and Evaluation Protocol. In *LNCS 2688, 4th Int. Conf. Audio- and Video-Based Biometric Person Authentication, AVBPA 2003*. Springer-Verlag, 2003.

[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.

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

[Bis07] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2007.

[BM98] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *COLT: Proceedings of the Workshop on Computational Learning Theory, Morgan Kaufmann Publishers*, pages 92–100, 1998.

[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`.

[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.

[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.

[FMJ+00] Corinne Fredouille, Johnny Mariéthoz, Cédric Jaboulet, Jean Hennebert, Chafik Mokbel, and Frédéric Bimbot. Behavior of a bayesian adaptation method for incremental enrollment in speaker verification. In *ICASSP2000 - IEEE International Conference on Acoustics, Speech, and Signal Processing*, Istanbul, Turkey, June 5–9 2000.

[HTF01] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer-Verlag, 2001.

[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.

[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.

[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.

[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.

[PBK09] N. Poh, T. Bourlai, and J. Kittler. Quality-based score normalisation with device qualitative information for multimodal biometric fusion. *IEEE Trans. on Systems, Man, and Cybernatics (part B)*, 2009. accepted for publication.

[PBK10] N. Poh, T. Bourlai, and J. Kittler. Biosecure ds2: A score-level quality-dependent and cost-sensitive multimodal biometric test bed. *Pattern Recognition Journal*, 2010. accepted for publication.

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

[PWKR09] N. Poh, R. Wong, J. Kittler, and F. Roli. Challenges and research directions for adaptive biometric recognition systems. In *LNCS 5558, Proc. of the 3rd Int'l Conf. on Biometrics*, pages 753–764, Sardinia, 2009.

[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.

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

[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.

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

[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 3.

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),

$$h(\mathbf{x}) = \ln(p(\mathbf{x} \mid \Omega^i_{client})) - \ln(p(\mathbf{x} \mid \Omega_{world})), \tag{10}$$

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:

$$\mathbf{m}_{(h,s)} = \mathbf{m} + \mathbf{D}\mathbf{y}_s + \mathbf{U}\mathbf{x}_{(h,s)}, \tag{11}$$

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 $\mathtt{I} = \tau\mathbf{D}^t\mathbf{\Sigma}^{-1}\mathbf{D}$ where τ is the *relevance factor* required in the standard MAP adaptation.

The client model is obtained by performing the decomposition of equation 11 and by retaining only the speaker dependent components:

$$\mathbf{m}_s = \mathbf{m} + \mathbf{D}\mathbf{y}_s, \tag{12}$$

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 (12), 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.

$$K(\mathcal{X}_\mathbf{s}, \mathcal{X}_{\mathbf{s}'}) = \sum_{g=1}^{M} \left(\sqrt{\alpha_g}\mathbf{\Sigma}_g^{-\frac{1}{2}}\mathbf{m}_s^g\right)^t \left(\sqrt{\alpha_g}\mathbf{\Sigma}_g^{-\frac{1}{2}}\mathbf{m}_{s'}^g\right). \tag{13}$$

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. 12, *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.