Multi-system Biometric Authentication: Optimal Fusion and User-Specific Information

THÈSE N^o 3555 (2006)

PRESENTÉE LE 31 MAI

À LA FACULTÉ SCIENCES ET TECHNIQUES DE L'INGÉNIEUR

Laboratoire de l'IDIAP

PROGRAMME DOCTORAL EN HORS PROGRAMME DOCTORAL

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

par

Norman POH

DEA d'Informatique,Université Louis Pasteur, France
et de nationalité malaisienne

acceptée sur proposition du jury:

Prof. J.R. Mosig, président du jury
Prof. H. Bourlard, Dr. S. Bengio, directeurs de thèse
Dr. A. Drygajlo, rapporteur
Prof. J. Kittler, rapporteur
rof. F. Roli, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2006

Abstract

Verifying a person's identity claim by combining multiple biometric systems (fusion) is a promising solution to identity theft and automatic access control. This thesis contributes to the state-of-the-art of multimodal biometric fusion by improving the understanding of fusion and by enhancing fusion performance using information specific to a user.

One problem to deal with at the score level fusion is to combine system outputs of different types. Two statistically sound representations of scores are probability and log-likelihood ratio (LLR). While they are equivalent in theory, LLR is much more useful in practice because its distribution can be approximated by a Gaussian distribution, which makes it useful to analyze the problem of fusion. Furthermore, its score statistics (mean and covariance) conditioned on the claimed user identity can be better exploited.

Our first contribution is to estimate the fusion performance given the class-conditional score statistics and given a particular fusion operator/classifier. Thanks to the score statistics, we can predict fusion performance with reasonable accuracy, identify conditions which favor a particular fusion operator, study the joint phenomenon of combining system outputs with different degrees of strength and correlation and possibly correct the adverse effect of bias (due to the score-level mismatch between training and test sets) on fusion. While in practice the class-conditional Gaussian assumption is not always true, the estimated performance is found to be acceptable.

Our second contribution is to exploit the user-specific prior knowledge by limiting the class-conditional Gaussian assumption to each user. We exploit this hypothesis in two strategies. In the first strategy, we combine a user-specific fusion classifier with a user-independent fusion classifier by means of two LLR scores, which are then weighted to obtain a single output. We show that combining both user-specific and user-independent LLR outputs always results in improved performance than using the better of the two.

In the second strategy, we propose a statistic called the user-specific F-ratio, which measures the discriminative power of a given user based on the Gaussian assumption. Although similar class separability measures exist, e.g., the Fisher-ratio for a two-class problem and the d-prime statistic, F-ratio is more suitable because it is related to Equal Error Rate in a closed form. F-ratio is used in the following applications: a user-specific score normalization procedure, a user-specific criterion to rank users and a user-specific fusion operator that selectively considers a subset of systems for fusion. The resultant fusion operator leads to a statistically significantly increased performance with respect to the state-of-the-art fusion approaches. Even though the applications are different, the proposed methods share the following common advantages. Firstly, they are robust to deviation from the Gaussian assumption. Secondly, they are robust to few training data samples thanks to Bayesian adaptation. Finally, they consider both the client and impostor information simultaneously.

Keywords: multiple classifier system, pattern recognition, user-specific processing

Version Abrégée

La vérification de l'identité d'une personne en combinant plusieurs systèmes biométriques est une solution prometteuse pour contrer le piratage d'identité et de contrôle d'accès. Cette thèse contribue à l'état de l'art de la fusion multimodale biométrique. Elle améliore la compréhension du mécanisme de fusion et augmente la performance de ces systèmes en exploitant l'information spécifique d'un utilisateur donné.

Cette thèse se concentre sur le problème de fusion au niveau de la sortie de plusieurs systèmes de vérification d'identité biométrique. En particulier deux différentes représentations sont utilisées comme valeur de sortie de ces sytèmes : les probabilités et le ratio de vraisemblances (Log-Likelihood Ratio, LLR). Même si en théorie, les deux représentations sont équivalentes, les LLRs sont plus facile à modèliser car leur distribution est approximativement normale. En plus, les statistiques (moyenne et covariance) pour un utilisateur donné peuvent être mieux exploitées.

Les contributions de cette thèse sont présentées en deux parties.

Tout d'abord, nous proposons un modèle pour prédire la performance optimale de fusion étant donné les statistiques dépendant des clients et des imposteurs, ainsi qu'un opérateur de fusion. Grâce à ce modèle, nous pouvons prédire la performance avec une précision acceptable, identifier les conditions qui favorisent un opérateur de fusion donné, analyser la corrélation entre les différentes fonctions de classification et analyser l'effet du biais engendré par la différence de distribution des données d'entraînement et de test. Le nouveau modèle paramétrique est fondé sur l'hypothèse que la distribution des scores, étant donnée la classe, suit une loi Gaussienne. Bien que cette hypothèse ne soit pas toujours vraie en pratique, la valeur estimée de l'erreur de performance est acceptable. Afin de pouvoir introduire des connaissances à priori pour chaque utilisateur, nous limitons l'hypothèse Gaussienne à chaque personne.

En deuxième partie, nous avons exploité cette hypothèse en utilisant deux stratégies différentes. La première consiste à combiner l'utilisation de connaissances à priori pour chaque utilisateur et celle commune à tous, par le biais de deux scores LLRs. Ceux-ci sont ensuite pondérés pour obtenir un seul score. Ce cadre générique peut être utilisé pour une ou plusieurs fonctions de classification. Nous montrons qu'en exploitant ces deux sources d'informations, l'erreur est plus petite qu'en exploitant le meilleur des deux.

La deuxième stratégie consiste à utiliser une statistique dit «F-ratio» qui indique le degré de discrimination pour un utilisateur donné en supposant l'hypothèse Gaussienne. Bien que cette statistique ressemble beaucoup au ratio de Fisher pour un problème à deux classes et le d-prime, seul le F-ratio est une fonction directement liée au taux d'erreur égal (Equal Error Rate). Nous avons exploité cette statistique dans différentes applications qui se montrent plus efficaces que les techniques classiques, à savoir, une procédure pour normaliser les scores pour chaque utilisateur, un critère pour trier les utilisateurs selon leur indice de discrimination et un nouvel opérateur qui sélectionne un sous-ensemble de systèmes pour chaque utilisateur. Bien que ces applications soient différentes, elles partagent des avantages similaires : elles sont robustes à la déviation de l'hypothèse Gaussienne, elles sont robustes à la faible disponibilité des données grâce à l'adaptation Bayesienne, enfin, elles exploitent simultanément l'information du client et des imposteurs.

Mots Clef : combinaison de plusiers fonctions de classification, reconnaissance de forme, traitement utilisateur-spécifique

Contents

List of Figures

List of Tables

Notation

Notations	Descriptions
$i \in \{1, \ldots, N\}$	index of systems from 1 to a total of N systems
$j \in \{1, \ldots, J\}$	user index from 1 to a total of J users
$y \in \mathcal{Y}$	a realization of score from a system and $\mathcal{Y}$ is a set of scores
Δ	threshold in the decision function
$k = \{C, I\}$	client or impostor class
$\mu, \boldsymbol{\mu}$	mean and mean vector
$\sigma, \boldsymbol{\Sigma}$	standard deviation and covariance matrix
γ, ω	model parameters to be tuned
$P(\cdot)$	probability
$p(\cdot)$	probability density function
$E[\cdot]$	expectation of a random variable
$Var[\cdot], \sigma$	variance of a random variable
$\mathcal{N}\big(\mathbf{y}\vert\boldsymbol{\mu}, \boldsymbol{\Sigma}\big)$	a normal (Gaussian) distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ evaluated at the point $\mathbf{y}$. The distribution is written as $\mathcal{N}\big(\boldsymbol{\mu}, \boldsymbol{\Sigma}\big)$
$\mathbf{a}'$	the transpose of the vector $\mathbf{a}$

Note that:

- No distinction is made between a variable and its realization so that $p(Y < \Delta) \equiv p(y < \Delta)$ where Y is a variable of $y \in \mathcal{Y}$. Similarly, $E_{y \in \mathcal{Y}}[Y] \equiv E[y]$.
- Subscripts and superscripts are used for conditioning a variable. The conditioning of class label k is written as a superscript, i.e., $y^k \equiv y|k$, and the user-specific conditioning (user index) is used as a subscript, i.e., $y_j \equiv y|j$.

Acronyms and Abbreviations

Acronyms	Descriptions
DCT	Discrete Cosine Transform
DET	Decision Error Trade-off
EER	Equal Error Rate
EPC	Expected Performance Curve
FAR	False Acceptance Rate
FRR	False Rejection Rate
GMM	Gaussian Mixture Model
HTER	Half Total Error Rate
LDA	Linear Discriminant Analysis
LLR	Log-Likelihood Ratio
LPR	Log-Prior Ratio
MAP	Maximum *A Posterriori*
MLP	Multi-Layer Perceptron
PCA	Principal Component Analysis
QDA	Quadratic Discriminant Analysis
ROC	Receiver's Operating Characteristic
SVM	Support Vector Machine
WER	Weighted Error Rate

Acknowledgements

I would like to thank: Dr. Samy Bengio for his constant supervision, timely response and open-mindedness to various propositions; Johnny Mariéthoz for his unbiased insights and constructive opinions; Prof. Hervé Bourlard for making extremely useful recommendations to the structure of the thesis; Prof. Hynek Hermansky, Dr. Conrad Sanderson and Dr. Samy Bengio for an important turning-point meeting about the research directions to pursue in August 2003; Julian Fierrez-Aguilar for generously sharing with me the potential research directions; the administration of IDIAP for providing an excellent computing environment; Mrs. Nadine Rousseau and Mrs. Sylvie Millius for efficiently and effectively ensuring that the administrative issues are taken care of; Romain Herault and Johnny Mariéthoz for correcting the text in French; and Dr. Conrad Sanderson for correcting parts of this thesis.

I thank the following persons for generously hosting me in their laboratories: Prof. David Zhang at the Biometric Lab of Hong Kong Polytechnic University (HKPolyU) in 2004, Dr. John Garofolo and Dr. Alvin Martin at NIST, and Prof. Anil Jain at PRIP lab, Michigan State University (MSU), both in 2005. I also thank the following persons for insightful discussions in various occasions during my visit: Dr. Arun Ross at West Virginia University; Dr. Michael Schuckers and Dr. Stephanie Schuckers at Clarkson University; Dr. Sarat Dass at MSU; Prof. Tsuhan Chen and Dr. Todd Stephenson at Carnegie Mellon University; and Dr. Ajay Kumar at HKPolyU.

Special thanks go to Prof. Jerzy Korczak at LSIIT (Laboratoire des Sciences de l'Images, de l'Informatique et de la Télédétection), Strasbourg, France, for having initiated me into the domain of pattern recognition and for having supervised me during my MSc. studies on multimodal biometric authentication during 1999-2002. I also thank University Science of Malaysia for providing a fellowship during the program.

I thank the following persons for providing precious data so much needed to study the subject of fusion: all the members of the verification group at IDIAP, especially, Fabien Cardinaux, Sébastien Marcel, Christine Marcel, Guillaume Heusch, Yan Rodriguez for the match scores of BANCA and XM2VTS; all the members of PRIP lab, MSU, especially Chenyoo Roo, Yi Chen, Yongfang Zhu and Xiaoguang Loo and for generously sharing fingerprint, iris and 3D face match scores; all the members of speech processing group at NIST, especially Mark Przybocki for preparing a subset of NIST evaluation scores; and Dr. Ajay Kumar for providing palmprint features.

I thank my mother Geraldine Tay for helping me with the arrival of my youngest son Bernard while I was in the midst of writing my thesis. Special thanks go to my wife Wong Siew Yeung for her constant moral support, and my sons François and Bernard for coloring my life.

Last but not least, I thank the following people for making my stay memorable in Switzerland: all the members of Dejeuné Priere, especially Alain Léger and Sophie Bender, all the members of Solitude Myriam, especially Anne-Marie Soudan, and all my colleagues at IDIAP.

Norman Poh

Martigny, May 2006.

Chapter 1

Multi-system Biometric Authentication

1.1 Problem Definition

Biometric authentication is a process of verifying an identity claim using a person's behavioral or physiological characteristics [62]. Biometric authentication offers many advantages over conventional authentication systems that rely on possessions or special knowledge, e.g., passwords. It is convenient and is widely accepted in day-to-day applications. Typical scenarios are access control and authentication transaction. This field is evolving fast due to the desire of governments to provide a better homeland security and due to the market demand to protect privacy in various forms of transactions.

Authentication versus Identification

This thesis is about biometric *authentication* (also known as verification) and not about biometric *identification*. In the latter, there is no identity claim, but rather the goal of the system is to output the most probable identity. If there are J persons in the database, then J matchings are needed. In a closed set identification, this task is to forcefully classify a biometric sample as one of the J known persons. In an open set identification, the task is to classify the sample as either one of the J persons or an unknown person. In some applications, particular in access control with a limited population size, biometric authentication is operated in the open set identification mode. In this scenario, an authorized user simply presents his/her biometric sample prior to accessing a secured resource, *without* making any identity claim [86][1]. Hence, in terms of applications, there needs no clear distinction between authentication and identification, i.e., techniques developed in one application scenario can be applied to another.

Error Rates

Upon presentation of a biometric sample, a system should grant access (if the person is a client/user) or reject the request (if the person is an impostor). In general terms, this decision is made by comparing the system output with an *operating threshold*. In this process, two types of error can be committed: falsely rejecting a genuine user or falsely accepting an impostor. The error rates are respectively called False Rejection Rate (FRR) and False Acceptance Rate (FAR). These two errors are important measures to assess the system performance which is visualized using a Detection Error Trade-off (DET) curve. A special point called Equal Error Rate (EER), where FAR=FRR, is also commonly used for application independent assessment.

Desired Operational Characteristics of Biometric Authentication

It is desirable that biometric authentication be performed *automatically, quickly, accurately and reliably*. Using multimedia sensors and ever increasingly powerful computers, the first two criteria can certainly be

[1] In this case, the original authentication system has to be modified so that the accept/reject decision is not made for each enrolled user. This is because there could be multiple accept decisions.

fulfilled. However, *accuracy* and *reliability* are two issues not fully resolved. Due to sensor technologies and external manufacturing constraints, no single biometric trait can achieve a 100% authentication performance. By *accuracy*, we mean that both FAR and FRR have to be reduced. Often, decreasing one error type by changing the operational threshold will only increase the other error type. Hence, in order to truly improve the accuracy, there must be a *fundamental improvement*. By *reliability*, we mean that the *same* result in terms of score should be expected each time a system processes a biometric sample during testing.

The Challenges in Biometric Authentication

Person authentication is a difficult problem because of the following properties:

- **Unbalanced classification task:** At least in a typical experimental setting, the number of genuine (client) attempts is much smaller than that of impostor attempts[2].
- **Unbalanced risk:** Depending on applications, the *cost* of falsely accepting an impostor and that of falsely rejecting a client can differ by one or two orders of magnitude.
- **Scarce training data:** At the initial (enrollment) phase, a biometric system is allowed to have very few biometric samples (less than four or so; in order not to annoy the user). Building a statistical model or a feature template is thus a challenging machine-learning problem.
- **Vulnerability to noise:** It is known that biometric samples are vulnerable to “noise”. Examples are, but not limited to, (i) occlusion, e.g., glasses occluding a face image; (ii) environmental noise, e.g., view-based capturing devices are particularly susceptible to change of illumination, and speech is susceptible to external noise sources [118] as well as distortion by the transmission channel; (iii) user’s interaction with the device, e.g, non-frontal face [128]; (iv) the deforming nature of biometrics, as beneath physiological biometric traits are often muscles or living tissues that are subject to minor changes over both short and long time-span; (v) detection algorithms, e.g., inaccurate face detectors [147]; and (vi) the ageing effect [46] in the sense that the duration can span from days (e.g., growth of beards and mustaches for face recognition) or weeks (e.g., hair) to years (e.g., appearance of wrinkles). Increasing the system reliability implies decreasing the influence of these noise sources.

Multi-System Biometric Authentication

The system accuracy and reliability can be increased by combining two or more biometric authentication systems. According to a yet-to-published standard report (ISO 24722) entitled “Technical Report on Multi-Modal and other Biometric Fusion” [149], these approaches can be any of the following types:

- **Multimodal:** Different sensors capturing different body parts
- **Multi-sensor:** Different sensors capturing the same body part
- **Multi-presentation:** Several sensors capturing several similar body parts, e.g., ten-fingerprint biometric system
- **Multi-instance:** The same sensor capturing several instances of the same body part
- **Multi-algorithmic:** The same sensor is used but its output is proposed by different feature extraction and classifier algorithms

This thesis concerns fusion of any of these types, i.e., a *multi-system* biometric authentication. For this reason, the term “multi-system” was used in this thesis title. In the general pattern recognition problem, our chosen approach can also be called a *Multiple Classifier System* (MCS). As this thesis focuses on the above-mentioned approaches, the classical ensemble algorithms such as bagging, boosting and error-correction output-coding [31] which rely on *common features* will not be discussed. This issue was examined elsewhere, e.g., [95].

[2] Such prior probabilities are unknown in real applications and are often set to be equal.

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 [14], 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 [74].

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* [11, Sec. 8.6]. In brief, it states that combined information (feature) 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.

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, e.g., combining the coefficients of Principal Component Analysis (PCA) and that of Linear Discriminant Analysis (LDA). Under such 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.

1.2 Motivations

Combining several systems has been investigated elsewhere, e.g., in general pattern recognition [138]; in applications related to audio-visual speech processing [76, Chap. 10] [77, 19]; in speech recognition – examples of methods are multi-band [17], multi-stream [38, 55], front-end multi-feature [136] approaches and the union model [85]; in the form of ensemble [13]; in audio-visual person authentication [127]; and, in multi-biometrics [125, 88] (and references herein), among others. In fact, one of the earliest works addressing multimodal biometric fusion was reported in 1978 [39]. Therefore, biometric fusion has a history of nearly 30 years. Admittedly, the subject of classifier combination is somewhat mature. However, below are some motivations for yet another thesis on the topic:

- **Justification of why fusion works:** Although this topic has been discussed elsewhere [57, 67, 68, 133], there is still a lack of theoretical understanding, particularly with respect to *correlation* and *relative strength* among systems in the context of fusion. While these two factors are well known in regression problems [13], they are not well-defined in classification problems [135]. As a result, many "diversity" measures exist while no one measure is a satisfactory predictor of the fusion performance – they are too weakly correlated with the fusion performance and are highly biased.

- **User-induced variability:** When biometric authentication was first used for biometric authentication [48], it was observed that scores from the output of a system are highly variable from one user to another. 17 years later, this phenomenon was statistically quantified [33]. As far as user-induced variability is concerned, several issues need to be answered: whether this phenomenon exists in *all* biometric systems or it is limited to the speaker verification systems; methods to mitigate this phenomenon; and to go one step further, methods to consider the claimed user identity in order to improve the overall performance.

- **Different modes of fusion:** The *de facto* approach to fusion is by considering the output of all sub-systems [125] (and references herein). However, in a practical application, e.g., [86], one rarely uses all the sub-systems simultaneously. This suggests that an efficient and accurate way of selecting sub-systems to combine would be beneficial.

- **On the use of chimeric users:** Due to lack of real large multimodal biometric datasets and privacy concerns, the biometric trait of a user from a database is often combined with another different biometric trait of yet another user, thus creating a so-called *chimeric user*. Using a chimeric database can thus effectively generate a multimodal database with a large number of users, e.g., up to a thousand [137]. While this practice is commonly used in the multimodal literature, e.g., [44, 124, 137] among others, it was questioned whether this was a right thing to do or not during the 2003 Workshop on Multimodal User Authentication [36]. While the privacy problem is indeed solved using chimeric users, it is still an open question of how such chimeric database can be used effectively.

1.3 Objectives

The objective of this thesis is two-fold: to provide a better understanding of fusion and to exploit the claimed identity in fusion.

Due to the first objective, proposing a new specialized fusion classifier is not the main goal but a consequence of a better understanding of fusion. To ensure *systematic* improvement, whenever possible, we used a relatively large set of fusion experiments, instead of one or two case studies as often reported in the literature. For example in this thesis as few as 15 experiments are used. In our published paper, e.g., [113], as many as 3380 were used. None of the experiments used are chimeric databases (unless constructed specifically to study the effect of chimeric users). Our second objective, on the other hand, deals with how the information specific to a user can be exploited. Consequently, novel strategies have to be explored.

1.4 Original Contributions Resulting From Research

The original contributions resulting from the PhD research can be grouped in the following ways:

1. **Fusion from a parametric perspective:** Several studies [57, 67, 68, 133] show that combining several system outputs improves over (the average performance of) the baseline systems. However, the justifications are not directly related to the reduction of classification performance, e.g., EER, FAR and FRR. Furthermore, one or more unrealistic and simplifying assumptions are often made, e.g., independent system outputs, common class-conditional distributions across system outputs and common distribution across (client and impostor) class labels. We propose to model scores to be combined using a class-conditional multivariate Gaussian (one for the client scores; the other for the impostor scores). This model is referred to as a "parametric fusion model" in this thesis. Although being simple, this model does not make any of the three assumptions just stated above. A well known Bayes error bound (or the upper bound of EER) based on this model is called the Chernoff bound [35].

   Our original idea is to derive the *exact* EER (instead of its bound) given the parametric fusion model and given a particular fusion operator thanks to a derived statistic called the "F-ratio" [103]. Although in practice the Gaussian assumption inherent in the parametric fusion model is not always true, the error of the estimated EER is acceptable in practice. We used the F-ratio to show the reduction of classification error due to fusion in [103], to study the effect of correlation of system outputs in [109], to predict fusion performance in [102] and to compare the performance of commonly used fusion operators (e.g, $\min$, $\max$, mean and weighted sum) in [107].

2. **On exploiting user-specific information:** While assuming that class conditional scores are Gaussian is somewhat naive, this approach is much more acceptable when one makes such an assumption on the user-specific scores, where the client (genuine) scores are scarce. Two different approaches are proposed to exploit user-specific information in fusion.

   The first approach, called a *user-specific compensation framework* [105], linearly combines the outputs of both user-specific and user-independent fusion classifiers. This framework also generalizes to a user-specific score normalization procedure when only a single system is involved. The advantage of this framework is that it compensates for the possibly unreliable but still useful user-specific fusion classifier.

   The second approach makes use of the *user-specific F-ratio*, which is in the following techniques:

   - A novel user-specific score normalization procedure called F-norm.
   - A user-specific performance criterion to rank users according to their ease of recognition.
   - A novel user-specific fusion operator called an "OR-Switcher" which works by selecting only a subset of system to combine on a per person basis.

   These techniques can be found in our publications [108, 115, 112], respectively. Although the applications are different, they all are related to F-norm and hence share the following properties:

   - Robustness to the Gaussian assumption.
   - Robustness to extremely few genuine accesses via Bayesian adaptation, which is a unique advantage not shared by existing methods in user-specific score/threshold normalization, e.g. [18, 48, 52, 64, 75, 92, 126].
   - Client-impostor centric – making use of both the genuine and impostor scores.

3. **Exploring different modes of score-level fusions:** We also propose several new paradigms to fusion, namely:

   - A novel multi-sample multi-source approach – whereby multiple samples of different biometric modalities are considered.
   - Fusion with virtual samples by random geometric transformation of face images – whereby the novelty lies on applying virtual samples during test as opposed to during training.
   - A robust multi-stream (multiple speech feature representations) scheme. This scheme relies on a fusion classifier that is implemented via a Multi-Layer Perceptron and takes the outputs of the speaker verification systems. While being trained with artificial white noise, the fusion classifier is shown to be empirically robust to different realistic additive noise types and levels.

   These three subjects can be found in our publications [114, 116, 100], respectively.

4. **On incorporating both user-specific and quality information sources:** Several studies on fusion [10, 44, 129, 141] as well as on other biometric modalities, e.g., speech [49] and fingerprint [21, 134], iris [20] and face [70], have demonstrated that quality index, also known as confidence, is an important information source. In the mentioned approaches, a quality index is derived from the features or raw biometric data. We propose two ideas to improve the existing techniques. The first one aims at directly deriving the quality information from the score, based on the concept of margin used in boosting [47] and Support Vector Machine (SVM) [146], [26]. The second one aims at combining user-specific and quality information in fusion using a discriminative approach. The resultant techniques based on these two ideas were published in in [110] and [111][3], respectively.

5. **On the merit of chimeric users:** To the best of our knowledge, no prior work is done on the merits of chimeric users in experimentation. We examined this issue from two perspective: whether or not the performance measured on a chimeric database is a good predictor of that measured on a real-user

[3]This paper is the winner the best student poster award in Int'l Conf. on Audio- and Video-Based Biometric Person Authentication (AVBPA2005) for contribution on "biometric fusion".

database; and whether or not a chimeric database can be exploited to *improve* the generalization performance of a fusion operator on a real-user database. Based on a considerable amount of empirical biometric person authentication experiments, we conclude that the answer is unfortunately "no" to the first question[4] and no statistical significant improvement or degradation to the second question. However, considering the lack of real large multimodal database, it is still useful to construct a trainable fusion classifier using a chimeric database. These two investigations were published in [104] and [113], respectively.

6. **On performance prediction/extrapolation:** Due to user-induced variability, the system performance is often database-dependent, i.e., the system performance differs from one database to the other. Working towards this direction, we address two issues: establishing confidence interval of a DET curve such that the effect due to different composition of users is taken into account [117]; and modeling the performance change (over time) on a per user basis so as to provide an explanation to the trend of the system performance.

7. **Release of a score-level fusion benchmark database and tools:** Motivated by the fact that multi biometric fusion score-level is an important subject and yet such a benchmark database does not exist, the XM2VTS fusion benchmark dataset was released to the public[5]. Together with this database come the state-of-the-art evaluation tools such as DET (Detection Error Trade-off), ROC (Receiver's Operating Characteristic) and EPC (Expected Performance Curve) curves. The work was published in [106].

The above contributions (except topic 7) can be divided into two categories, i.e., user-independent processing (topics 1, 3 and 5) and user-specific processing (topics 2, 4 and 6). User-specific processing, as opposed to user-independent processing, takes into account the label of the claimed identity for a given access request, e.g., user-specific fusion classifier, user-specific threshold and user-specific performance estimation. Topics 1 and 2 are the *most representative* and also the *most important* subject in its category. We therefore give much more emphasis on these two topics.

1.5 Publications Resulting From Research

The publications resulting from this thesis are as follows:

1. **Fusion from a parametric perspective.**

   - N. Poh and S. Bengio. Why Do Multi-Stream, Multi-Band and Multi-Modal Approaches Work on Biometric User Authentication Tasks? In *IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, pages vol. V, 893–896, Montreal, 2004.
   - N. Poh and S. Bengio. How Do Correlation and Variance of Base Classifiers Affect Fusion in Biometric Authentication Tasks? *IEEE Trans. Signal Processing*, 53(11):4384–4396, 2005.
   - N. Poh and S. Bengio. Towards Predicting Optimal Subsets of Base-Experts in Biometric Authentication Task. In *LNCS 3361, 1st Joint AMI/PASCAL/IM2/M4 Workshop on Multimodal Interaction and Related Machine Learning Algorithms MLMI*, pages 159–172, Martigny, 2004.
   - N. Poh and S. Bengio. EER of Fixed and Trainable Classifiers: A Theoretical Study with Application to Biometric Authentication Tasks. In *LNCS 3541, Multiple Classifiers System (MCS)*, pages 74–85, Monterey Bay, 2005.

2. **On exploiting user-specific information.**

   - N. Poh and S. Bengio. F-ratio Client-Dependent Normalization on Biometric Authentication Tasks. In *IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, pages 721–724, Philadelphia, 2005.

[4] This implies that if one fusion operator outperforms another fusion operator on a chimeric database, one *cannot guarantee* that the same observation is repeatable in a true multimodal database of the same size.

[5] Accessible at http://www.idiap.ch/∼norman/fusion

- N. Poh, S. Bengio, and A. Ross. Revisiting Doddington's Zoo: A Systematic Method to Assess User-Dependent Variabilities. In *Workshop on Multimodal User Authentication (MMUA 2006)*, Toulouse, 2006.
- N. Poh and S. Bengio. Compensating User-Specific Information with User-Independent Information in Biometric Authentication Tasks. Research Report 05-44, IDIAP, Martigny, Switzerland, 2005.

3. **On exploring different modes of score-level fusions.**

- 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.
- N. Poh, S. Bengio, and J. Korczak. A Multi-Sample Multi-source Model for Biometric Authentication. In *IEEE International Workshop on Neural Networks for Signal Processing (NNSP)*, pages 275–284, Martigny, 2002.
- N. Poh, S. Marcel, and S. Bengio. Improving Face Authetication Using Virtual Samples. In *IEEE Int'l Conf. Acoustics, Speech, and Signal Processing*, pages 233–236 (Vol. 3), Hong Kong, 2003.
- N. Poh and S. Bengio. Noise-Robust Multi-Stream Fusion for Text-Independent Speaker Authentication. In *The Speaker and Language Recognition Workshop (Odyssey)*, pages 199–206, Toledo, 2004.

4. **On incorporating both user-specific and quality information sources.**

- N. Poh and S. Bengio. Improving Fusion with Margin-Derived Confidence in Biometric Authentication Tasks. In *LNCS 3546, 5th Int'l. Conf. Audio- and Video-Based Biometric Person Authentication (AVBPA)*, pages 474–483, New York, 2005.
- N. Poh and S. Bengio. A Novel Approach to Combining Client-Dependent and Confidence Information in Multimodal Biometric. In *LNCS 3546, 5th Int'l. Conf. Audio- and Video-Based Biometric Person Authentication (AVBPA 2003)*, pages 1120–1129, New York, 2005 ((winner of the Best Student Poster award)).

5. **On the merit of chimeric users.**

- N. Poh and S. Bengio. Can Chimeric Persons Be Used in Multimodal Biometric Authentication Experiments? In *LNCS 3869, 2nd Joint AMI/PASCAL/IM2/M4 Workshop on Multimodal Interaction and Related Machine Learning Algorithms MLMI*, pages 87–100, Edinburgh, 2005.
- N. Poh and S. Bengio. Using Chimeric Users to Construct Fusion Classifiers in Biometric Authentication Tasks: An Investigation. In *IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Toulouse, 2006.

6. **Other subjects.**

- N. Poh, A. Martin, and S. Bengio. Performance Generalization in Biometric Authentication Using Joint User-Specific and Sample Bootstraps. IDIAP-RR 60, IDIAP, Martigny, 2005.
- 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.
- N. Poh, C. Sanderson, and S. Bengio. An Investigation of Spectral Subband Centroids For Speaker Authentication. In *LNCS 3072, Int'l Conf. on Biometric Authentication (ICBA)*, pages 631–639, Hong Kong, 2004.

1.6 Outline of Thesis

This thesis is divided into two parts which correspond to two major contributions. Chapter 2 is devoted to explaining the common databases and evaluation methodologies used in both parts of thesis.

Part I focuses on the score-level user-independent fusion. It contains two chapters. Chapter 3 reviews the state-of-the-art techniques in score-level fusion. Our original contribution, to be presented in Chapter 4, is on providing a better understanding based on the class-conditional Gaussian assumption of scores to be combined – the so-called *parametric fusion model*.

Part II focuses on user-specific fusion. All the discussions in Part I can directly be extended to Part II by conditioning the parametric fusion model on a specific user. For this reason, Part I and II are complementary. Part II contains three chapters. Chapter 5 is the first survey written on the subject of *user-specific processing*. The next two chapters are our original contributions. Chapter 6 proposes a compensation scheme that balances between user-specific and user-independent fusion. Chapter 7 presents a user-specific fusion classifier as well as a user-specific normalization procedure based on F-norm.

Finally, Chapter 8 summarizes the results obtained so far and outlines promising future research directions.

Chapter 2

Database and Evaluation Methods

This chapter is divided into two sections: Section 2.1 describes the databases used in this thesis and Section 2.2 describes the adopted evaluation methodologies. The second section deals with issues such as threshold selection, performance evaluation, visualization of pooled performance (from several experiments) and significance test.

2.1 Database

There are currently many multimodal person authentication databases that are reported in the literature, for examples (but not limited to):

- BANCA [5] – face and speech modalities[1].
- XM2VTS [78] – face and speech modalities[2].
- VidTIMIT database [25] – contains face and speech modalities[3].
- BIOMET [15] – contains face, speech, fingerprint, hand and signature modalities.
- NIST Biometric Score Set – contains face and fingerprint modalities[4].
- MYCT [90] – ten-print fingerprint and signature modalities[5].
- UND – face, ear profile and hand modalities acquired using visible, Infrared-Red and range sensors at different angles[6].
- FRGC – face modality captured using camera at different angles and range sensors in different controlled or uncontrolled settings[7].

However, not all these databases are true multi-biometric modalities, i.e., from the same user. To the best of our knowledge, BANCA, XM2VTS, VidTIMIT, FRGC and NIST are true multimodal databases whereas the rest are *chimeric* multimodal databases. A chimeric user is composed of at least two biometric modalities originated from two (or more) individuals. BANCA and XM2VTS are preferred because:

- They are publicly available.

[1] http://www.ee.surrey.ac.uk/banca
[2] http://www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb
[3] http://users.rsise.anu.edu.au/∼conrad/vidtimit
[4] http://www.itl.nist.gov/iad/894.03/biometricscores/bssr1_contents.html
[5] http://turing.ii.uam.es/bbdd_EN.html
[6] http://www.nd.edu/∼cvrl/UNDBiometricsDatabase.html
[7] http://www.frvt.org/FRGC

Table 2.1: The Lausanne and fusion protocols of the XM2VTS database. Numbers quoted below are the number of samples.

Data sets	Lausanne Protocols		Fusion Protocols
	LP1	LP2	
LP Train client accesses	3	4	NIL
LP Eval client accesses	600 (3 × 200)	400 (2 × 200)	Fusion dev
LP Eval impostor accesses	40,000 (25 × 8 × 200)		Fusion dev
LP Test client accesses	400 (2 × 200)		Fusion eva
LP Test impostor accesses	112,000 (70 × 8 × 200)		Fusion eva

- They come with well defined experimental configurations, called *protocols*, which define clearly the training and test sets such that different algorithms can be benchmarked.
- They contain behavioral and physiological biometric traits.

2.1.1 XM2VTS Database and Its Score-Level Fusion Benchmark Datasets

The XM2VTS database [83] contains synchronized video and speech data from 295 subjects, recorded during four sessions taken at one month intervals. On each session, two recordings were made, each consisting of a speech shot and a head shot. The speech shot consisted of frontal face and speech recordings of each subject during the recital of a sentence.

The Lausanne Protocols

The 295 subjects were divided into a set of 200 clients, 25 evaluation impostors and 70 test impostors. There exists two configurations or two different partitioning approaches of the training and evaluation sets. They are called Lausanne Protocol I and II, denoted as LP1 and LP2. One can distinguish three data sets, namely train, evaluation and test sets (labeled as "Train", "Eval" and "Test", respectively). For each user, these three sets contain $(3, 3, 2)$ samples for LP1 and $(4, 2, 2)$ for LP2. The training set is used *uniquely* to build a user-specific model. Any hyper-parameter of the model can be tuned on the Eval set. Thus the Eval set is *reserved* uniquely as a validation set. An *a priori* threshold has to be calculated on the Eval set and this threshold is used when evaluating the system performance on the Test set in terms of FAR and FRR (to be described in Section 2.2). Note that in both protocols, the test set remains the same. Table 2.1 is the summary of the LP1 and LP2 protocols. The last column of Table 2.1 shows the fusion protocol. Note that as long as fusion is concerned, only two types of data sets are available, namely fusion development and fusion evaluation sets[8]. These two sets have $(3, 2)$ samples for LP1 and $(2, 2)$ samples for LP2, respectively, on a per user basis. More details about the XM2VTS database can be found in [78].

The Score-Level Fusion Datasets

As for the score fusion datasets, we collected match scores of seven face systems and six speech systems. This data set is known as the "XM2VTS score-level fusion benchmark dataset" [106][9]. The label assigned to each system (Table 2.2) has the format Pn:m where n denotes the protocol number (1 or 2) and m denotes the order in which the respective system is invoked. For MLP-based classifiers, their associated class-conditional scores have a skewed distribution due to the use of the logistic activation function in the output layer. Note that LP1:6 and LP1:8 are MLP systems with hyperbolic tangent output whereas LP1:7 and LP1:9 are the same systems but whose outputs are transformed into LLR by using an inverse hyperbolic

[8]Note that at the fusion level, only scores are available. The fusion *development* set is derived from the LP Eval set whereas the fusion *evaluation* set is derived from the LP Test set. The term "development" is consistently referred to as the training set; and "evaluation" as the test set.

[9]Available at http://www.idiap.ch/∼norman/fusion. There are nearly 100 downloads at the time of this thesis publication.

Table 2.2: The characteristics of 12 (+2 modified) systems taken from the XM2VTS benchmark fusion database.

Labels	Modalities	Features	Classifiers
P1:1	face	DCTs	GMM
P1:2	face	DCTb	GMM
P1:3	speech	LFCC	GMM
P1:4	speech	PAC	GMM
P1:5	speech	SSC	GMM
P1:6	face	DCTs	MLP
P1:7	face	DCTs	MLPi
P1:8	face	DCTb	MLP
P1:9	face	DCTb	MLPi
P1:10	face	FH	MLP
P2:1	face	DCTb	GMM
P2:2	speech	LFCC	GMM
P2:3	speech	PAC	GMM
P2:4	speech	SSC	GMM

MLPi denotes the output of MLP converted to LLR using inverse hyperbolic tangent function. P1:6 and P1:7 (resp. P1:8 and P1:9) are the *same* systems except that the scores of the latter are converted.

tangent function. This is done to ensure that the scores are once again linear. More explanation about the motivation and the post-processing technique can be found in Section 3.3.2[10].

The Participating Systems in the Fusion Datasets

Note that each system in Table 2.2 can be characterized by a feature representation and a classifier. All the speech systems are based on the state-of-the-art Gaussian Mixture Models (GMMs) [121]. They differ only by their feature representations, namely Linear Frequency Cepstral Coefficients (LFCC) [119], Phase-AutoCorrelation (PAC) [59] and Spectral Subband Centroids (SSC) [91, 118]. These feature representations are selected such that they exhibit different degree of tolerance to noise. Highly tolerant feature representation performs worse in clean conditions. The face systems are based on a downsized raw Face images concatenated with color Histogram information (FH) [81] and Discrete Cosine Transform (DCT) coefficients [131]. The DCT procedure operates with two sizes of image block, i.e., small (s) or big (b), and are denoted by DCTs or DCTb, respectively. Hence, the matching process is local as opposed to the holistic matching approach. Both the face and speech systems are considered the-state-of-the-art systems in this domain. Details of the systems can be found in [106].

2.1.2 BANCA Database and Score Datasets

The BANCA database [5] is the principal database used in this paper. It has a collection of face and voice biometric traits of up to 260 persons in 5 different languages. We used only the English subset, containing only a total of 52 persons; 26 females and 26 males. The 52 persons are further divided into two sets of users, which are called g1 and g2, respectively. Each set of users contains 13 males and 13 females. According to the experimental protocols, when g1 is used as a development set (to build the user's template/model), g2 is used as an evaluation set. Their roles are then switched. In this thesis, g1 is used as a development set; and g2 an evaluation set.

[10] In some fusion experiments, especially in user-specific fusion, P1:10 is excluded from study because for some reasons, it contains scores more than 1 or less than -1 (which should not in theory!). When converting these border scores using the inversion process, they result in overflow and underflow. While we tried different ways to handle this special case, using P1:10 only complicates the analysis without bring additional knowledge.

Table 2.3: Usage of the seven BANCA protocols (C: client, I: impostor). The numbers refer to the ID of each session.

Test Sessions	Train Sessions			
	1	5	9	1,5,9
C: 2-4 I: 1-4	Mc			
C: 6-8 I: 5-8	Ud	Md		
C: 10-12 I: 9-12	Ua		Ma	
C: 2-4,6-8,10-12 I: 1-12	P			G

The BANCA Protocols

There are altogether 7 protocols, namely, Mc, Ma, Md, Ua, Ud, P and G, each simulating matched control, matched adverse, matched degraded, uncontrolled adverse, uncontrolled degraded, pooled and grant test, respectively. For protocols P and G, there are 312 client accesses and 234 impostor accesses. For all other protocols, there are 78 client accesses and 104 impostor accesses. Table 2.3 describes the usage of different sessions in each configuration. Note that the data is acquired over 12 sessions and spanned over several months.

The Score Files

For the BANCA score data sets, there are altogether 1186 score files containing single modality experiments as well as fusion experiments, thanks to a study conducted in [80][11]. The classifiers involved are Gaussian Mixture Models (GMMs) (514 experiments), Multi-Layer Perceptrons (MLPs) (490 experiments) and Support Vector Machines (SVMs) (182 experiments).

Differences Between BANCA and XM2VTS

The BANCA database differs from the XM2VTS database in the following ways:

- BANCA contains more realistic test scenarios.
- The population on which the hyper-parameter of a baseline system is tuned is different for the development and evaluation sets, whereas in XM2VTS the genuine users are the same (the impostor populations are different in both cases). In both cases, there are no “inter-template” match scores, i.e., match scores resulting from comparing the biometric data of two genuine users, which are used frequently in databases with identification setting.
- The number of client and impostor accesses are much more balanced in BANCA than in XM2VTS.

Pre-defined BANCA Fusion Tasks

We selected a subset of BANCA systems to constitute a set of fusion tasks. These systems are from University of Surrey (2 face systems), IDIAP (1 speaker system), UC3M (1 speaker system) and UCL (1 face system)[12]. The specific score files used are as follow:

- `IDIAP_voice_gmm_auto_scale_33_200`
- `SURREY_face_svm_auto`

[11] Available at “ftp://ftp.idiap.ch/pub/bengio/banca/banca_scores”
[12] Available at “ftp://ftp.idiap.ch/pub/bengio/banca/banca_scores”

- SURREY_face_svm_man
- UC3M_voice_gmm_auto_scale_34_500
- UCL_face_lda_man

for each of the 7 protocols. By combining each time two systems from the same protocol, one can obtain 10 fusion tasks, given by ${}^5C_2$ (5 "choose" 2). This results in a total of 70 experiments for all 7 protocols.

These experiments can be divided into two types: multimodal fusion (fusion of two different modalities, i.e, face and speech systems) and intramodal fusion (of two face systems *or* two speech systems). We expect multimodal fusion to be less correlated while intramodal fusion to be more correlated. This is an important aspect so that both sets of experiments will cover a large range of correlation values.

2.1.3 NIST Speaker Database

The NIST yearly speaker evaluation plans [89] provide many data sets for examining different issues that can influence the performance of a speaker verification system, notably with respect to handset types, transmission channels and speech duration [148, Chap. 8]. The 2005 (score) datasets are obtained from 24 systems that participated in the evaluation plan. These scores are resulted from using testing the 24 systems on the speech test data sets as defined by the NIST experimental protocols. However, for the purpose of fusion, there exists no fusion protocol so we define one that suits our needs.

In compliance to the NIST's policy, the identity of the participants are concealed, so are the systems which the participants submitted. Most systems are based on Gaussian Mixture Models (GMMs) but there exists also Neural Network-based classifiers and Support Vector Machines. A few systems are actually combined systems using different levels of speech information. Some systems combine different type of classifiers but each classifier uses the same feature sets. We use a subset of this database which contains 124 users.

2.2 Performance Evaluation

2.2.1 Types of Errors

A fully operational biometric system makes a decision using the following *decision function*:

$$\text{decision}(\mathbf{x}) = \begin{cases} accept & \text{if } y(\mathbf{x}) > \Delta \\ reject & \text{otherwise,} \end{cases} \tag{2.1}$$

where Δ is a threshold and $y(\mathbf{x})$ is the output of the underlying system supporting the hypothesis that the extracted biometric feature of the query sample, $\mathbf{x}$, belongs to the *target* client, i.e., whose identity is being claimed. Note that in this case, the decision is *independent* of any identity claim. A more thorough discussion of user-specific decision making can be found in Section 5. For the sake of clarity, we write y instead of $y(\mathbf{x})$. The same convention applies to all variables derived from y. Because of the accept-reject outcomes, the system may make two types of errors, i.e., false acceptance (FA) and false rejection (FR). The normalized versions of FA and FR are often used and called False Acceptance Rate (FAR) and False Rejection Rate (FRR)[13], respectively. They are defined as:

$$\text{FAR}(\Delta) = \frac{\text{FA}(\Delta)}{N^I} , \tag{2.2}$$

$$\text{FRR}(\Delta) = \frac{\text{FR}(\Delta)}{N^C} . \tag{2.3}$$

where FA and FR count the number of FA and FR accesses, respectively; and N^k are the total number of accesses for class $k = \{C, I\}$ (client or impostor). To obtain the FAR and FRR curves, one sweeps over different Δ values.

[13] Also called False Match Rate (FMR) and False Non-Match Rate (FNMR). In this thesis, we are interested in algorithmic evaluation (as opposed to scenario or application evaluation), hence other errors such as Failure to Enroll and Failure to Acquire do not contribute to FAR and FRR. As a result, FAR and FRR are taken to be the same as FMR and FNMR, respectively. **[[]** reference?]

2.2.2 Threshold Criterion

To choose an "optimal threshold" Δ, a threshold criterion is needed. This criterion has to be optimized on a development set. Two commonly used criteria are Weighted Error Rate (WER) and Equal Error Rate (EER). WER is defined as:

$$\text{WER}(\alpha, \Delta) = \alpha \text{FAR}(\Delta) + (1 - \alpha)\, \text{FRR}(\Delta), \tag{2.4}$$

where $\alpha \in [0, 1]$ balances between FAR and FRR. The WER criterion discussed here is a generalization of the criterion used in the yearly NIST evaluation plans [148, Chap. 8] (known as C_{DET}) and that used in the BANCA protocols [5]. This is justified in Section B.

Let Δ^*_α be the optimal threshold that *minimizes* WER on a *development set*. It can be calculated as follows:

$$\Delta^*_\alpha = \arg\min_{\Delta} |\alpha \text{FAR}(\Delta) - (1 - \alpha)\, \text{FRR}(\Delta)|. \tag{2.5}$$

Note that one could have also used a second minimization criterion:

$$\Delta^*_\alpha = \arg\min_{\Delta} \text{WER}(\alpha, \Delta). \tag{2.6}$$

In theory, these two minimization criteria should give identical results. This is because FAR is a decreasing function while FRR is an increasing function of threshold. In practice, however, they do not, since FAR and FRR are empirical functions and are not smooth. (2.5) ensures that the difference between weighted FAR and weighted FRR is as small as possible while (2.6) ensures that the sum of the two weighted terms are minimized. By taking advantage of the shape of FAR and FRR, (2.5) can estimate the threshold more accurately and is used for evaluation in this study.

Note that a special case of WER where $\alpha = 0.5$ is known as the EER criterion. The EER criterion makes the following two assumptions: the costs of FA and FR are equal and the prior probabilities of client and important class are equal.

2.2.3 Performance Evaluation

Having chosen an optimal threshold using the WER threshold criterion discussed in Section 2.2.2, the final performance is measured using Half Total Error Rate (HTER). Note that the threshold (Δ^*_α) is found with respect to a given α. The HTER is defined as:

$$\text{HTER}(\Delta^*_\alpha) = \frac{\text{FAR}(\Delta^*_\alpha) + \text{FRR}(\Delta^*_\alpha)}{2}. \tag{2.7}$$

It is important to note that the FAR and FRR do not have the same *resolution*. Because there are more simulated impostor accesses than the client accesses in most benchmark databases, FRR changes more drastically than does FAR. Hence, when comparing the performance using $\text{HTER}(\Delta^*_\alpha)$ from two systems (at the *same* cost α), the question of whether a given HTER difference is statistically significant or not has to take into account the highly unbalanced numbers of client and impostor accesses. This is discussed in Section 2.2.4.

Note that the key idea advocated here is that the threshold has to be fixed *a priori* using a threshold criterion (optimized on a development set) before measuring the system performance (on an evaluation set). The system performance obtained this way is called *a priori*. On the other hand, if one *optimizes* a criterion and quotes the performance on the *same* data set, the performance is called *a posteriori*. The *a posteriori* performance is thus overly optimistic because one assumes that the class-conditional score distributions are completely known in advance. In an actual operating system, the class-conditional score distributions as well as the class prior probabilities are unknown; yet the decision threshold has to be fixed *a priori*. Quoting *a priori* performance thus reflects better the application need. This subject is further discussed in Section 2.2.6. It is for this reason that the NIST yearly evaluation plans include two sets of performance for C_{DET}: one *a priori* and another *a posteriori* (called minimum C_{DET}). In this thesis, only *a priori* HTER is quoted.

2.2.4 HTER Significance Test

Although there exists several statistical significance tests in the literature, e.g., the McNemar's Test [30], it has been shown that the HTER significance test [9] better reflects the unbalanced nature of precision in FAR and FRR.

A two-sided significance test for HTER was proposed in [9]. Under some reasonable assumptions, it has been shown [9] that the difference of HTER of two systems (say A and B) is normally distributed with the following variance:

$$\sigma^2_{\text{HTER}} = \frac{\text{FAR}_A(1-\text{FAR}_A)+\text{FAR}_B(1-\text{FAR}_B)}{4\cdot N^I} + \frac{\text{FRR}_A(1-\text{FRR}_A)+\text{FRR}_B(1-\text{FRR}_B)}{4\cdot N^C} \tag{2.8}$$

where HTER_A, FAR_A and FRR_A are HTER, FAR and FRR of the first system labeled A and these terms are defined similarly for the second system labeled B. N^k is the number of accesses for class $k = \{C, I\}$. One can then compute the following z-statistic:

$$z = \frac{\text{HTER}_A - \text{HTER}_B}{\sigma_{\text{HTER}}}. \tag{2.9}$$

Let us define $\Phi(z)$ as the cumulative density of a normal distribution with zero mean and unit variance. The significance of z is calculated as $\Phi^{-1}(z)$. In a standard two-sided test, $|z|$ is used. In (2.9), the sign of z is retained so that $z > 0$ (resp. $z < 0$) implies that $\text{HTER}_A > \text{HTER}_B$ (resp. $\text{HTER}_A < \text{HTER}_B$). Consequently, $\Phi^{-1}(z) > 0.5$ (resp. $\Phi^{-1}(z) < 0.5$).

Note that the HTER significance test [9] does not consider the fact that scores from the same user template/model are correlated. As a result, the confidence interval can be under-estimated. There exists a more advanced technique that considers such dependency and it is called the bootstrap subset technique [12]. Note that the usage of the HTER significance test and that of the bootstrap subset technique are different. If one is interested in comparing two algorithms evaluated on the *same* database (hence of the same population and size), the HTER significance test is adequate. However, if one is interested in comparing two algorithms evaluated on two different databases (hence *different* sets of population) the bootstrap subset is more appropriate.

2.2.5 Measuring Performance Gain And Relative Error Change

This section presents the "gain ratio". This measure is aimed at quantifying the performance gain obtained due to fusion with respect to the baseline systems. Suppose that there are $i = 1, \ldots, N$ baseline systems. HTER_i is the HTER evaluation criterion (measured on an *evaluation* set) associated to the output of system i and HTER_{COM} is the HTER associated to the combined system. The "gain ratio" β has two definitions, as follows:

$$\beta_{mean} = \frac{\text{mean}_i(\text{HTER}_i)}{\text{HTER}_{COM}}, \tag{2.10}$$

$$\beta_{min} = \frac{\text{min}_i(\text{HTER}_i)}{\text{HTER}_{COM}}, \tag{2.11}$$

where β_{mean} and β_{min} are the proportion of the HTER of the combined (fused) system with respect to the mean and the minimum HTER of the underlying systems $i = 1, \ldots, N$. In order that $\beta_{min} \geq 1$, several conditions have to be fulfilled (see Section C.3).

Another measure that we use often is the relative error change. It is defined as:

$$\text{relative HTER change} = \frac{\text{HTER}^{new} - \text{HTER}^{old}}{0 - \text{HTER}^{old}} = \frac{\text{HTER}^{new}}{\text{HTER}^{old}} - 1,$$

where the zero in the denominator is made explicit to show that the relative error change compares the amount of error reduction with respect to the maximal reduction possible, i.e., zero in this case. This measure is useful because it takes into account the fact that when an error rate is already very low, making some more progress becomes very difficult.

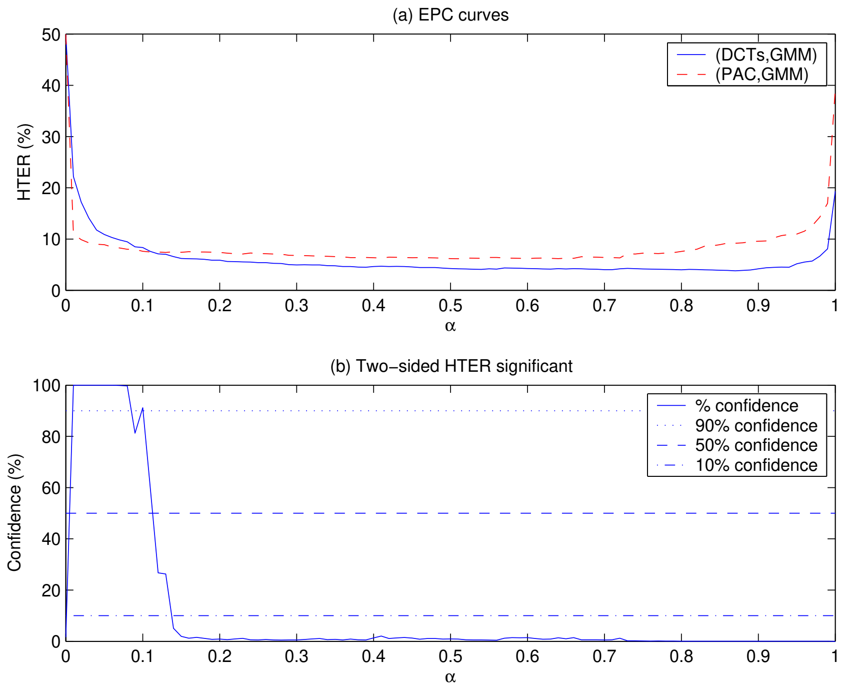


Figure 2.1: An Examples of two EPC curves and their corresponding significance level of HTER difference. (a): Expected Performance Curves (EPCs) of two experiments: one is a face system (DCTs,GMM) and the other is speech system (PAC,GMM). (b) HTER significance test of the two EPC curves. Confidence more than 50% implies that the speech system is better and vice-versa for confidence less than 50%. This is a two-tailed test so two HTERs of a given *cost* α are considered significantly different when the level of confidence is below 10% or above 90% (for a significance level of 20%, in this case for illustration).

2.2.6 Visualizing Performance

Perhaps the most commonly used performance visualizing tool in the literature is the Detection Error Trade-off (DET) curve [82], which is actually a Receiver Operator Curve (ROC) curve plotted on a scale defined by the inverse of a cumulative Gaussian density function. It has been pointed out [8] that two DET curves resulted from two systems are not comparable because such comparison does not take into account how the thresholds are selected. It was argued [8] that such a threshold should be chosen *a priori* as well, based on a given criterion such as WER in (2.5). As a result, the Expected Performance Curve (EPC) [8] was proposed. We will adopt this evaluation method, which is also in coherence with the original Lausanne Protocols defined for the XM2VTS and the BANCA databases.

The EPC curve simply plots HTER (in (2.7)) versus α (as found in (2.4)), since different values of α give rise to different HTER values. The EPC curve can be interpreted in the same manner as the DET curve, i.e., the lower the curve is, the better the performance but for the EPC curve, the comparison is done at a given cost (controlled by α). Examples of DET and EPC curves can be found in Figure 6.3.

We show in Figure 2.1 how the statistical significance test discussed in Section 2.2.4 can be used in conjunction with an EPC curve. Figure 2.1(a) plots the EPC curves of two systems and Figure 2.1(b) plots their degree of significance. In this case, (DCTs,GMM) is system A whereas (PAC,GMM) is system B. Whenever the EPC curve of system B is lower than that of system A (B is better than A), the corresponding significance curve is more than 50%. Below 10% of confidence (or above 90% of confidence) indicates that system B is statistically significantly worse than A (or system A is statistically significantly worse than B).

2.2.7 Summarizing Performance From Several Experiments

It is often necessary to pool several DET/EPC curves together. For instance, when two algorithms exhibit very similar performance on an experiment, by using N databases, one is interested to know if one system is better than the other by using only a single visualization curve via DET or EPC. Two of these reasons are: (i) to summarize the curves; (ii) to obtain a *significant* statistics. Often, due to fusion, FAR and FRR measures can be very small and can reach 100% accuracy. By pooling the curves, this problem can be avoided. It is due to this problem that an *asymptotic performance* procedure [42] was proposed. This procedure first fits the conditional scores with a chosen distribution model and then the smoothed FAR and FRR curves can be generated. While such a model-based approach is well accepted in the medical fields (where the data is not continuous but rank-ordered) [84], it is not commonly used in biometric authentication. This is because the empirical FAR and FRR values in biometric authentication can be linearly interpolated. The composite FAR and FRR measures hence is a practical solution *without* any model-fitting (whose model and hyper-parameter tuning are subject to discussion).

The main idea in pooling several curves together is by establishing a global coordinate such that the pair of FAR and FRR values from different curves are comparable. Examples of such coordinates are DET angle [2], LLR unique to each DET [54] and the α value used in WER as shown in (2.5), among others. We use the α parameter because it inherits the property that the corresponding threshold is *unbiased*, i.e., the threshold is set without the knowledge of the score distribution of the test set. The pooled FAR and FRR across $i = 1, \dots, N$ experiments for a given $\alpha \in [0, 1]$ is defined as follow:

$$\text{FAR}^{pooled}(\Delta^*_\alpha) = \frac{\sum_{i=1}^{N} \text{FA}(\Delta^*_\alpha)[i]}{N^I \times N}, \tag{2.12}$$

and

$$\text{FRR}^{pooled}(\Delta^*_\alpha) = \frac{\sum_{i=1}^{N} \text{FR}(\Delta^*_\alpha)[i]}{N^C \times N}, \tag{2.13}$$

where FA$(\Delta^*_\alpha)[i]$ counts the number of false acceptances of system i due to using the threshold Δ^*_α at the cost α, N^C is the number of accesses for class $k\{C, I\}$. FR$(\Delta^*_\alpha)[i]$ that counts the number of client is defined similarly. The pooled HTER is defined similarly as in (2.7) by using the pooled versions of FAR and FRR.

2.3 Summary

In this chapter, we discussed the databases and the evaluation techniques that will be used throughout this thesis. In particular, we highlight the following issues:

- **A priori performance**: We quote only *a priori* performance, where the decision threshold is fixed after optimizing a criterion on a separate development set as a function of α. In contrast, quoting *a posteriori* performance measured on an evaluation set is *biased* because such performance assumes that the class-conditional distribution of the test score is completely known in advance. For this reason, all DET/EPC curves in this thesis are plotted with *a priori* performance given (some equally spaced and sampled values of) $\alpha \in [0, 1]$[14].

- **HTER significance test:** We choose to employ the HTER significance test that considers the unbalanced numbers of client and impostor accesses, thereby obtaining a more realistic confidence interval around the performance difference involving two systems.

- **Pooled performance evaluation:** We adopt a strategy to visualize a composite EPC/DET curve that is summarized from several experiments.

In this chapter, we also made available a score-level fusion benchmark fusion benchmark dataset which was published in [106].

[14]The DET curve plotted with *a priori* FAR and FRR values is hence a discrete version of the original DET curve. This is not a weakness as a fine sampling of α values will compensate for the discontinuities. The advantage, however, is that when "comparing two DET curves", we actually compare two HTERs given the same α value. In this sense, the α value establishes an unambiguous coordinate where points on two DET curves can be compared.

Part I

Score-Level Fusion From the LLR Perspective

Chapter 3

Score-Level Fusion

3.1 Introduction

Fusing information at the score level is interesting because it reduces the problem complexity by allowing different classifiers to be used independently of each other. Since different classifiers are used, a fusion classifier will have to take into consideration the fact that the scores to be combined are of different types, e.g., a fingerprint which outputs scores in the range $[0, 1000]$, a correlation based face classifier which outputs scores in the range $[-1, 1]$, etc. In this respect, there exists two fusion strategies. In the first strategy, the system outputs are mapped into a common *score representation* – a process called score normalization – before they are combined using (very often) simple rules, e.g., $\min$, $\max$, mean, etc. Learning takes place at the score normalization stage. In the second strategy, a fusion classifier is learnt from the scores to be combined directly. Examples of fusion classifiers are Support Vector Machines, Logistic Regression, etc. Both the fusion strategies are analyzed in this chapter.

While there exists many score representations, only two score representations are statistically sound: probability and Log-Likelihood Ratio (LLR). While in theory, both representations are equivalent, using LLR has the advantage that the corresponding scores can be conveniently characterized by the first- and second-order moments. Furthermore, these moments can be conditioned on a particular user, thus providing a means to introduce the statistics associated to a particular user.

This chapter is presented with the goal to prepare the reader to better understand our original contributions on better understanding the fusion problem (Chapter 4 in Part I) and on user-specific processing (Part II).

Chapter Organization

This chapter contains the following sections: Section 3.2 introduces the notations to be used through out this thesis and presents some of the basic concepts, e.g., levels of information fusion and decision functions. Section 3.3 emphasizes the importance of mapping the system outputs into a common domain since the system outputs are *heterogeneous* (of different types). Section 3.4 includes a survey of existing fusion techniques. Section 3.5 emphasizes the benefits of working on the LLR representation of system outputs from the fusion perspective. These benefits will be concretely shown in Chapter 4 using a parametric fusion model, as well as in Chapters 6 and 7, where scarce user-specific information is exploited.

In order to support some of the claims in this chapter, several experiments have been carried out. However, in the interest to keep this chapter concise, none of the experimental results (in terms of DET/EPC curves) are included here. Most of these results can be found in [101].

3.2 Notations and Definitions

3.2.1 Levels of Fusion

According to [132] (and references herein), biometric systems can be combined at several architectural levels, as follow:

- **sensor**, e.g., weighted sum and concatenation of raw data,
- **feature**, e.g., weighted sum and concatenation of features,
- **score**, e.g., weighted sum, weighted product, and post-classifiers (the conventional machine-learning algorithms such as SVMs, MLPs, GMMs and Decision Trees/Forests); and
- **decision**, e.g., majority vote, Borda count, Behavioral Knowledge Space [138], Bayes fusion [74], AND and OR.

The first two levels are called pre-mapping whereas the last two levels are called post-mapping. Algorithms working in-between the two mappings are called midst-mapping [132]. We are concerned with the *score* level fusion (hence post-mapping) in this thesis. Note that we do not work on the decision level fusion but the score level fusion because much richer information is available at the score level, e.g., user-specific score statistics. In fact, an experimental study in [74] shows that the decision level fusion does not generalize as well as the score level fusion (although this was the objective of the paper).

3.2.2 Decision Functions

Let us denote C (for client) and I (for impostor) as the two class labels the variable k can take, i.e., $k \in \{C, I\}$. Note that class C is also referred to as the *genuine* class. We consider a "person" as a composite of data for various biometric modalities, which can be captured by biometric devices/sensors, i.e.,

$$\text{person} = \{\mathbf{t}_{face}, \mathbf{t}_{speech}, \mathbf{t}_{fingerprint}, \ldots\},$$

where $\mathbf{t}_i$ is the raw data, i.e., 1D, 2D and multi-dimensional signals, of the i-th biometric modality.

To decide whether to accept or reject an access requested by a person, one can evaluate the *posterior probability ratio* in logarithmic domain (called log-posterior ratio, LPR):

$$\begin{aligned}
\text{LPR} &\equiv \log\left(\frac{P(C|\text{person})}{P(I|\text{person})}\right) = \log\left(\frac{p(\text{person}|C)P(C)}{p(\text{person}|I)P(I)}\right), \\
&= \underbrace{\log\frac{p(\text{person}|C)}{p(\text{person}|I)}} + \underbrace{\log\frac{P(C)}{P(I)}}, \\
&= \log\frac{p(\text{person}|C)}{p(\text{person}|I)} - \log\frac{P(I)}{P(C)} \equiv y^{llr} - \Delta, \qquad (3.1)
\end{aligned}$$

where we introduced the term y^{llr} – also called a Log-Likelihood Ratio (LLR) score – and a threshold $\Delta \equiv \log\frac{P(I)}{P(C)}$ to handle the case of different priors. This constant also reflects the different *costs* of false acceptance and false rejection. In both cases, the threshold Δ has to be fixed *a priori*. The decision of accepting or rejecting an access is then:

$$\text{decision(LPR)} = \begin{cases} \text{accept} & \text{if LPR} > 0 \\ \text{reject} & \text{otherwise,} \end{cases} \qquad (3.2)$$

or

$$\text{decision}_\Delta(y^{llr}) = \begin{cases} \text{accept} & \text{if } y^{llr} > \Delta \\ \text{reject} & \text{otherwise,} \end{cases} \qquad (3.3)$$

where in (3.3), the adjustable threshold is made explicit.

Let y^{prob} be the probability of being a client, i.e., $y^{prob} \equiv P(C|\text{person})$ and using the definition of LPR $\equiv \log\left(\frac{P(C|\text{person})}{P(I|\text{person})}\right)$, the decision function of (3.2) can be written as $P(C|\text{person}) > P(I|\text{person})$ or $P(C|\text{person}) > 0.5$, since $P(C|\text{person}) + P(I|\text{person}) = 1$. In terms of y^{prob}, this decision function is:

$$\text{decision}_\Delta(y^{prob}) = \begin{cases} \text{accept} & \text{if } y^{prob} > 0.5 \\ \text{reject} & \text{otherwise,} \end{cases} \tag{3.4}$$

Note that the prior probability has already been absorbed, i.e, $P(C|\text{person}) \propto p(\text{person}|C)p(C)$.

We call y^{llr} an LLR score whereas y^{prob} a probability[1]. In theory, the decision functions of (3.3) and (3.4) are equivalent because both can be derived from (3.2). However, in practice, the explicit presence of a threshold in (3.3) is *more convenient* because the prior probabilities ($P(C)$ and $P(I)$) can be adjusted *separately* from the LLR score. For this reason, (3.3) is *more commonly used* in the literature. For the rest of the discussion, we will write $y \equiv y^{llr}$ so that we consistently use LLR in our discussion unless stated otherwise.

3.2.3 Different Contexts of Fusion

From an architectural view point, the (LLR) score y can be explicitly written as:

$$y \equiv f_\theta(f_e(s(\mathbf{t}))), \tag{3.5}$$

where, s is a sensor capturing a particular biometric trait $\mathbf{t}$, f_e is a feature extractor, θ is a set of classifier parameters associated to the classifier f_θ. We also denote $\mathbf{x} \equiv f_e(s(\mathbf{t}))$ when only the extracted features are concerned.

When considering different fusion contexts, the score y is associated to a subscript i, which takes on a different meaning. The score can be summarized as follows:

$$y_i(\text{person}) = \begin{cases} f_\theta(f_e(s(\mathbf{t}[i]))) & \text{if multi-sample} \\ f_\theta(f_e(s_i(\mathbf{t}_i))) & \text{if multi-modal} \\ f_\theta(f_{e,i}(s(\mathbf{t}))) & \text{if multi-feature} \\ f_{\theta,i}(f_e(s(\mathbf{t}))) & \text{if multi-classifier,} \end{cases} \tag{3.6}$$

where $\mathbf{t}$ denotes any given one of the $\mathbf{t}_i$ biometric traits for $i \in \{face, speech, \ldots\}$, $\mathbf{t}[i]$ denotes the i-th instance (in time) of the biometric trait $\mathbf{t}$, and $\mathbf{t}_i$ denotes the i-th biometric trait. As in common biometric applications, we assume that a dedicated sensor is designed to capture a specific biometric trait, i.e., $s_i(\mathbf{t}_i)$.

Note that the index i takes on a different meaning in any of the four contexts in (3.6). For example, i denotes the i-th instance in the multi-sample case, the i-th biometric modality in the multi-modal case, the i-th feature set in the multi-feature case, and the i-th classifier in the multi-classifier case.

To simplify the notation, we write y_i instead of $y_i(\text{person})$, while bearing in mind that y_i is always dependent on the "person" (in the sense of composite 1D or 2D signals as captured by biometric devices) who makes an access request. Without loss of generality, we assume that for each access request, there are $y_i|i \in \{1, \ldots, N\}$ scores available. We further write y to refer to the output of any of the arbitrary systems $i \in \{1, \ldots, N\}$.

Let $\mathbf{y} = [y_1, \ldots, y_N]'$ be the vector of system outputs to be combined. To decide if an access should be granted or not, a fusion classifier $f_{COM} : \mathbb{R}^N \rightarrow \mathbb{R}$ must be defined. This can be expressed by $y_{COM} = f_{COM}(\mathbf{y})$. Note that the decision function in (3.3) can still be used for the score y_{COM}. The different types of fusion classifiers of the form f_{COM} will be discussed in Section 3.4. In the next Section we will examine different score types commonly used in the literature.

[1] There is an increase use of $y^{prob\prime} = P(C|y)$ in fusion, e.g., [60], where y is an output score and $P(C|y)$ is considered a *score-normalization procedure* intended to approximate the ideal probability $y^{prob} = P(C|\mathbf{t})$ and $\mathbf{t}$ is a biometric trait. While y^{prob} is a true probability, $y^{prob\prime}$ can, at best, be the *score-level approximation* of y^{prob}. No distinction is made between y^{prob} and $y^{prob\prime}$ in this thesis.

3.3 Score Types and Conversion

3.3.1 Existing Score Types

In biometric authentication, there are several types of output, depending on the underlying system, which are listed as follows:

- **Distance metric:** $y \in \mathbb{R}^+$ (a positive number). This is often an output of a template matching system using $y = dist(\mathbf{x}, \mathbf{x}_{tmplt})$, where $dist$ is a distance function comparing a stored template $\mathbf{x}_{tmplt}$ and a query biometric sample $\mathbf{x}$. Some fingerprint recognition system outputs an index between 0 and 1000 using the function INT$(y \times 1000)$ where INT converts any real number to its nearest integer value.

- **Probability** $y \in [0, 1]$. This is a typical output of a Multi-Layer Perceptron (MLP) with a sigmoid activation output.

- **Similarity index:** $y \in [-1, 1]$. This is a typical output of a Multi-Layer Perceptron (MLP) with a hyperbolic tangent activation function.

- **Correlation index:** $y \in [0, 1]$. Similar to a distance, the correlation index measures the closeness of two biometric samples.

- **LLR score:** $y \in \mathbb{R}$ (a real number). This type of output is typical for systems relying on LLR test, i.e., Bayes classifier. The state-of-the-art speaker verification system based on the Gaussian Mixture Models (GMMs) output an LLR.

- **Direction from the decision plane:**. The classical Linear Discriminant Analysis and the more recent Support Vector Machines (SVMs) for instance output a score that can be interpreted as a geometric perpendicular direction from the decision hyper plane in the feature (or kernel) space. Based on the direction (positive or negative), a decision function classifies a sample as either one class or the other. The distance (magnitude) of this direction can be associated with the level of confidence in classifying a given query sample.

Although there are many types of scores, they can be categorized roughly by their types of class-conditional distribution, i.e., approximately normally (Gaussian) distributed or not. By *approximately normally distributed*, we mean that the scores can be summarized by the first order (mean) and second order (covariance) statistics. Obviously probability and similarity index ($[-1, 1]$) have extremely skewed class-conditional distributions. The rest of the scores are approximately normally distributed [109] (see also Section C.1). Fortunately, by converting the probability scores (and similarly the similarity scores) to LLR scores, the process that causes such a skewed class conditional (score) distribution can be reversed. This subject is discussed in Section 3.3.2.

3.3.2 Score Conversion Prior to Fusion

Given the heterogeneous system outputs listed in Section 3.3.1, the first challenge is to convert them into a common representation. We survey here a family of score-normalization procedures here, namely, conversion to probability and to LLR, non-linear score conversion, linear score conversion with the $[0, 1]$ range constraint and linear score conversion without the $[0, 1]$ range constraint While these score normalization procedures are not new, e.g., [60], our somewhat original contribution here is to propose algorithms to systematically convert any score types into probability and LLR.

Conversion Between Probability and LLR

According to the decision functions discussed in Section 3.2.2, there are only two types of score, i.e., probability and LLR. We will discuss the conversion between both types of scores here.

Algorithm 1 Conversion to probability: $f_{prob}(y)$

- If y is an LLR score, $f_{prob}(y) = \text{sigmoid}(y)$.
- If y is $P(C|\mathbf{x})$, $f_{prob}(y) = y$.
- If y is $\tanh$, $f_{prob}(y) = \frac{1+y}{2}$.
- If y is a distance metric, a similarity index, a correlation or any other score type not considered, two solutions can be used:
  1. $f_{prob}(y) = \text{sigmoid}(f_{LLR}(y) - \Delta)$ where $\Delta = \frac{P(I)}{P(C)}$. See Algorithm 2 for $f_{LLR}(y)$.
  2. $f_{prob}(y) = \text{sigmoid}(\frac{y-B}{A})$ where A and B have to be empirically adjusted using algorithms such as logistic regression [56]. This is a more *ad hoc* form and was reported in [60, 127] for instance.

Algorithm 2 Conversion to LLR: $f_{LLR}(y)$

- If y is an LLR score, $f_{LLR}(y) = y$.
- If y is $P(C|\mathbf{x})$, $f_{LLR}(y) = \text{sigmoid}^{-1}(y)$.
- If y is $\tanh$, $f_{LLR}(y) = \tanh^{-1}(y)$.
- If y is a distance metric, a similarity index, a correlation or any other score type not considered, $f_{LLR}(y) = \log \frac{p(y|C)}{p(y|I)} - \log \frac{P(C)}{P(I)}$.

Let $y = P(C|\mathbf{t})$. By using the definition of LPR appeared in (3.2), LLR and probability can be converted into one another by:

$$\text{LPR} = \log \frac{P(C|\mathbf{t})}{P(I|\mathbf{t})} = \log \frac{y}{1-y} \quad \text{or} \quad \text{sigmoid}(z)^{-1} = \log \frac{z}{1-z} \tag{3.7}$$

$$y = \frac{1}{1+\exp(\text{LPR})} \quad \text{or} \quad \text{sigmoid}(z) = \frac{1}{1+\exp(z)}, \tag{3.8}$$

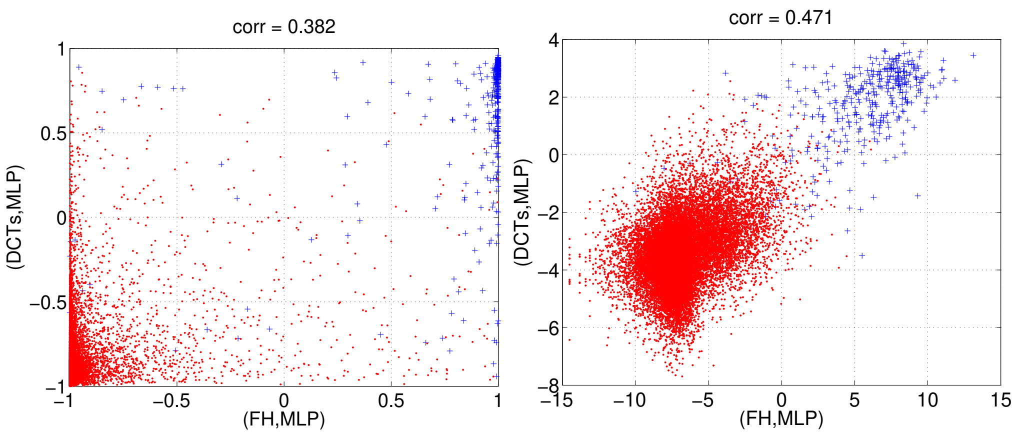
where we explicitly show that a probability can be converted to an LPR using an inverse sigmoid function and the process can be reversed using a sigmoid function. In a similar fashion, an MLP output y with a hyperbolic tangent activation function, $\tanh(z) = \frac{\sinh(z)}{\cosh(z)}$, can be mapped into LLR by its inverse, i.e.,

$$\tanh^{-1}(y) = \frac{1}{2} \log \left(\frac{1+y}{1-y} \right). \tag{3.9}$$

The algorithms that convert from *any score type* (including those not considered in Section 3.3.1) to probability and LLR are shown in Algorithm 1 and 2, respectively.

An Example to Illustrate the Differences Between Probability and LLR

To motivate why converting from one score type to another is important, we consider a fusion task consisting of two systems in the XM2VTS database (see Section 2.1.1). These two systems are based on outputs of two MLP classifiers with non-linear activation functions. The scores *before* and *after* transformation into LLR are plotted in Figure 3.1. Because these two systems use the same face image as input (but different feature representations), their system outputs are expected to be somewhat correlated. Their corresponding correlations before and after LLR transformation are measured to be 0.382 and 0.471, respectively. As can be seen, the supposedly observed correlation is *underestimated* using the original scores (due to hyperbolic



(a) similarity index or probability (b) LLR

Figure 3.1: Conversion between probability and LLR. Scatter plots of two systems (a) before and (b) after probabilistic inversion. The X-axis is a face system based on histogram features and an MLP classifier, labeled as (FH,MLP). The Y-axis is also a face system based on DCTMod2 features and an MLP classifier, labeled as (DCTs,MLP).

tangent transformation) than using the transformed scores in LLR. Furthermore, the transformed scores can better be characterized by the first and second-order moments (the second order moment, variance, is proportional to correlation). More about the merits of working in probability and LLR will be discussed in Section 3.5.

Non-Linear Score Conversion

In [60], several variants of sigmoid-like functions are proposed, namely double-sigmoid and tanh-estimator. While the techniques mentioned thus far are parametric approaches that convert any score type to probability, in [101], we proposed a non-parametric approach. It is defined as:

$$f_{prob}(y) = \mathrm{FRR}(y) - \mathrm{FAR}(y). \tag{3.10}$$

where FRR and FAR are estimated curves from the scores.

Linear Score Conversion with [0,1] Output Range Constraint

There exists also a family of linear transformation functions, all of the form

$$f_{lin}(y) = \frac{y - B}{A}, \tag{3.11}$$

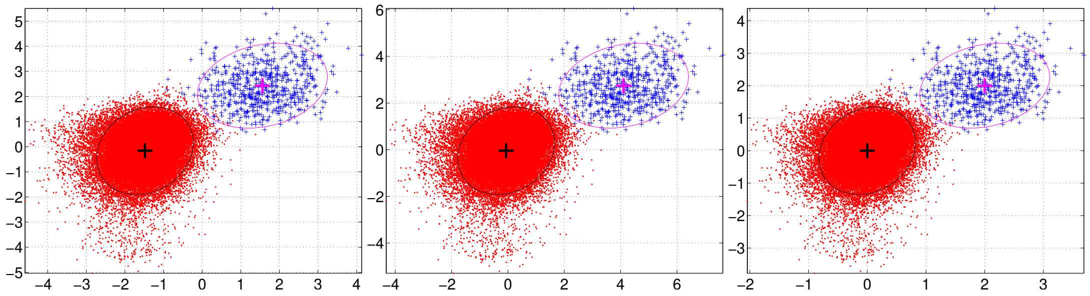
such that

$$f_{lin} : \mathbb{R} \to [0, 1]. \tag{3.12}$$

The terms $\{A, B\}$ are called scaling factor and bias, respectively. Examples of normalization procedures [60] are:

- decimal-scaling, i.e., $\{\left(10^{\log_{10}\max y},\right)^{-1}, 0\}$
- min-max, i.e., $\{(\max(y) - \min(y))^{-1}, \min(y\}$,
- median, i.e., $\{\mathrm{median}(|y - \mathrm{median}(y)|)^{-1}, \mathrm{median}(y)\}$

Note that imposing the range to be $[0, 1]$ does not guarantee that the normalized scores are probability. For instance, $f_{prob}(y) > 0.5$ can be a sensible decision rule where as $f_{lin}(y) > 0.5$ is not guaranteed to be optimal.



(a) original (b) Z-norm (c) F-norm

Figure 3.2: Effects of some linear score transformations. Scatter plots of one of the fusion data sets using (a) the original score, (b) Z-norm and (c) F-norm. The X- and Y-axes are the outputs of two systems. For each sub-figure and each class of scores, a bi-variate Gaussian fit is also depicted whose mean is marked by a big plus sign and whose width is displayed with an oval. The client cluster of scores (small plus signs) are on the upper right corner and the those of impostor (small dots) are on the lower left corner. Note that for (b), the impostor centers are always zero for the two systems whereas the client centers could take on any value. In (c), not only the impostor centers are always zero, the client centers are also fixed to 2 in this case (or any number desired). Due to being linear transformations, both Z- and F-norms *preserve* the score distribution linearly.

Linear Score Conversion without [0,1] Output Range Constraint

Another commonly used normalization also having the form of (3.12) is called z-score normalization (or Z-norm), except that $f_z : \mathbb{R} \to \mathbb{R}$. The following choice of parameters $\{A, B\}$ can be used:

(1) **Unconditioned Z-norm:** i.e., $\{\mu, \sigma\}$, where $\mu \equiv E[y]$ and $\sigma \equiv \sqrt{Var[y]}$. These parameters are motivated by the assumption that the unconditional scores y are normally distributed. In reality, this assumption is violated (even if the *class-conditional* scores are normally distributed!) but practically it still works.

(2) **Impostor-conditioned Z-norm:** i.e., $\{\mu^I, \sigma^I\}$, where $\mu^I \equiv E_{y \in \mathcal{Y}|I}[y]$ and $\sigma^I \equiv \sqrt{Var_{y \in \mathcal{Y}|I}[y]}$. In doing so, one applies the parameters conditioned only on the impostor distribution. The rationale is that the parameters of the client distribution are less informative (due to the relatively less data points on which the parameters are estimated) compared to that of the impostor distribution.

(3) **F-norm:** i.e., $\{\mu^I, \mu^C - \mu^I\}$ which *relaxes* the conditional Gaussian assumption because the second-order statistic $\sigma^k|\forall_k$ are not used. Note that in this case, both the client and impostor parameters are used, i.e., F-norm is considered "client-impostor centric".

Unless stated otherwise, the term "Z-norm" refers to the impostor-conditioned Z-norm in this thesis, especially Chapter 7. While Z-norm is commonly used in the literature, F-norm is our original idea and is presented here for convenience. The rationale for its parameters is justified in Section 4.4.3.

Figure 3.2 shows the effect of impostor-conditioned Z-norm and F-norm. Preliminary experiments using both these normalization procedures show that their fusion performance, using the mean operator, are not statistically significantly different [101]. However, as will be illustrated in Chapter 7, a modified version of F-norm that limits the hypothesis to each user is superior over Z-norm. This is because F-norm is client-impostor centric, whereas (the impostor-conditioned) Z-norm is (necessarily) impostor centric.

3.4 Fusion Classifiers

This section contains a brief survey of the commonly used fusion techniques in pattern recognition. Section 3.4.1 discusses the various ways fusion classifiers can be categorized. We then identify three distinctive types of fusion classifiers each adopting a different philosophy. They are discussed in Sections 3.4.2–3.4.4.

3.4.1 Categorization of Fusion Classifiers

In the literature, there are several ways one can categorize score level fusion classifiers:

- **In probability or in LLR:** To the best of our knowledge, the majority of literature converts scores to probabilities before combining them using sum or product rules [60, 72, 66, 123, 138, 58]. The use of LLR as a score normalization, although equally important, especially in predicting the fusion performance, e.g., [54, 1], is somewhat downplayed. This thesis focuses on LLR.

- **Trainable or non-trainable (classification or combination) [37]:** A fusion classifier needs training if it contains free parameters that have to be optimized given some training data. A trainable fusion classifier can be viewed as a second-level classifier. For this reason, it is also called a stack-generalizer[150] or a supervisor [10]. Examples are any machine-learning based classifier, i.e., SVMs, MLPs, GMMs, etc. On the other hand, since a non-trainable fusion classifier does not have any free parameter, it does not need training. Instead, the training takes place at the score normalization stage, which is an essential part of a non-trainable fusion classifier. Non-trainable classifiers are known as fixed fusion operators here. Examples are mean, $\max$, $\min$, median, majority vote, etc.

- **Dependent or independent [65]:** – Whether one assumes the system outputs to be dependent or not. When they are their probabilities are jointly estimated; otherwise, their probabilities can be separately estimated and combined using a product rule.

- **Adaptive or non-adaptive [132]:** A fusion classifier is considered adaptive if it changes its strategy for each observed sample based on the sample quality. Empirical studies in [127, 10] show that by exploiting the quality information appropriately, the adaptive methods can be superior over the conventional non-adaptive methods.

- **User-specific or user-independent:** In the former, a fusion classifier (or its weight parameters) differs from one user to another. In the latter, all users share the same fusion classifier.

- **Discriminative or generative [145]:** In the former, one introduces a parametric model for the posterior probabilities and infers the values of the parameters from a set of labelled data. In the latter, one models the joint label and feature distributions. This is done by learning the class prior probabilities and the class-conditional densities, separately for each class.

- **Parallel or serial combination [65]:** In the parallel case, each participating system performs the same classification task hence each of them can *also* be used independently. In the serial case, the systems work together in a collaborative manner. One example is a hierarchical classification scheme. Under such a scheme, when a top-level classifier cannot make a decision, it passes the decision making process to the next available level of classifier and so on. A hierarchical approach was reported in [152] to combine multiple feature representations of palmprint. It was shown that the first level of classifier can already achieve 80% of accuracy, leaving the 20% to be fine-tuned by other more computationally demanding classifiers. Note that deciding when to delegate the decision making process to another level of classifier is still an open research problem. We consider only the parallel case in this thesis.

Figure 3.4.1 shows one way to categorize score level fusion classifiers and sections in which they are discussed.

3.4.2 Fusion by the Combination Approach

Having mapped the system outputs to an appropriate space, i.e., probability, LLR or $[0, 1]$ space, combining scores assuming that system outputs are independent become: $\prod_i f_{prob}(y_i)$ and $\sum_i f_{LLR}(y_i)$, respectively for probability and LLR. In the linear space, the theoretical justification for combining scores using simple rules such as sum ($\sum_i (f_{lin}(y_i))$) and product ($\prod_i (f_{lin}(y_i))$) is unclear. In fact, combining scores using simple rules with f_{lin} often results in sub-optimal performance compared to transforming them into probability and LLR [101].

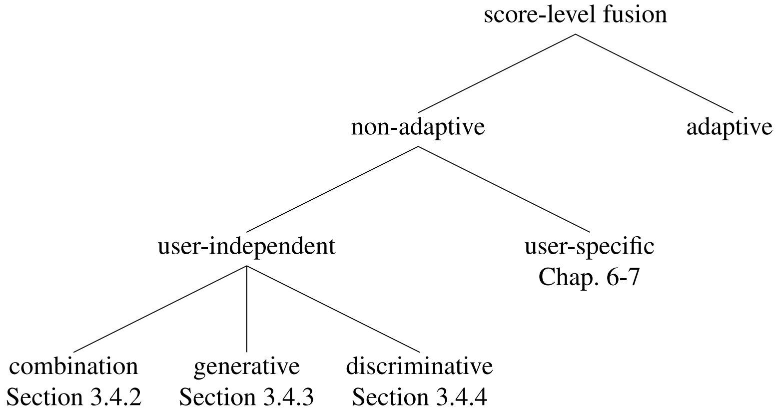


Figure 3.3: Categorization of score-level fusion classifiers.

Simple Fusion Operators (Fixed Rules)

Several operators are commonly used in the literature, namely $\min$, $\max$, median, weighted sum and weighted product, defined as follow:

$$
\begin{aligned}
y_{min} &= \min_i(y_i), &(3.13)\\
y_{max} &= \max_i(y_i), &(3.14)\\
y_{med} &= \mathrm{median}_i(y_i), &(3.15)\\
y_{wsum} &= \sum_{i=1}^{N} w_i y_i, &(3.16)\\
y_{wprod} &= \prod_{i=1}^{N} y_i^{w_i}, &(3.17)
\end{aligned}
$$

respectively, where $w_i|\forall_i$ are parameters that need to be estimated. The mean operator is a special case of weighted sum with $w_i = \frac{1}{N}$. Similarly, the product operator is a special case of weighted product with $w_i = 1$. The $\min$, $\max$ and median operators are sometimes collectively known as *Order Statistics* (OS) combiners because they consider the ordering of scores. The order statistics, mean, sum and product combiners are collectively known as *simple fixed rules* because they do not contain any adjustable parameter.

Kittler *et al* [66] provided an explanation on how these fusion rules can arise as approximations to the product and sum rules in a Bayesian framework. In particular, $\min$ estimates product and $\max$ estimates sum. In the case the estimate of probability y (or y_i, for any i) is *biased* (inaccurate due to mismatch between training and test sets), they showed that the sum rule outperforms the product rule. Note that the so-called “biased” estimate of probability is due to the underlying mismatch between training and test sets. Extending Kittler *et al*’s work, Lucey [76, Chap. 10] provided an interesting noise mismatch framework in probability for independent fusion classifiers. Working towards this direction, we will provide a parametric view in LLR in Section 4.6. Note that in reality, the weighted product rule is more commonly found in *adaptive* fusion where each weight is a function of a quality index [127]. We will thus not discuss further the weighted product rule here.

Specialized Fusion Classifiers Based On the Combination Approach

Two other specialized fusion classifiers should be mentioned here, namely Bayesian expert conciliation [44] and Decision Template (DT) [72]. The expert conciliation is based on the assumption that the conditional scores are normally distributed and is more appropriate to be carried out in LLR. One can implement DT using many types of distance measures such as Dempster-Shafer rules, fuzzy rules and geometric distances. Among them, the most common one is the Euclidean distance, which has the following form:

$$y_{COM} = -\left(\|\mathbf{y} - \boldsymbol{\mu}^C\| - \|\mathbf{y} - \boldsymbol{\mu}^I\|\right) \tag{3.18}$$

where $\|\mathbf{z}\|$ is $\sqrt{\sum_i (z_i^2)}$, z_i is an element of the vector $\mathbf{z}$, $\boldsymbol{\mu}^k$ is the mean vector of system outputs (or a "class prototype"). A negative sign is introduced here so that the measure is interpreted as similarity (the larger it is, the closer $\mathbf{y}$ is to the client prototype). Our empirical studies [101] show that this classifier works best using probability scores. We conjecture that this is due to the unimodal nature of scores in this space. However, its generalization, in most fusion experiments, is worse than the general purpose classifiers that will be discussed in Sections 3.4.3 and 3.4.4.

3.4.3 Fusion by the Generative Approach (in LLR)

Let us define the *joint* system output in the LLR domain by $\mathbf{y}^{llr} \equiv [y_1^{llr}, \ldots, y_N^{llr}]'$ and $y_i^{llr} \equiv f_{LLR}(y_i)$. Then, the classical approach to establish an LLR test between the client and impostor classes, i.e., $k = \{C, I\}$, is defined as:

$$y_{dep}^{llr} = \log \frac{p(\mathbf{y}^{llr}|C)}{p(\mathbf{y}^{llr}|I)} \qquad \text{or} \qquad y_{dep}^{llr} = \log \frac{p(\mathbf{y}|C)}{p(\mathbf{y}|I)}, \tag{3.19}$$

for the dependent assumption[2] and

$$y_{indep}^{llr} = \log \frac{\prod_i p(y_i|C)}{\prod_i p(y_i|I)} = \sum_i f_{LLR}(y_i), \tag{3.20}$$

for the independent assumption. The approximations to (3.19) and (3.20) using GMM [11, Chap. 2], for any $\mathbf{y}^{llr}$ and $y_i^{llr}|\forall_i$ (or $\mathbf{y}$ and $y_i|\forall_i$, i.e., in the original score domain), can be written as follow:

$$\hat{p}(\mathbf{y}|k) = \sum_{c=1}^{N_{cmp}^k} w_c^k \mathcal{N}(\mathbf{y}|\boldsymbol{\mu}_c^k, \boldsymbol{\Sigma}_c^k), \tag{3.21}$$

$$\hat{p}(y|k) = \sum_{c=1}^{N_{cmp}^k} w_c^k \mathcal{N}\left(y|\mu_c^k, (\sigma_c^k)^2\right), \tag{3.22}$$

for any $y \in \{y_i|i = 1, \ldots, N\}$, respectively, where, the c-th component of the class conditional (denoted by k) mean vector is $\boldsymbol{\mu}^k = [\mu_1^k, \ldots, \mu_N^k]'$, its covariance matrix of dimension $N \times N$ is $\boldsymbol{\Sigma}_c^k$ and there are N_{cmp}^k components for each $k = \{C, I\}$. The mean and variance in the mixture $p(y|k)$, i.e., μ_c^k and $(\sigma_c^k)^2$ are defined similarly except that they are single dimensional. The GMM parameters can be optimized using the Expectation-Maximization algorithm [11] for instance and the number of components can be tuned by validation or optimization of a criterion, e.g., minimum description length [45].

There are two remarks regarding the generative classifiers discussed here:

- **Special cases of generative classifier:** Note that when the number of Gaussian components $N_{cmp}^k = 1$ for $k = \{C, I\}$, the resultant classifier is a Quadratic Discriminant Analysis (QDA) classifier. The Linear Discriminant Analysis (LDA) (also called Fisher linear discriminant) classifier is obtained by further imposing the *common* covariance $\boldsymbol{\Sigma}$. This can be done by taking the linear interpolation of the two covariance matrices, i.e., $\boldsymbol{\Sigma} = \gamma\boldsymbol{\Sigma}^C + (1-\gamma)\boldsymbol{\Sigma}^I$, where γ is parameter to be tuned. We also used two preset values of γ that give acceptable generalization performance. They are $\gamma = P(C)$ (the prior probability of the client class $P(C)$) and $\gamma = 0$ (making the contribution of the client covariance matrix to be zero). The rationale for the second version, $\gamma = 0$, is that the client covariance matrix cannot be estimated reliably. Our empirical results on XM2VTS (not reported here) show that the second version ($\gamma = 0$) generalizes well, especially in the user-specific context (see Section 6). This phenomenon is further confirmed in Section 7.2.

[2]We make no distinction between the first form of (3.19) (on the left) and the second form (on the right) as GMM is a general purpose algorithm. However, by converting scores to LLR (the first form) can ensure that the data is in linear scale. As a result, the LLR scores can be more appropriately summarized by a mixture of Gaussian distributions. In practice, we observe that using the first or second form has no significant influence on the generalization performance.

- **Robustness of naive Bayes classifiers:** Our preliminary fusion experiments (carried out on the XM2VTS database) show that the generalization performance between the fusion classifiers based on (3.19) and that based on (3.20) (also called Naive Bayes Classifier) is *not* statistically significantly different (figure not shown here), even though the system outputs are known to be correlated (e.g., in the context of intra-modal fusion). This is because there are no "outliers" – samples that are found extremely far from the rest. This is not entirely surprising following the observation from [34], which confirms that Naive Bayes classifiers (as in (3.20)) are robust to the underlying system outputs dependency.

3.4.4 Fusion by the Discriminative (Classification) Approach

There exists a handful of discriminative algorithms for score-level fusion. However, one must be careful to take into account the fact that the amount of training samples for each class can be highly unbalanced. We will pay particular attention to linear classifiers as non-linear classifiers such as QDA and reduced polynomial classifier [140] are not known to perform *statistically significantly* better than its linear counterpart[3]. Before doing so, it is important to point out that the bias in the linear classifier, even though is available, is not used directly to make the accept/reject. The externally optimized threshold Δ replaces the actual bias used (see (2.1)). All linear classifiers, in our context, have the following form:

$$y_{COM} = \sum_{i=1}^{N} w_i y_i - \Delta = \mathbf{w}'\mathbf{y} - \Delta \tag{3.23}$$

where, Δ is a bias. For convenience, we introduced the vector representation $\mathbf{w} = [w_i, \ldots, w_N]'$ and $\mathbf{y} = [y_i, \ldots, y_N]'$. The discussion that follows will consider three classifiers in this category: Support Vector Machine, Logistic Regression and Linear (Fisher) Discriminant Analysis.

- **Support Vector Machine:** Among the existing classifiers, SVM [146] is undoubtedly the most popular for two reasons: (i) it relies on minimizing the empirical risk (or maximizing the margin) and (ii) it does not make any assumption about the data (score) distribution. Suppose that $\mathbf{y}^{(j)}$ and $t^{(j)} \in \{-1, 1\}$ (positive or negative class) are the input and target output of example j and $\omega^{(j)}$ is its associated *embedding strength* obtained after SVM training. Large $\omega^{(j)}$ implies that the associated example is difficult to classify, and vice-versa for small $\omega^{(j)}$. Examples with $\omega^{(j)} > 0$ are known as support vectors. The linear solution proposed by an SVM with a linear kernel is:

$$f(\mathbf{y}) = \sum_j \omega^{(j)} t^{(j)} \left\langle \mathbf{y}^{(j)}, \mathbf{y} \right\rangle = \underbrace{\left(\sum_j \omega^{(j)} t^{(j)} \mathbf{y}^{(j)\prime} \right)} \mathbf{y} = \mathbf{w}'\mathbf{y}, \tag{3.24}$$

  where $\langle \cdot, \cdot \rangle$ is the linear kernel and the underbraced term forms is the solution to the weight vector $\mathbf{w}'$.

- **Logistic Regression:** In [56], another algorithm called Logistic Regression (LR) is compared to SVM. According to [56], LR shares many similar characteristics with SVM. Our past empirical experiments show that LR and SVM perform equally well in biometric fusion tasks [113]. LR is defined as:

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

  where

$$g(\mathbf{y}) = \underbrace{\sum_{i=1}^{M} \beta_i y_i} + \beta_0.$$

  One should recognize that $g(\mathbf{y})$ is LPR, the underbraced term is LLR and the bias β_0 is replaced by Δ. The weight parameters β_i are optimized using gradient ascent to maximize the likelihood of the

[3] As no statistical significance test was reported in [140].

data given the LR model [32]. Note that the LR classifier used here is more general than the one used in [94]. The former is the *standard* approach as described in [56] whereas the latter assumes class-conditional Gaussian assumption as well as common covariance of both client and impostor distributions.

SVM and the standard LR classifier are attractive because they do not make any assumption about the distribution of the system outputs and thus are good general purpose algorithms for classification. In practice, using any transformed y, e.g., $f_{lin}(y)$, $f_{LLR}(y)$, or $f_{prob}(y)$, for any $y \in \{y_1, \ldots, y_N\}$ cannot affect the generalization performance of SVM and LR (see [101]). For the case of $f_{lin}(y)$, we illustrate this property theoretically in Section D.1.

- **LDA as a discriminative classifier:** The classical LDA as well as QDA classifier which was discussed in Section 3.4.3 can also be considered a discriminative classifier. This is because LDA can be written as a linear function as in (3.23). Similarly, QDA can be written as a quadratic discriminative function. We will consider the LDA case here because we found its use in user-specific processing (to be used in Chapter 6). Using the class-conditional mean and covariance (i.e., $\boldsymbol{\mu}^k$ and $\boldsymbol{\Sigma}^k$ for each $k = \{C, I\}$) as described in Section 3.4.3, let us define the within-class covariance matrix as:

$$S_w = \sum_{k=\{C,I\}} \boldsymbol{\Sigma}^k$$

The Fisher linear discriminant solution of the weight vector $\mathbf{w}$ for a two-class problem (see [11]) is:

$$\mathbf{w} = S_w^{-1} \left(\boldsymbol{\mu}^C - \boldsymbol{\mu}^I \right) \tag{3.25}$$

Note that the solution w_i can take on any value and their sum is not necessarily equal to 1. As can be seen, LDA turns out to be both generative and discriminative.

Note that LDA and QDA both rely on the Gaussian assumption. As a result, they are inferior in performance compared to SVM and LR which do not make such an assumption. This is confirmed by our empirical studies in [101]. While this assumption seems to be a limitation, converting scores into LLR scores prior to applying LDA can *improve* the generalization performance of LDA.

3.4.5 Fusion of Scores Resulting from Multiple Samples

This section describes two trainable methods to combine scores resulting from multiple samples. This fusion problem is more commonly solved using fixed fusion rules as discussed in Section 3.4.2. Trainable approaches are proposed here because we conjecture that it can give better results since the parameters of the fusion classifier can further be adjusted to suite the data.

Although trainable fusion classifiers as discussed in Sections 3.4.3 and 3.4.4 can be used, they are not suitable for combining scores resulting from multiple samples for two reasons: the ordering in which the samples are presented is not important and the number of samples per access can be different for different accesses.

We choose here two fusion strategies for combining scores from multiple samples. Below are two intuitive rationales for each of the two strategies:

- If one considers the fact that the scores are drawn from a distribution that can be estimated, then, matching can be done by comparing two distributions. This inspires us to use a distribution-based matcher via the *relative entropy*, which is also known as the *Kullback-Leibler distance*. The "relative entropy" method evaluates the difference of two relative entropies: the relative entropy between distribution of the sample scores and that of client scores; and the relative entropy between distribution of the sample scores and that of impostor scores.

- If one treats the scores like a sequence, then classifiers that compare sequence can be used. This inspires us to use GMM, in a similar way that the state-of-the-art speaker verification system [121] is used. The "GMM" method calculates the average log-likelihood ratio of the sample samples between a GMM modeling the client scores and another one modeling the impostor scores.

Both these methods are further described below (for readers who want to probe further):

- **Combining Sample Scores by Relative Entropy:** Relative entropy is used to compare two probabilistic density functions (*pdf*s). In our case, one *pdf* is derived from a global model (client or impostor), denoted as $p_k(y)$, for $k = \{C, I\}$ and the other *pdf* is derived from scores resulting from multiple samples, denoted as $q(y)$. Both *pdf*s can be estimated using any density estimator discussed in [11, Chap. 2], e.g., GMM (as in (3.22)) or the Parzen window. The relative entropy of a given access distribution $q(y)$ with respect to $p_k(y)$ can then be defined as:

$$L(p_k, q) = -\int p_k(y) \ln \frac{q(y)}{p_k(y)}\, dy. \tag{3.26}$$

  In practice, we sample the distribution of p_k and q in fine steps of y so that the integral is approximated by a sum operator over the sampled y space. Relative entropy can be regarded as a distance as to how much $q(y)$ is from $p_k(y)$ but not the other way round, i.e., this distance is not symmetric. This alone does not give discriminative information. To do so, the relative entropy of a client and impostor models should be used together, as follows:

$$y_{COM} = -\big(L(p_C, q) - L(p_I, q)\big) \tag{3.27}$$

  Note that the negative sign is introduced so that $E[y_{COM}|C] > E[y_{COM}|I]$. In this way, the decision function as in (2.1) can be used.

- **Combining Sample Scores by GMM:** This is an extension of GMM (discussed in Section 3.4.3) used in the general context of fusion. In the context of combining multiple samples, one can safely assume that the samples (scores) are drawn from the same distribution $p(y|k)$ estimated using (3.22) for each k. The LLR test can thus be constructed using $f_{LLR}(y_i) = \log \frac{p(y_i|C)}{p(y_i|I)}$ for each sample i. By *naively* assuming that the scores are independent given the user[4], the joint score is:

$$y_{COM} = \sum_i f_{LLR}(y_i). \tag{3.28}$$

  In general, mean is used in place of $\sum$ so that y_{COM} is not biased towards the access characterized by a larger number of samples. In this way, we consider the “average LLR”.

3.5 On the Practical Advantage of LLR over Probability in Fusion Analysis

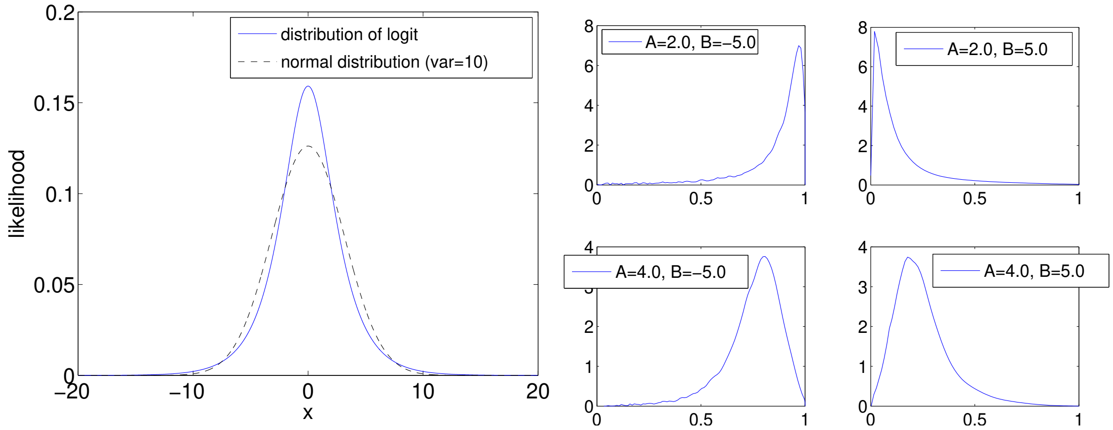
While working in probability and LLR are theoretically equivalent, we have shown intuitively that the statistics of the LLR scores follow approximately a normal distribution. In [23], the logit transform, i.e., $x = \log \frac{z}{1-z}$, was used to post-process randomly generated numbers z to obtain another set of numbers (x) having the following form of distribution:

$$p(x) = \frac{1}{\pi(\exp(x/2) + \exp(-x/2)))} \tag{3.29}$$

One can recognize that x corresponds to an LLR score. We drew 10,000 random samples according to $p(x)$ and re-approximated the sample distribution using a Gaussian. The distribution $p(x)$ and its approximation using a Gaussian distribution are shown in Figure 3.4(a). As can be observed, both the distributions are similar. In this case, both the distributions have zero mean. The approximated Gaussian has a variance fixed to 10 [5]. Using the same generated samples, we applied the sigmoid function with some chosen scale (A) and shift (B) values. The distributions of the resultant transformed probability scores are shown in Figure 3.4(b). Note that the scale value determines the width (variance) of the distribution whereas the

[4] These scores are expected to be dependent because their corresponding biometric samples are closely related in time.

[5] We drew the samples several times and found that the expected variance was about 10.



(a) LLR (b) Probability

Figure 3.4: (a) The distribution of LLR scores and its approximation using a Gaussian distribution. The mean of both distributions are zero. (b) The distribution of probability scores for several shift (B) and scale (A) values using 10,000 sample data generated by the LLR distribution.

shift value determines the center (location) of the distribution. Only when $B = 0$, the score distribution become central since the generated samples have zero mean. Although Figures 3.4(a) and (b) are drawn from the same distribution as shown in (3.29), the LLR scores can be more *conveniently* approximated using a normal distribution whereas the transformed probability scores may have to be described using a non-central distribution, e.g., a gamma distribution. Summarizing the LLR scores using a Gaussian is convenient because a Gaussian distribution is *closed* under a linear transformation. For instance, if a score vector $\mathbf{y}$ follows a multivariate normal distribution and $\mathbf{w}$ is a weight vector, $\mathbf{w}'\mathbf{y}$ will also follow a one-dimensional normal distribution [120]. For this reason, by working on LLR scores, we deviate from the mainstream literature in terms of analysis (where probability is a popular choice), e.g. [135, 67, 57, 76], and fusion methodology (where scores are transformed into probability prior to combination), e.g., [60, 72, 66, 123, 138, 58]. It should be noted that the use of LLR for performance prediction was reported in [54, 1] whereas its use in fusion is more common, e.g., [27, 65].

3.6 Summary

This chapter discussed the following issues:

- **Fusion modes:** Several ways of combining scores are discussed, i.e., using multiple samples, biometric modalities, features and classifiers.

- **Score types:** Some commonly used score types in biometric systems are discussed: probability, LLR, distance, correlation, similarity index, direction from the decision plan, etc.

- **Score normalization:** This issue aims at mapping scores into a common domain so that scores can be combined using simple combination rules. The two statistically sound representations of score are discussed: probability and LLR. Another family of functional transformation of scores having the form $\mathbb{R} \to [0, 1]$ is also discussed. However, this family of functional approaches does not have a sound justification and in practice do not perform as well as converting scores into the probability or the LLR space.

- **Types of score-level fusion classifiers:** Three categories of fusion classifiers are identified: fusion by combination (using simple rules), by the generative approach (using the LLR test) and by the discriminative approach.

While none of the materials presented here is novel, we conclude that between the two statistical representations of score, i.e., LLR and probability, LLR is the *preferred* choice because scores in this domain can be summarized by using the first- and second-order moments. This deviates from the mainstream whereby scores are almost always systematically converted into probability scores prior to fusion using simple rules [60, 72, 66, 123, 138]. The choice of using LLR has important consequences to this thesis. In fact, almost all the contributions in this thesis, as found in Chapters 4–7, essentially demonstrate the usefulness of LLR.

Chapter 4

Towards a Better Understanding of Score-Level Fusion

4.1 Introduction

There have been a growing number of works that empirically show that combining multiple system outputs is beneficial, e.g., [125] (and many references herein). However, admittedly, relatively much less works were reported on the theoretical understanding of fusion, e.g., [66, 73, 57, 143, 123, 76]. Such an understanding is important because the empirical approach to studying fusion *cannot* explain why or when a combined system fails to achieve the desired performance. This is because there are simply too many factors to be considered, e.g., the type of system output, the dependency among system outputs, the relative performance of systems, the choice of decision threshold, the presence of noise and the choice of fusion classifier.

Previous studies on the understanding of fusion rely on one or more of the following simplifying (and unfortunately unrealistic) assumptions:

- **Independence of system outputs:** that the system outputs are independent of each other. In intramodal fusion, where several biometric systems rely on the same biometric capturing device, the system outputs are likely to be correlated. In this case, this assumption is violated.
- **Common class-conditional distributions:** that the client and impostor distributions are the same.
- **Common output distributions:** that the scores of all the system outputs follow a common distribution.

We will consider LLR scores in this chapter so that it is adequate to summarize the LLR scores to be combined using a class-conditional multivariate Gaussian. The resultant client and impostor multivariate Gaussian models are referred to as a "parametric fusion model" since the model essentially summarizes the fusion problem. Although relying on the class-conditional score Gaussian assumption seems to be restrictive, the model is powerful because it does not make use of any of the three simplifying assumptions. Furthermore, we will show that in the context of classification, deviation from this assumption cannot severely influence the precision of the estimated Equal Error Rate (EER).

We will revisit in this chapter a well known upper bound of the minimal classification (Bayes) error, i.e., the Chernoff bound [35], given the parametric fusion model. Although this bound is useful for classification, it does not estimate EER, a measure that is *far more important* as long as performance evaluation is concerned. Our original contribution in this chapter is to propose an exact EER solution given any linear fusion classifier (with mean as a special case) or any order-statistic fusion operators (e.g., $\min$, $\max$ and median). Thanks to the parametric fusion model, we can justify the reduction of classification error due to fusion, study the effect of correlation of system outputs, predict fusion performance and compare the performance of commonly used fusion operators.

Chapter Organization

This chapter is organized as follows: Section 4.2 is purely an empirical study to show that "the combined system is *never* worse than the average performance of its underlying systems". Section 4.4, as opposed to Section 4.2, is a theoretical study that explains the above phenomenon using the parametric fusion model. Section 4.5 demonstrates the real potential of the parametric fusion model by applying the proposed parametric model to determine an optimal subset of systems for fusion.

The next two sections are extended studies based on the parametric model presented in Section 4.4. These are advanced topics and can be skipped for readers who are more interested in user-specific processing (treated in Chapter 6 and 7). Section 4.6 analyzes whether or not correlation is a necessary and sufficient factor to predict the fusion performance (the answer turns out to be necessary but *not sufficient*!), the effect of unbalanced system performance and the effect of noise (or bias) to the fusion performance. Section 4.7 then extends the proposed parametric model to other fusion operators based on order-statistics. Thanks to the extended parametric model, one can now identify the conditions which favor $\min$, $\max$, mean or weighted sum. As a summary, Section 4.8 highlights the original contributions of this chapter with respect to the state-of-the-art in fusion.

Because this chapter is theoretical in nature, most experiments that are designed to support our claims are put in Section C. Readers who are more concerned with the practical applicability of the proposed parametric fusion model are strongly encouraged to refer to the mentioned Section. Finally, a collection of proves, all needed to support the proposed model, can be found in Section D.

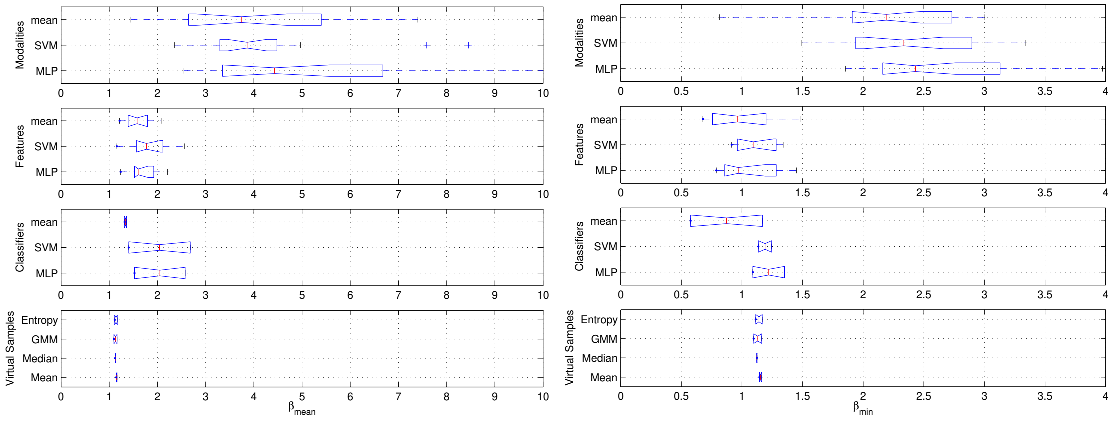
4.2 An Empirical Comparison of Different Modes of Fusion

From (3.6), we know that there are different ways one can create diverse systems, i.e, using different modalities, different classifiers, different feature representations and different samples. We design a set of experiments containing these four scenarios, based on the XM2VTS score-level fusion benchmark database. In each fusion tasks, only two systems are involved. In the first three scenarios, the system outputs are combined using MLP, SVM and the mean operator as in $\mathrm{mean}_i(f_z(y_i))$ (using Z-norm). For the last scenario, we did not have multiple samples per access but we could generate "virtual samples" by randomly introducing geometric transformation to the images (translation, rotation, scaling). In order to combine the scores due to virtual samples, apart from using non-trainable fusion operators, e.g., mean and median, we also used two trainable order-insensitive fusion classifiers: the *relative entropy* and *GMM* approaches as discussed in Section 3.4.5.

From the available 13 systems, we combined each time two systems according to the following modes of fusion:

- multi-modal (21 fusion tasks)
- multi-feature (9 fusion tasks)
- multi-classifier (2 fusion tasks)
- virtual samples (2 fusion tasks)

Details of these experiments can be found in our publication [98]. The results are shown in Figure 4.1. The performance is measured by the gain of *a priori* HTER (as discussed in Section 2.2.5) whose threshold is optimized using WER with $\alpha = 0.5$ (see Section 2.2). As can be observed, all systems achieve $\beta_{mean} \geq 1$, without exception. On the other hand, not all systems achieve $\beta_{min} \geq 1$ – suggesting than fusion may not be always useful. By comparing all four ways of generating diversity, the performance gain is most evident using multimodal fusion. This is expected because richer and more complementary information is available than the other fusion modes. It is interesting to observe that fusion with virtual samples can help improve the performance, albeit statistically insignificantly. Note that higher diversity (as in multimodal case) incurs higher computation/hardware costs. Ideally, one wishes to keep the cost low. This suggests that selecting a subset of systems may be more beneficial, i.e., trading off statistically insignificant performance gain for lower computation. This will be discussed in Section 4.5.



(a) β_{mean} (b) β_{min}

Figure 4.1: An empirical study of relative performance of different modes of fusion. Boxplot of (a) β_{mean} and β_{min}. Each bar shows the relative improvement in terms of β (defined in (2.10) and (2.11)) within 95% of confidence. The vertical line around the middle of each bar is the median of β_{mean}. Dotted lines at each end of a bar are extreme values found outside the 95% confidence interval. For fusion with virtual samples, β_{real} is used in place of β_{mean}. The x-axis of all the boxplots are aligned so that β_{mean} across different techniques of generating diversity are comparable. For virtual samples, the classifier "Entropy" refers to the relative entropy strategy whereas "GMM" refers to the GMM classifier discussed in Section 3.4.5.

4.3 Estimation of Fusion Performance

4.3.1 Motivations

The study of fusion is very often complicated by various factors. Some of these factors are:

1. The type of output of classifier of the base-systems
2. The dependency among features of base-systems
3. The relative performance of base-systems
4. The choice of fusion operator
5. The choice of decision threshold
6. The presence of noise

An empirical approach to understanding fusion is to study one factor by varying its parameters while fixing the rest of the factors. Unfortunately, such an approach is not appropriate since these factors may be dependent on a particular experimental setting and thus cannot be controlled.

We propose to study these factors by first modeling the scores to be combined. To give an intuition, one can summarize the class-conditional scores to be combined using a multivariate Gaussian whose dimension corresponds to the number of systems to combine. This is shown in Figure 4.2. Factor 1, i.e., different types of classifier output, can be considered by mapping scores into a domain where the scores can be more easily summarized by the first- and second-order moments. For example, if scores are probabilities, they can be transformed into LLR using Algorithm 2. Factor 2, i.e., the dependency among system outputs, can be captured by measuring the class-conditional pair-wise correlation among the system outputs. Note that this information has already been captured by the covariance matrix of the class-conditional multivariate Gaussian (since a correlation matrix can be derived from a covariance matrix in a close form). By modeling the scores, factor 3, i.e., the relative performance among systems, will be captured. This point will become clear later. By summarizing the scores using two class-conditional multivariate Gaussians, we will show

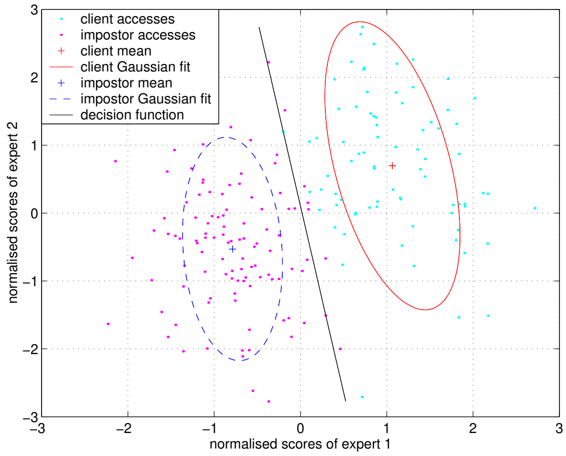


Figure 4.2: A geometric interpretation of a parametric model in fusion. A real fusion task whose samples are fitted by two class-conditional bi-variate Gaussian distribution. System 1 is IDIAP's voice system and system 2 is Surrey's automatic face authentication system, applied on the Ud-g1 BANCA data set.

that it is possible to estimate analytically the distribution of the combined score, for a given fusion operator. Factor 4 is thus considered by repeating the estimate of the combined score distribution *for each fusion operator*. Since the distribution of the combined score can be estimated, its corresponding FAR and FRR curves which are functions of a decision threshold can also be estimated analytically. Therefore, Factor 5 is taken into consideration. Finally, Factor 6 can be considered by assuming that the noise has a known effect on the multivariate class-conditional distributions, e.g., introducing a bias to the mean vector. Therefore, we justify that in order to analyze the problem of fusion, the scores to be combined must be summarized. For a tractable analysis, the use of a multivariate Gaussian distribution is a practical choice.

Section Organization

In Section 4.3.2, we will explain how the scores to be combined in a more formal way, using a so-called "parametric fusion model". Section 4.3.3 then presents a very well known approach – the Chernoff bound – to estimate the *minimal classification (Bayes) error* given the parametric fusion model. In contrast to the Bayes error, Section 4.3.4 explains how the EER of a linear classifier can be estimated given the parametric fusion model. Note that EER plays a somewhat *more important role* in biometric performance evaluation than the minimal Bayes error. Section 4.3.5 then outlines the differences between the minimal Bayes error and EER. Because the proposed parametric model relies on the Gaussian assumption, Section 4.3.6 verifies the adequacy of the model when applied to the real (score) data. By doing so, we examine how well the estimate of EER is when the Gaussian assumption is violated[1].

4.3.2 A Parametric Fusion Model

Let us assume that the i-th system output (out of N participating systems) is composed of a deterministic component μ_i^k, and a noise component η_i^k, and that their relation is additive, i.e.,

$$y_i^k = \mu_i^k + \eta_i^k, \tag{4.1}$$

for $k \in \{C, I\}$. The deterministic component is due to the discrete binary classification task whereas the noise component is due to some random processes during biometric acquisition (e.g. degraded situation

[1]Section 4.3.6 essentially summarizes the experimental results reported in Section C.1 and Section C.2. These two sections are not required to understand the proposed parametric fusion model but are *important* to illustrate empirically that the fusion model is still useful even if the Gaussian assumption is violated. Section C.1 empirically examines the effect of violating the Gaussian assumption. Section C.2 not only relaxes the Gaussian assumption but also improves the experimental design of Section C.1 so that classification errors other than EER, e.g., low FAR and low FRR, are also considered.

due to light change, miss-alignment, etc) which in turn affect the quality of extracted features. Indeed, it has a distribution governed by the extracted feature set $\mathbf{x}$ under some unknown conditions $c \in \mathcal{C}$ such as geometric distortion. The unconditioned noise variance $\eta_i^2(\mathbf{x})$ is related to the conditioned noise variance $\eta_i^2(\mathbf{x}|c)$ by:

$$\begin{aligned} E[\eta_i^2(\mathbf{x})] &= \int_{\mathbf{x}\in\mathbb{R}^N} \int_{c\in\mathcal{C}} \eta_i^2(\mathbf{x}|c)p(\mathbf{x})\, dc\, d\mathbf{x} \\ &= \int_{\mathbf{x}\in\mathbb{R}^N} \eta_i^2(\mathbf{x})p(\mathbf{x})\, d\mathbf{x}. \end{aligned}$$

We generally write η_i instead of $\eta_i(\mathbf{x})$ since the noise component is always dependent on the biometric feature $\mathbf{x}$. This is also true for its class-conditioned counterpart, η_i^k. Note that the same convention applies to y_i and μ_i (so as y_i^k and μ_i^k).

By ignoring the source of distortion in the (extracted) biometric feature space, we actually assume that the noise component is random (while in fact they may be not if we were able to systematically control the conditions c). As before, we write y instead of y_i when referring to any of the participating systems. The noise component is drawn from an unknown distribution W with zero mean and $(\sigma^k)^2$ variance, i.e., $\eta^k \sim W\left(0,(\sigma^k)^2\right)$. It follows that $y_i^k \sim\equiv W\left(\mu^k,(\sigma^k)^2\right)$. Due to the noise model in (4.1), one can characterize the system by the first- and second-order moments, i.e., μ^k and σ^k. While it is unnecessary to assume that the noise is normally distributed at this point of discussion, we will assume so when the integral of the distribution (cumulative density function) is involved. If the system output is not in the LLR domain, one can convert the output to LLR using $f_{LLR}(y)$ (Algorithm 2) in order to ensure that (4.1) is adequate.

Extending from a single system to N systems, the system output vector can be written as $\mathbf{y}^k = [y_1^k,\ldots,y_N^k]'$ whose class-conditional distribution is a multi-variate Gaussian $\mathcal{N}(\mathbf{y}|\boldsymbol{\mu}^k,\boldsymbol{\Sigma}^k)$. The parameters $\boldsymbol{\mu}^k,\boldsymbol{\Sigma}^k$) for $k=\{C,I\}$ are the so-called parametric fusion model. It is a model because it summarizes the problem of fusion. The next two Sections, 4.3.3 and 4.3.4, will rely uniquely on these parameters as input in order to predict the fusion performance. Note that Section 4.3.3 aims to predict the minimal classification error whereas Section 4.3.4 predicts EER. Their difference will be presented in Section 4.3.5.

4.3.3 The Chernoff Bound (for Quadratic Discriminant Function)

Analytically estimating the Bayes error is a classical problem in machine-learning [35]. In a two class problem, following the decision function of (3.4), the probability of making an error given the observation $\mathbf{y}$ is:

$$\begin{aligned} P(error|\mathbf{y}) &= \begin{cases} P(I|\mathbf{y}) & \text{if decision is } accept \\ P(C|\mathbf{y}) & \text{if decision is } reject \end{cases} \\ &= \min\left[p(I|\mathbf{y}),p(C|\mathbf{y})\right]. \end{aligned} \tag{4.2}$$

Note that this is the *minimal* possible error, or *minimal Bayes error* since the decision function

$$P(C|\mathbf{y}) \;>\; P(I|\mathbf{y})$$

(for an accept decision) is optimal. The probability of error is thus:

$$\begin{aligned} P(error) &= \int P(error,\mathbf{y}) \\ &= \int P(error|\mathbf{y})P(\mathbf{y}) \\ &= \int \min\left[p(I|\mathbf{y}),p(C|\mathbf{y})\right]P(\mathbf{y}) \\ &= \int \min\left[\frac{p(\mathbf{y}|I)P(I)}{P(\mathbf{y})},\frac{p(\mathbf{y}|I)P(C)}{P(\mathbf{y})}\right]p(\mathbf{y})d\mathbf{y} \\ &= \int \min\left[p(\mathbf{y}|I)P(I),p(\mathbf{y}|I)P(C)\right]d\mathbf{y}. \end{aligned} \tag{4.3}$$

The probability of error can be expressed in terms of risk as follows:

$$\text{WER} \quad = \quad \int \min\big[p(\mathbf{y}|I)\alpha, p(\mathbf{y}|C)(1-\alpha)\big]d\mathbf{y}, \tag{4.4}$$

where we explicitly introduce WER as defined in (2.4). Note that α includes the dual factor of prior probability (between client and impostor classes) and *normalized costs* (between FRR and FAR) which sum to one. By making use of $\min[a,b] \leq a^\beta b^{1-\beta}$ for $a,b>0$ and $\beta \in [0,1]$, $P(error)$ can be written as:

$$P(error|\beta) \quad \leq \quad P^\beta(I)P^{1-\beta}(C)\underbrace{\int p(\mathbf{y}|I)^\beta p(\mathbf{y}|C)^{1-\beta}d\mathbf{y}}, \tag{4.5}$$

or in terms of risk:

$$\text{WER} \quad \leq \quad \alpha^\beta(1-\alpha)^{1-\beta}\underbrace{\int p(\mathbf{y}|I)^\beta p(\mathbf{y}|C)^{1-\beta}d\mathbf{y}}\,. \tag{4.6}$$

If the class-conditional probabilities are normal, the underbraced term can be evaluated analytically, i.e., $\int p(\mathbf{y}|C)^\beta p(\mathbf{y}|I)^{1-\beta}d\mathbf{y} = \exp(-k(\beta))$, where

$$\begin{aligned} k(\beta) \quad &= \quad \frac{\beta(1-\beta)}{2}(\boldsymbol{\mu}^C - \boldsymbol{\mu}^I)'[\beta\boldsymbol{\Sigma}^I + (1-\beta)\boldsymbol{\Sigma}^C]^{-1}(\boldsymbol{\mu}^C - \boldsymbol{\mu}^I) \\ &\quad + \frac{1}{2}\log\frac{|\beta\boldsymbol{\Sigma}^I + (1-\beta)\boldsymbol{\Sigma}^C|}{|\boldsymbol{\Sigma}^I|^\beta|\boldsymbol{\Sigma}^C|^{1-\beta}}. \end{aligned} \tag{4.7}$$

This quantity is called the Chernoff bound. The minimal Bayes error is given by $\min_\beta P(error|\beta)$. On the other hand, the minimal Bayes error, assuming equal prior (or cost), i.e., $\alpha = 0.5$, is given by $\min_\beta k(\beta)$. The advantage of introducing an upper bound via β is that the search is not dependent on the N dimensional spaces of $\mathbf{y}$ but on a single dimension spanned by β. A special case of error bound, called the Bhattarcharyya bound is given by $k(0.5)$. This quantity is *practical* because it does not involve any numerical search but suffers from a looser estimate of the minimal Bayes error [35, Chap. 2]. Note that these statistics give an *upper* bound of the minimal Bayes error a QDA fusion classifier.

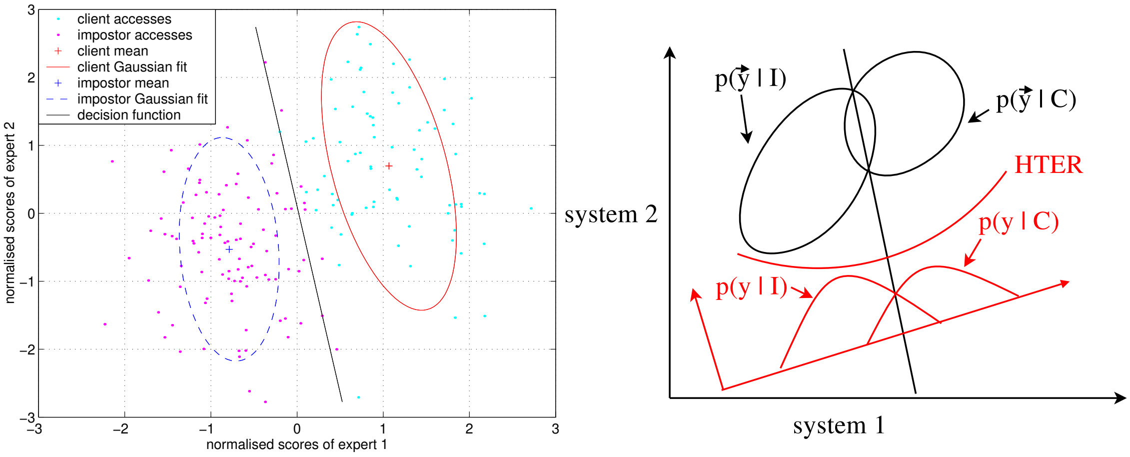
4.3.4 EER of A Linear Classifier

However, in reality, QDA is not used as a fusion classifier. The most commonly used one is perhaps mean or weighted sum, i.e., a linear discriminant function or a linear opinion pool.

To quickly give an intuitive picture, we consider a fusion task consisting of two system outputs after transforming them into the LLR space. The scatter plot of scores are shown in Figure 4.3(a) using the XM2VTS data of one of the fusion tasks described in Section 2.1.1. By summarizing the class-conditional scores (for each class) using a multivariate Gaussian, our goal is to predict the fusion performance. There are two sub-problems to solve. Firstly, one needs to determine the fusion classifier to be used (including its parameters). Having chosen a fusion operator, the second problem consists of calculating the EER analytically. Because of the class-conditional Gaussian assumption, obviously the optimal fusion classifier, according to the LLR test, is to use Quadratic Discriminant Analysis (QDA). We consider the less fortunate (but realistic!) case whereby the parameters of the distribution may not be estimated correctly due to the lack of genuine data and hence QDA is not necessary optimal.

For the case of a linear classifier, Figure 4.3(b) shows that it is possible to project each class-conditional multivariate Gaussian to a single Gaussian. This single Gaussian represents the class-conditional distribution of the *combined* scores.

We will propose a procedure that finds the *exact* solution in terms of EER analytically *without any numerical search*. However, calculating the operational errors other than EER requires a single dimensional numerical search in the combined score space (threshold). In this case, the solution is still *exact* contrary to the Chernoff bound. Then, we will extend such an analysis to other fusion operators, e.g., min, max,



(a) A real example (b) An illustration

Figure 4.3: A geometric interpretation of a parametric model in fusion. (a) A real fusion task whose samples are fitted by two class-conditional bi-variate Gaussian distribution. System 1 is IDIAP's voice system and system 2 is Surrey's automatic face authentication system, applied on the Ud-g1 BANCA data set. (b) A schematic interpretation of projecting from a class-conditional multivariate Gaussian to a single Gaussian.

etc. An application of such analytical technique will be illustrated in Section 4.5 in the user-independent context and its full potential in the user-specific context will be developed in Chapter 7.

To begin, we suppose that a system output may be pre-processed by a linear transformation f_{lin} as in (3.11) so that

$$\mathbf{y}^{norm} = (\mathbf{y} - \mathbf{B}) \;./\; \mathbf{A},$$

where "./" is an element-by-element division and the resultant combined score is

$$y_{COM} = \mathbf{w}'\mathbf{y}^{norm}. \tag{4.8}$$

This generalizes the case where there is no such pre-processing, i.e., the normalizing terms of *each* system take on the values $B_i = 0$ and $A_i = 1$ for all $\mathbf{B} = [B_1, \ldots, B_N]'$ and $\mathbf{A} = [A_1, \ldots, A_N]'$ and $i \in \{1, \ldots, N\}$.

The class-conditional distribution of the combined score y_{COM} using a linear opinion pool as appeared in (4.8) can be written as $\mathcal{N}(\mu^k_{COM}, (\sigma^k_{COM})^2)$ where,

$$\mu^k_{COM} = \sum_{i=1}^{N} \frac{w_i}{A_i} \left(\mu^k_i - B_i\right) \tag{4.9}$$

and

$$(\sigma^k_{COM})^2 = \sum_{m=1}^{N} \sum_{n=1}^{N} \frac{w_m w_n}{A_m A_n} E\left[\eta^k_m \eta^k_n\right] \tag{4.10}$$

respectively, for any $k \in \{C, I\}$, where $E\left[\eta^k_m \eta^k_n\right]$ is the m-th and n-th element of the class-conditional covariance matrix $\mathbf{\Sigma}^k$. The derivations can be found in Section D.2.

If the class-conditional $\mathbf{y}^{norm}$ follows a multivariate Gaussian distribution, then the class-conditional y_{COM} must be 1D Gaussian distribution [120]. It follows that the corresponding FRR and FAR are integrals of Gaussian. We will write y instead of y_{COM} to emphasize the fact that this equation is generally applicable to *any* system output. The derived statistics from y, e.g., μ^k and σ^k, follow the same convention.

The resultant FRR and FAR can be written as:

$$\begin{aligned}
\text{FRR}(\Delta) &= P(\Delta > y|C) = \int_{-\infty}^{\Delta} p(y|C)dy \\
&= \int_{-\infty}^{\Delta} \frac{1}{\sigma^C\sqrt{2\pi}} \exp\left[\frac{-(y-\mu^C)^2}{2(\sigma^C)^2}\right] dy \\
&= \frac{1}{2} + \frac{1}{2}\text{erf}\left(\frac{\Delta - \mu^C}{\sigma^C\sqrt{2}}\right), \text{ and}
\end{aligned} \tag{4.11}$$

$$\begin{aligned}
\text{FAR}(\Delta) &= 1 - P(\Delta > y|I) = \int_{\Delta}^{\infty} P(y|I)dy \\
&= 1 - \left[\frac{1}{2} + \frac{1}{2}\text{erf}\left(\frac{\Delta - \mu^I}{\sigma^I\sqrt{2}}\right)\right] \\
&= \frac{1}{2} - \frac{1}{2}\text{erf}\left(\frac{\Delta - \mu^I}{\sigma^I\sqrt{2}}\right),
\end{aligned} \tag{4.12}$$

where

$$\text{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z \exp\left[-t^2\right] dt,$$

which is known as the "error function" in statistics.

The minimal error happens when FAR(Δ) = FRR(Δ) = EER, i.e., the Equal Error Rate. Making these two terms equal – (4.11) and (4.12) – and using the property that $\text{erf}(-z) = -\text{erf}(z)$, we can deduce that:

$$\Delta = \frac{\mu^I\sigma^C + \mu^C\sigma^I}{\sigma^I + \sigma^C}. \tag{4.13}$$

By introducing (4.13) into (4.12) (or equivalently into (4.11)), we obtain:

$$\text{EER} = \frac{1}{2} - \frac{1}{2}\text{erf}\left(\frac{\text{F-ratio}}{\sqrt{2}}\right) \equiv \text{eer}(\text{F-ratio}), \tag{4.14}$$

where we introduced F-ratio, defined as:

$$\text{F-ratio} = \frac{\mu^C - \mu^I}{\sigma^C + \sigma^I}. \tag{4.15}$$

Note that the use of an error function similar to F-ratio was reported in [22], but with differences in the definition of the error function. In another similar work (but in the context of combining multiple samples) [67], EER was not calculated explicitly.

Other Class-Separability Measures

It should be noted that the term "F"-ratio is used here because this value is somewhat *similar to* the standard Fisher ratio, but not defined exactly in the same way. In a two-class problem, the Fisher ratio [11, pg. 107] is defined as:

$$\frac{\mu^C - \mu^I}{(\sigma^C)^2 + (\sigma^I)^2}. \tag{4.16}$$

F-ratio is used here just to underpin the idea that the degree of separability of the class distribution affects the authentication performance measured by EER. There exists similar measures such as the *d-prime* metric proposed by Daugman [29]. It measures how separable the client distribution is from its impostor counterpart. It is defined as:

$$d' = \frac{|\mu^C - \mu^I|}{\sqrt{\frac{1}{2}(\sigma^C)^2 + \frac{1}{2}(\sigma^I)^2}}. \tag{4.17}$$

Besides the abovementioned quantities, in [71], three other similar quantities used in texture classification were also considered for biometric authentication, i.e.,:

$$J_1 = \frac{\mu^C}{\mu^I}, J_2 = \frac{(\mu^C - \mu^I)^2}{\mu^C \mu^I} \text{ and } J_3 = \frac{(\mu^C - \mu^I)^2}{(\sigma^C)^2 + (\sigma^I)^2}.$$

F-ratio will be used throughout this thesis because it is directly related to EER by (4.14).

Summary of Results

We gather here several important results presented so far. From (4.9) and (4.10), one knows how to calculate the first- and second-order moments of the combined score y_{COM}, i.e., μ^k_{COM} and $(\sigma^k_{COM})^2$. Based on these four Gaussian parameters $\{\mu^k_{COM}, \sigma^k_{COM}\}$ for both $k = \{C, I\}$, the F-ratio of the combined score y_{COM}, according to (4.15) is:

$$\text{F-ratio}_{COM} = \frac{\mu^C_{COM} - \mu^I_{COM}}{\sqrt{V^C_{diag} + V^C_{ndiag}} + \sqrt{V^I_{diag} + V^I_{ndiag}}}, \tag{4.18}$$

where

$$V^k_{diag} = \sum_{i \in [1,N]} \frac{w_i w_i}{A_i A_i} E\left[\eta^k_i \eta^k_i\right]$$

and

$$V^k_{ndiag} = \sum_{i,j \in [1,N], i \neq j} \frac{w_m w_n}{A_m A_n} E\left[\eta^k_m \eta^k_n\right]$$

are respectively the diagonal and non-diagonal sum of the covariance matrix $\mathbf{\Sigma}^k$ whose i-th and j-th element is denoted as $E\left[\eta^k_m \eta^k_n\right]$. The corresponding theoretical EER will be $\mathrm{eer}(\text{F-ratio}_{COM})$ as defined in (4.14).

From (4.18), three factors can be identified to influence the performance of the fusion performance. They are:

1. **The mean difference ($\mu^C_{COM} - \mu^I_{COM}$) :** Higher mean difference improves the system performance.
2. **The diagonal component (V^k_{diag}):** This term measures, on average, how good the base-systems are, when acting alone. Note that by definition, $V^k_{diag} \geq 0$. Lower variance is desirable.
3. **The non-diagonal component (V^k_{ndiag}):** This term is influenced by the pair-wise correlations $\rho^k_{m,n}$ for $m, n \in \{1, \ldots, N\}$ and therefore can be positive or negative since $-1 \leq \rho^k_{m,n} \leq 1$ for any pair of systems m, n. Lower covariance or even negative V^k_{ndiag} is desirable.

Note that the second and third factors cannot be separated since they are tied by a common square-root. The reason we separated the weighted sum of the covariance matrix into V^k_{diag} and V^k_{ndiag} is to show explicitly that V^k_{ndiag} is directly dependent on the pair-wise correlation. Therefore, correlation is a required but not sufficient condition to predict the fusion performance. This claim is verified in Section C.3 using real datasets.

4.3.5 Differences Between the Minimal Bayes Error and EER

It is important to distinguish between the Chernoff bound presented in Section 4.3.3 and our proposed EER calculated based on the F-ratio in Section 4.3.4. They differ in the following ways:

- **Definition:** Figure 4.4 illustrates the difference between the minimal Bayes error and EER from their definitions. From (4.5), the minimal Bayes error is:

$$\int \min\left[p(\mathbf{y}|I)P(I), p(\mathbf{y}|C)P(C)\right] d\mathbf{y}.$$

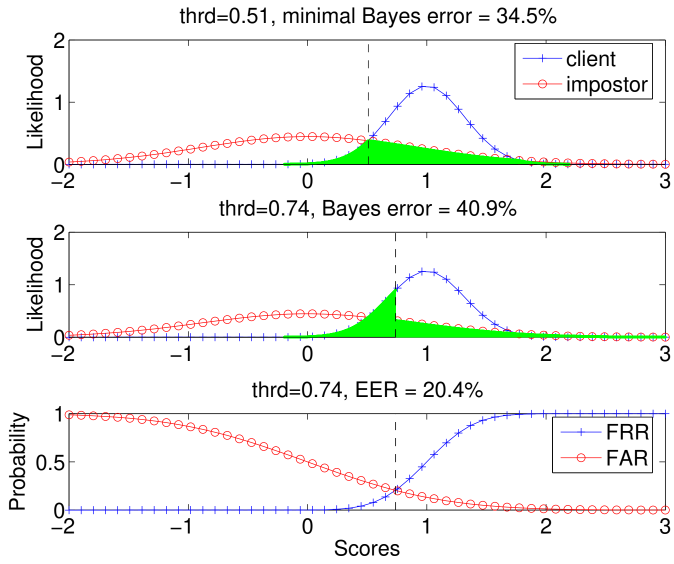


Figure 4.4: The difference between minimal Bayes error and EER. For this example, equal class prior probabilities are assumed, i.e., $P(C) = P(I)$. For all the figures, the Y-axis is the score combined score and the X-axis is likelihood or probability. The top figure shows the minimal Bayes error. The middle figure shows the Bayes error due to EER. The bottom figure shows how the EER criterion, i.e., FAR=FRR, is fulfilled. Due to equal prior probabilities, in this case, the Bayes error at EER is $2 \times$ EER.

Therefore, this expression minimizes the *overlap* of the two posterior distributions, i.e., $P(k|\mathbf{y}) \propto P(\mathbf{y}|k)P(k)$, for $k = \{C, I\}$. On the other hand, EER by definition is FAR$(\Delta) =$ FRR(Δ) or

$$\int p(\mathbf{y}|I)d\mathbf{y} = 1 - \int p(\mathbf{y}|C)d\mathbf{y}.$$

The constraint ensures that the overlap between the two class-conditional distributions are equal. Note that EER does not take the class prior probability into consideration whereas the Bayes error does. For the example in Figure 4.4, equal class prior probabilities are assumed, i.e., $P(C) = P(I)$. In this case, the Bayes error at EER is $2 \times$ EER.

- **Bound or exact error:** The Chernoff bound is, at best, only an upper bound of the theoretically minimal classification error. On the other hand, the EER is an exact estimate.

- **Quadratic or linear classifier:** The Chernoff bound is only indicative of the Bayes error of a quadratic classifier (which includes LDA as a special case). On the other hand, the proposed EER applies to *any* linear classifier, e.g., SVM with a linear kernel, logistic regression, the Perceptron algorithm, the LDA classifier (based on the Fisher ratio), etc. This is thank to the property that a multivariate Gaussian is closed under a linear transform, as discussed in Section 3.5.

To the best of our knowledge, this is the first time in the literature where an analytical expression of EER for fusion is proposed.

4.3.6 Validation of the Proposed Parametric Fusion Model

Since F-ratio is based on the class-conditional Gaussian assumption – an assumption that is likely to be violated –, it is thus important to verify if the EER calculated based on F-ratio is acceptable or not. The "level of acceptability" can be quantified by the difference between the *theoretical* EER (due to applying (4.15)) and the *empirical* EER (that is measured directly on the observed data). This experiment is reported in Section C using 1186 BANCA score sets. We summarize the findings here:

- Despite deviation from the Gaussian assumption, the theoretical EER (derived from F-ratio) correlates well with the empirical EER, i.e., 0.957 for all the 1186 datasets.
- The error estimates at the extreme ends (low FAR or high FAR costs) are less accurate than EER.
- Relaxing the class-conditional Gaussian assumption improves the error estimates.

4.4 Why Does Fusion Work?

4.4.1 Section Organization

This section aims to explain theoretically the phenomenon observed in Section 4.2, i.e., the combined system works better than the average performance of systems working individually. Section 4.4.2 summarizes the literature that attempts to explain theoretically the mentioned phenomenon and explains why the current literature is not adequate. In the justification, an additional step is required to align the system outputs. This step is explained in Section 4.4.3 and has important consequences on Chapter 7. Section 4.4.4 then demonstrates the reduction of classification error due to combining several systems using the mean operator and a brief explanation of how this can be done for the weighted sum case.

4.4.2 Prior Work And Motivation

Although fusion in the context of biometric authentication has been discussed elsewhere, there is still a lack of theoretical understanding, particularly with respect to correlation. The *correlation* here refers to the *pairwise class-conditional correlation* between the outputs of any two participating systems. We review several theoretical studies here:

- In [57], it was demonstrated that combining several multimodal system scores using AND and OR will result in improved performance. The underlying assumption is that multimodal system scores are independent. As we understood, the issue of relative performance among systems and the strategy of choosing the decision threshold *prior to fusion* were not thoroughly considered.
- In [73], the theoretical classification error of six classifiers are thoroughly studied for a two-class problem. This study assumes that the underlying classifier outputs are probabilities, i.e., $P(y|C)$ using our notation (see Chapter 3.2). Therefore, regardless of the cost of FAR or FRR, the optimal threshold is always set to 0.5. The study also assumes that all the participating system outputs follow a common distribution. Gaussian and uniform distributions were used in this study. This assumption is unfortunately unrealistic in most situations, particularly in multimodal fusion. This is because the (class-dependent) score distributions are often *different* across different systems.
- In [143], order statistics (OS) combiners, i.e., $\min$, $\max$ and median, are examined both theoretically and empirically. The authors introduced the concept of biased and unbiased classifier, which is the same as mismatch between training and test sets as observed by the system outputs. While the analysis in [143] is certainly interesting, there is no direct way of inferring the overall classification performance given a data set. It is also unclear how correlation affects the OS operators.
- In [66], sum and product rules were discussed in a Bayesian framework. According to this study, several fixed rules such as $\min$, $\max$, median and majority vote can be seen as approximations to the aforementioned rules. In particular, it was shown that the sum rule (or mean in our context) outperforms the rest of the fixed rules and even better than the single best underlying system. A further investigation showed that the sum rule is most resilient to estimation error of individual classifier than the product rule. Similar to [73] this study, too, assumes *common probability distribution* which is likely violated in reality.
- In [76, Chap.10], product and sum rules were studied by taking into condieration of the mismatch between training and test sets. The conclusion is similar to that of [66]. Again, the analysis assumes that the underlying classifier/system outputs are independent. This assumption is acceptable for multimodal fusion but inadequate for intramodal fusion.

- A more recent study, [123], considers correlation, unbalanced performance among participating systems and biased system outputs.

Note that these prior works, except [123], make simplifying assumptions in one way or another, e.g., common distribution for all the underlying systems and independence assumption of system outputs.

The goal of the following Section is to provide a very simple parametric fusion model that precisely takes the mentioned factors into consideration. This is done in LLR, instead of probability as in [57, 73, 143, 66, 76, 123].

4.4.3 From F-ratio to F-Norm

We now introduce a useful normalization derived from F-ratio that we call F-norm It is used to simplify the proof of EER reduction in Section 4.4.4. It is also used extensively in user-specific processing. F-norm is introduced here because of its frequent usage.

Motivation to Align Scores using Z-norm as An Example

Because different system types are used, the deterministic component μ_i^k for all $i = 1, \ldots, N$ and $k = \{C, I\}$ are not necessarily the same. As a result, the combined system output using simple fusion rules will be biased toward the system with large output values. This will cause a sub-optimal fusion performance. One way to align them using a linear function such as f_{lin} appeared in (3.11). For Z-norm, the scaling factor and bias are $A = \sigma_i^I$ and $B = \mu_i^I$, respectively for each i (see Section 3.3.2). By doing so, one obtains:

$$y_i^Z = \frac{y_i - \mu_i^I}{\sigma_i^I}.$$

Because y_i^k is (or assumed to be) approximately normally distributed, it follows that $y_i^{Z,k}$ is the case too, with the class-conditional mean and variance:

$$\mu_i^{Z,k} \equiv E[y_i^Z] = \frac{E[y_i|k] - \mu_i^I}{\sigma_i^I} = \frac{\mu_i^k - \mu_i^I}{\sigma_i^I}.$$

$$(\sigma_i^{Z,k})^2 = \frac{(\sigma_i^{Z,k})^2}{(\sigma_i^{Z,I})^2}.$$

Note that while the resultant impostor distribution is *standard* normal ($\mu_i^{Z,I} = 0$, $(\sigma_{Z,i}^I)^2 = 1$) for all i, the resultant client distribution varies from one system to another ($\mu_i^{Z,C} = \frac{\mu_i^C - \mu_i^I}{\sigma_i^I}, (\sigma_i^{Z,C})^2 = \frac{(\sigma_i^C)^2}{(\sigma_i^I)^2}$). As a result, such a normalization procedure is not satisfactory.

Derivation of F-norm

A reasonably good procedure should align the system outputs such that the expected means (the deterministic components) of the client and impostor distributions are the same. One way to achieve this is by imposing the following constraint, based on F-ratio:

$$\frac{\mu_i^C - \mu_i^I}{\sigma_i^C + \sigma_i^I} = \frac{1 - 0}{{\sigma'}_i^C + {\sigma'}_i^I}, \tag{4.19}$$

where the numerator of the RHS term is the *desired* difference in mean and the denominator is the sum of standard deviations as a result of the desired transformation. Solving this constraint yields:

$${\sigma'}_i^k = \alpha \sigma_i^k, \tag{4.20}$$

where $\alpha = (\mu_i^C - \mu_i^I)^{-1}$. Using the definition of variance and taking the square of (4.20), we obtain:

$$({\sigma'}_i^k)^2 = E\left[\left(\alpha(y_i - \mu_i^k)\right)^2\right]. \tag{4.21}$$

Note that the factor α is not dependent on y_i. This implies that the desired transformation due to the constraint of (4.19) should take the form $\frac{y}{\mu_i^C - \mu_i^I}$. However, this constraint does not guarantee zero impostor mean. To do so, we introduce a subtraction term μ_i^I to obtain F-norm:

$$y_i^F = \frac{y - \mu_i^I}{\mu_i^C - \mu_i^I}. \tag{4.22}$$

Characteristics of F-norm

We verify that the following constraints are fulfilled (by design):

$$\mu_i^{F,C} \equiv E[y^F|C,i] = \frac{E[y|C,i] - \mu_i^I}{\mu_i^C - \mu_i^I} = 1, \text{ for all } i \tag{4.23}$$

and

$$\mu_i^{F,I} \equiv E[y^F|I,i] = \frac{E[y|I,i] - \mu_i^I}{\mu_i^C - \mu_i^I} = 0, \text{ for all } i \tag{4.24}$$

The corresponding class-conditional standard deviation is: $\sigma_i^{F,k} = \frac{\sigma_i^k}{\mu_i^C - \mu_i^I}$ as implied by (4.20).

Differences Between Z-norm and F-norm

It is not immediately obvious why F-norm is *better* than Z-norm. Following our empirical experiments reported in [101], the generalization performance of Z-norm and F-norm are not statistically significantly different between the two procedures. However, the advantage will become apparent in the user-specific context (Chapter 7). One reason is that the alignment due to F-norm is client-impostor centric, i.e., making use of both the genuine and impostor distributions, whereas Z-norm is only impostor centric, i.e., making use of only the impostor distribution.

We introduce F-norm here so that after applying such procedure, one needs only to focus on $\sigma_i^{F,k}$ for all i without worrying about the alignment problem.

4.4.4 Proof of EER Reduction with Respect to Average Performance

We have demonstrated that making the class-conditional Gaussian assumption is somewhat acceptable on real biometric authentication problems, thanks to the robustness of Gaussian assumption. To the best of our knowledge, such a demonstration (using EER) has not been reported elsewhere in the literature for *classification problems* but is well known for *regression problems*, e.g., [11, Chap. 9]. It should be mentioned that in [123], a proof along similar line was reported for classification problems but the error term used in the demonstration is not EER but the so-called "added error"[2].

A Sketch of the proof

A sketch of the approach is shown in Figure 4.5. Suppose that F-norm is first applied to all system outputs so that their expected values are the same, i.e., $\mu_i^C = 1$ and $\mu_i^I = 0$ for any $i \in [1, \ldots, N]$. Then, we show that due to fusion, the class-conditional variance is reduced – which is the first part of the proof. Consequently, the resultant EER is reduced – which is the second part of the proof. For the proof, we will first consider the special case of mean operator and then provide a sketch for the general case of weighted sum.

[2] This term is due to bias between the approximated class posterior and the actual posterior not available during training. In this sense, the bias is due to mismatch between training and test sets. This subject of noise mismatch is treated in Section 4.6.5.

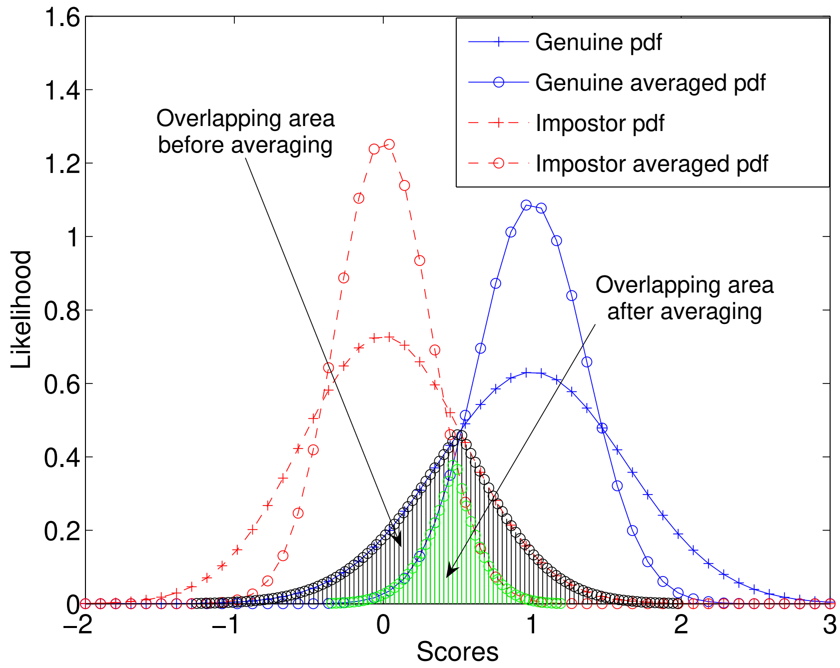


Figure 4.5: A sketch of EER reduction due to the mean operator in a two-class problem

Variance Reduction

Let us consider two cases here. In the first case, for each access, N system outputs are available and are used independently of each other. The *average of variance* of y_i^k over all $i = 1, \dots, N$, denoted as $(\sigma_{AV}^k)^2$ is, according to [103]:

$$\left(\sigma_{AV}^k\right)^2 \quad = \quad \frac{1}{N}\sum_{i=1}^{N}\frac{E[\eta_i^k \eta_i^k]}{(A_i)^2} \equiv \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\sigma_i^k}{A_i}\right)^2, \tag{4.25}$$

where $A_i = \mu_i^C - \mu_i^I$.

In the second case, all N responses are used together and are combined using the mean operator so that one obtains y_{COM}. Note that because $\mu_i^k = \mu_j^k$ for any $i, j \in [1, \dots, N]$, $\mu_{COM}^k = \mu_i^k$ for any i. The variance of y_{COM}^k, denoted as $(\sigma_{COM}^k)^2$, is called the *variance of average*. Based on (4.10) (with $w_i = \frac{1}{N}$), its value is:

$$\begin{aligned}
\left(\sigma_{COM}^k\right)^2 &= \frac{1}{N^2}\sum_{i=1}^{N}\left(\frac{\sigma_i^k}{A_i}\right)^2 + \frac{2}{N^2}\sum_{m=1,m<n}^{N}\frac{\rho_{m,n}^k\sigma_m^k\sigma_n^k}{A_m A_n}, \\
&= \underbrace{\frac{1}{N}\left(\sigma_{AV}^k\right)^2} + \underbrace{\frac{2}{N^2}\sum_{m=1,m<n}^{N}\frac{\rho_{m,n}^k\sigma_m^k\sigma_n^k}{A_m A_n}}, \\
&\equiv V_{diag}^k + V_{ndiag}^k, \qquad (4.26)
\end{aligned}$$

where we separated the matrix sum involving Σ^k (whose element is $E[\eta_m^k, \eta_n^k]$) into a diagonal term (V_{diag}^k) and a non-diagonal term (V_{ndiag}^k). Note that V_{diag}^k is always positive whereas V_{ndiag}^k can be a negative value. Note also that $\rho_{m,n}^k$ is the correlation coefficient between y_m^k and y_n^k for $k \in \{C, I\}$ and it is defined by:

$$\rho_{m,n}^k\sigma_m^k\sigma_n^k = E[\eta_m^k\eta_n^k], \tag{4.27}$$

with the property that $\rho^k_{n,n} = 1$ for $k \in \{C, I\}$. Because $V^k_{diag} \geq 0$, it follows that $(\sigma^k_{COM})^2 \geq V^k_{diag}$ or $(\sigma^k_{COM})^2 \geq \frac{1}{N}\left(\sigma^k_{AV}\right)^2$. We can also show that $(\sigma^k_{COM})^2 \leq (\sigma^k_{AV})^2$ (see Section D.3). As a result, we have:

$$\frac{1}{N}\left(\sigma^k_{AV}\right)^2 \leq \left(\sigma^k_{COM}\right)^2 \leq \left(\sigma^k_{AV}\right)^2. \tag{4.28}$$

Hence, by combining N responses using the mean operator, the resulting variance is assured to be smaller than the average (not the minimum) variance.

EER Reduction

In order to show that the EER of the combined scores is lower than the average EER over N outputs, i.e.,

$$\text{EER}_{COM} \leq \text{EER}_{AV}, \tag{4.29}$$

we first need to calculate μ^k_p and σ^k_p for $k = \{C, I\}$ and $p = \{COM, AV\}$. $\sigma^k_p|p = \{COM, AV\}$ have been defined by (4.25) and (4.26), respectively. μ^k_{AV} is the average of N responses when used separately. It is defined as:

$$\mu^k_{AV} \equiv \frac{1}{N}\sum_{i=1}^{N}\frac{\mu^k_i - B_i}{A_i}, \tag{4.30}$$

where A_i and B_i are the parameters due to F-norm. μ^k_{COM} is the mean of the combined scores of N responses (used simultaneously). It is defined as:

$$\begin{aligned} E[y^k_{COM}] \quad &\equiv \quad \mu^k_{COM} = \frac{1}{N}\sum_{i=1}^{N}\frac{E[y^k_i] - B_i}{A_i} \\ &= \quad \frac{1}{N}\sum_{i=1}^{N}\frac{\mu^k_i - B_i}{A_i} = \mu^k_{AV}. \end{aligned} \tag{4.31}$$

Hence, $\mu^k_{COM} = \mu^k_{AV}$. Since F-ratio is non-linearly and inversely proportional to EER as shown in (4.14), the inequality of (4.29) can be rewritten as:

$$\text{F-ratio}_{COM} \geq \text{F-ratio}_{AV}, \tag{4.32}$$

Replacing the two F-ratio terms using (4.31) and (4.30) into (4.32) and using the relation $\mu^k_{COM} = \mu^k_{AV}$, we obtain:

$$\begin{aligned} \frac{\mu^C_{COM} - \mu^I_{COM}}{\sigma^C_{COM} + \sigma^I_{COM}} &\geq \frac{\mu^C_{AV} - \mu^I_{AV}}{\sigma^C_{AV} + \sigma^I_{AV}} \\ \sigma^C_{COM} + \sigma^I_{COM} &\leq \sigma^C_{AV} + \sigma^I_{AV} \\ \sum_{\{C,I\}}\sigma^k_{COM} &\leq \sum_{\{C,I\}}\sigma^k_{AV} \end{aligned} \tag{4.33}$$

Hence, the inequality of (4.29) is true, i.e., fusing scores can reduce variance which results in reduction of EER (with respect to the case where scores are used separately). This formed the argument for why fusion using multiple modalities, features, and classifiers works for biometric authentication tasks. Note that this observation is in perfect agreement with the empirical experiments in Section 4.2, especially Figure 4.1(a).

Extending the Proof to Weighted Sum

Note that a similar proof for fusion using weighted sum instead of mean can be demonstrated as well. Such a proof will lead to the form:

$$\sum_{\{C,I\}}\sigma^k_{wsum} \leq \sum_{\{C,I\}}\sigma^k_{COM} \leq \sum_{\{C,I\}}\sigma^k_{AV}$$

where σ^k_{wsum} is the class-conditional variance due to weighted sum fusion. Note that such a proof requires that the weight parameters to be estimated correctly, a requirement that is quite restricted to have any practical value. An involved discussion can be found in [11, Chap. 9]. Instead, we will demonstrate that weighted sum is better than mean by simulation in Section 4.6.2.

4.5 On Predicting Fusion Performance

In order to demonstrate the potential of the parametric fusion model discussed so far, in this section, we outline an approach to *analytically* select a subset of systems for fusion. The weighted sum fusion classifier will be used as it is somewhat optimal for the data sets available to us, i.e., the same datasets as those used in Section 2.1.2. The task is to choose out of the $N = 5$ systems, a combination of them that will give an optimal result, without degrading the performance significantly compared to using *all* the sub-systems. In other words, we want to trade-off *insignificant performance gain* with lower computation cost. Note that this is a combinatory problem with $2^N - 1$ possibilities (minus one for the case where not choosing any system is not a valid option).

The brute-force approach to the solution, typically adopts the following procedure:

1. For each of the possible combinations:
   - Estimate the best (weight) parameters from the development set according to a criterion (such as Mean Squared Error)
   - Use the weights to evaluate the performance on the development set
2. Choose the best fusion candidate based on the evaluated performance.

Our proposed analytical solution works as follow:

1. Estimate μ^k, $\mathbf{\Sigma}^k$, for each $k \in \{C, I\}$.
2. For each of the possible combinations:
   - Estimate the weights $\mathbf{w}$ given $\{\mu^k, \mathbf{\Sigma}^k\}$ for $k \in \{C, I\}$. The weights can be found using the Fisher linear discriminant solution as appeared in (3.25).
   - Evaluate the F-ratio given $\mathbf{w}$ and the model parameters
3. Choose the best fusion candidate that maximizes F-ratio.

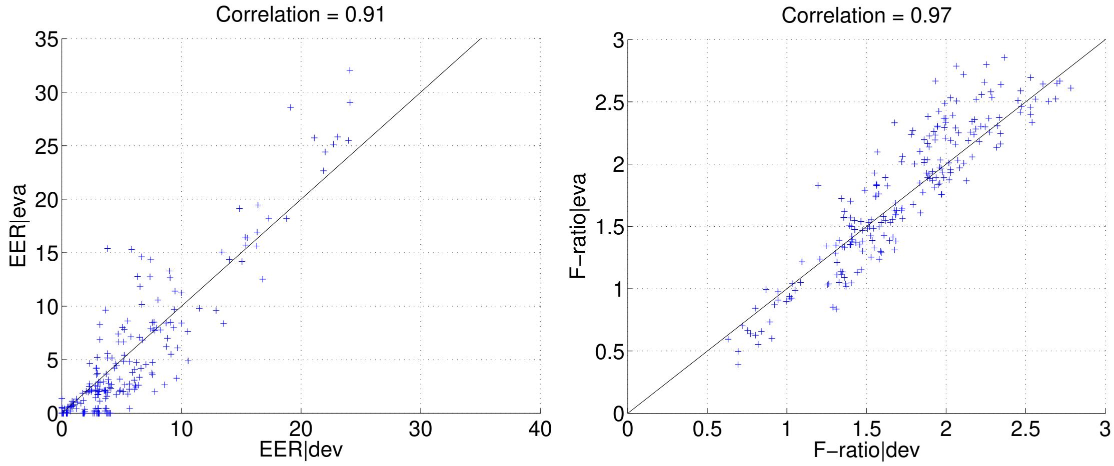
In the brute-force approach, to choose one best fusion candidate from all possible N base-systems, one would have to carry out the experiment $2^N - 1$ times (or 2^N). In each experiment, one has to loop through l examples. The complexity is thus:

$$O\left(l \times (2^N)\right). \tag{4.34}$$

In the proposed approach, one only has to loop through the data set once to derive all model parameters and then to evaluate the F-ratio criterion $2^N - 1$ times on the evaluation set. Hence, the complexity is thus:

$$O\left(l + 2^N\right). \tag{4.35}$$

To understand why such an analytical procedure can work, we measured the F-ratio of the combined scores of the development set versus its evaluation set counterpart. For comparison, we also performed the same experiments but this time empirically and the performance for both the development and evaluation sets are measured using *a posteriori* EER. Because there are 5 systems per experimental protocol (hence $2^5 - 1 = 31$ combinations) and there are altogether 7 BANCA protocols, there are altogether $31 \times 7 = 217$ F-ratio pairs. The results are plotted in Figure 4.6. As can be observed, compared to the empirical EER, F-ratio has a higher correlation than EER. Note that in the BANCA database, the development and evaluation datasets are from two *completely different* sets of population. Therefore, an additional advantage of F-ratio is its robustness to the population mismatch.



(a) Empirical EER (b) F-ratio

Figure 4.6: Comparison of empirical EER and F-ratio of the combined scores with respect to robustness to population mismatch between training and test data set. In both figure, the X-axis are EER or F-ratio of the development set whereas the Y-axis are the same measurements on the evaluation set. Each point is one of the 31 possible combinations per protocol and there are 7 protocols, hence, there are $31 \times 7 = 217$ data points. Note the improved correlation from (a) to (b).

In [102], we showed that the quality of prediction is satisfactory. Taking the evaluation set as the ground-truth, the top three proposed combination of fusion candidates *always* contain the ground truth combination, for all the seven BANCA protocols. It should be mentioned that the top three fusion candidates contain rather similar EER values. Hence, choosing any of the top three solutions cannot significantly influence the generalization performance.

4.6 An Extensive Analysis of Mean Fusion Operator

4.6.1 Motivations and Section Organization

The demonstration in Section 4.4 can only show that a combined system is better than the average performance of its underlying systems. Ideally, a more desirable result is to know when the combined system is better than the *best* system. To the best of our knowledge, such a more desirable result has not been found in the literature. While a general result is not possible, we will consider the special case of combining two system outputs using the mean fusion operator here. This is actually not a limitation as generalizing to more than two fusion operators is straightforward. Section 4.6.2 is our attempt to work towards identifying the necessary conditions.

We are also motivated by the improved understanding of noise mismatch in regression problems, e.g., [69, 144]. However, until now, the consequence of noise in classification, also known as bias, is not well known. Although this subject has been treated in [123], there is no way one can make use of the findings in regression to classification directly. By working in the LLR space, we will show that the noise mismatch model in regression, as proposed by [69, 144], can be used in binary classification problems. Working towards this direction, Sections 4.6.3 and 4.6.4 review the works of [69] on the ambiguity decomposition and of [144] on the bias-variance-covariance decomposition, respectively. Finally, Section 4.6.5 extends the noise mismatch model to binary classification by using the already proposed parametric fusion model in Section 4.3.2. A useful finding from our study is that the bias introduced by the noise can possibly be rectified.

4.6.2 Effects of Correlation and Unbalanced System Performance on Fusion

Suppose that the mean operator is used to combine scores under the following scenarios:

1. Combining 2 uncorrelated system outputs with very different performance
2. Combining 2 highly correlated system outputs with very different performance
3. Combining 2 uncorrelated system outputs with very similar performance
4. Combining 2 highly correlated system outputs with very similar performance

The first and third cases are often encountered in multimodal fusion while the second and fourth cases are encountered in intra-modal (multi-feature) fusion. Fusing systems of similar and different performances are encountered in almost all biometric authentication problems. It should be noted that empirical evidences of these scenarios have been examined in [133] but unfortunately there was a lack of theoretical explanation. To make analysis simple, let us assume that (i) the two base-systems have the same numerator of F-ratio and that (ii) for each base-system, the variance and covariance of client and impostor distributions are proportional. The first assumption can be taken care of by using F-norm (see Section 4.4.3). The second assumption implies that $\sigma_i^C \propto \sigma_i^I$ for system $i \in \{1,2\}$ as well as their covariance

$$\rho^C \sigma_1^C \sigma_2^C \propto \rho^I \sigma_1^I \sigma_2^I.$$

This simplifies the analysis so that one considers only one class at a time. An empirical justification of the second assumption can be found in Figure C.5(c). Hence, the class label k can be dropped. Without loss of generality, we assume $\sigma_1 \leq \sigma_2$ (i.e., system 1 is better than system 2).

For the first case, $\rho \simeq 0$. Hence, for the combination to be *better than the best system*, i.e., system 1, it is required that:

$$\begin{aligned} \sigma_{COM}^2 &< \sigma_1^2 \\ \frac{\sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2}{4} &< \sigma_1^2 \end{aligned} \tag{4.36}$$

σ_{COM}^2 is calculated using (4.26) with N=2.

We see that:

$$\sigma_2^2 < 3\sigma_1^2 - 2\rho\sigma_1\sigma_2.$$

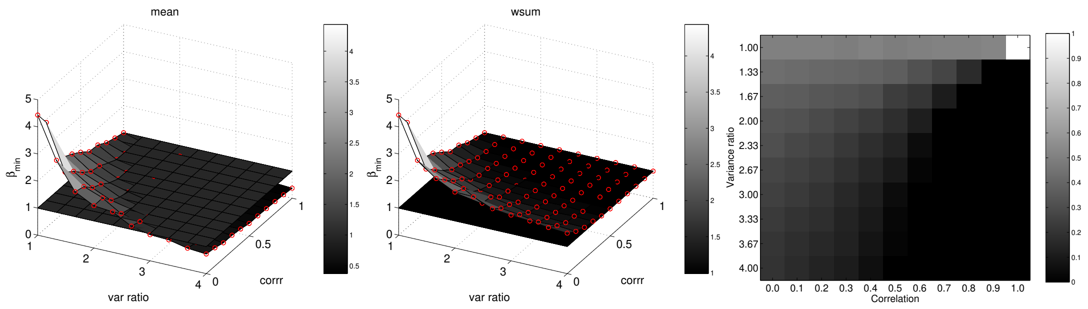
Note that in general, $\rho \geq 0$. For instance, in multimodal fusion, ρ is around zero while in multi-feature fusion, ρ is positive. Hence, the combined system will benefit from the fusion when σ_2^2 is *at most* less than 3 times of σ_1^2 since $\rho \simeq 0$.

Furthermore, correlation (or equivalently covariance; see (4.27)) between the two systems penalizes this margin of $3\sigma_1^2$. This is particularly true for the second case since $\rho > 0$. Also, it should be noted that $\rho \leq 0$ (which implies negative correlation) could allow for larger σ_2^2. As a result, adding another system that is negatively correlated, but with large variance (hence large EER) *will* improve fusion. Unfortunately, in biometric authentication, 2 systems are either positively correlated or not correlated, unless these systems are *jointly trained* together by algorithms such as negative correlation learning [13].

For the third and fourth cases, we have $\sigma_1^2 \simeq \sigma_2^2$. Hence, (4.36) becomes

$$\rho\sigma_2^2 < \sigma_1^2. \tag{4.37}$$

Note that for the third case, $\rho \simeq 0$ which will satisfy the constraint of (4.37). Therefore, fusion will *definitely* lead to better performance. On the other hand, for the fourth case where $\rho \simeq 1$, according to (4.37), fusion may not necessarily lead to statistically significantly better performance – suggesting that using only the better system may be appropriate.



(a) mean (b) wsum (c) weights

Figure 4.7: Comparison between the mean operator and weighted sum using synthetic data. Performance gain of in terms of EER with respect to the best underlying classifier, β_{min} (Z-axis), across different variance ratios (of two system outputs; X-axis) and different correlation values (Y-axis), as a result of fusing synthetic scores of two systems. The scores are combined using (a) mean and (b) weighted sum. (c): the weight of the *weaker* system found in the weighted sum after training. This can be thought of as the degree of "reliance on the weaker system".

Experimental Simulation

In order to support the theoretical analysis here, we performed a simulation. $\sigma_1 = 0.5$ whereas σ_2 varies from 0.5 to 2. The correlation value varies from 0 to 1 by a step of 0.1. While σ_i and ρ vary, the deterministic components are held constant $\mu^C = 1$ and $\mu^I = 0$ (the system outputs are aligned). This simulation produces a set of fusion tasks completely specified by the matrix $(\frac{\sigma_2}{\sigma_1}, \rho)$ (variance ratio and correlation). The first system has HTER between 5.3% and 6.2%, with a mean of 5.8% and the second system has HTER between 5.4% and 22% of HTER with a mean of 15% at the EER point. We then employ two fusion classifiers, mean and weighted sum whose weights are tuned to minimize EER empirically.

We plot the result (see Figure 4.7) as $(\frac{\sigma_2}{\sigma_1}, \rho, \beta_{min})$ where the Z-axis is the gain with respect to the single best system (see (2.11)). Note that $\beta_{min} <= 1$ implies that fusion results in worse performance. For the mean rule, we observe that at $(3, 0)$ (in the variance ratio and correlation space), $\beta_{min} = 0$. When $(3, \rho \geq 0)$, $\beta_{min} \leq 0$. On the other hand, the weighted sum operator does not suffer from such situation as the weight parameters can be adjusted accordingly. As a result, for the weighted sum operator, $\beta_{min} \geq 1$ in all possible values of $(\frac{\sigma_2}{\sigma_1}, \rho, \beta_{min})$. Of course, this is an overly optimistic result because we assume that the weight parameters can be estimated correctly.

4.6.3 Relation to Ambiguity Decomposition

We would like to link our findings with those of Krogh and Vedelsby [69] (see also [11, pages 368]). Note that the authors' finding *applies* only to the regression problem and does not directly offer an explanation to the same phenomenon in classification because in classification, the statistics of client and impostor distributions have to be considered *simultaneously*. Nevertheless, the authors' finding is an important precursor to the EER we proposed in Section 4.3.4. Using our notations, the authors showed that:

$$\begin{aligned} E[y^k_{COM'} - \mu^k_{COM'}]^2 &= \sum_i w_i E\left(y^k_i - \mu^k_{COM'}\right)^2 - \sum_i w_i E\left(y^k_i - y^k_{COM'}\right)^2 \\ (\sigma^k_{COM})^2 &\equiv \mathrm{acc}^k - \mathrm{div}^k, \end{aligned} \tag{4.38}$$

where w_i are the weights in weighted sum combination, $y^k_{COM'}$ is the output of the unnormalized combined system and $\mu^k_{COM'}$ is its expected value. Note that $w_i = 1/N$ because we are using the mean operator instead of weighted sum. The first term, denoted as acc (or "accuracy"), measures how accurate each base-system is with respect to the mean score of the combined mechanism. It depends only on the individual base-systems. The second term, denoted as div (or "divergence"), measures the spread of prediction of the base-systems relative to the score of combined mechanism.

Based on the definition of accuracy in (4.38), the accuracy of y^k_{COM} (after taking into account of the linear transformation A_i and B_i for all i) as defined by the fusion rule (4.8) is:

$$\begin{aligned}
\text{acc}^k &= \frac{1}{N}\sum_i E[\frac{y_i^k - B_i}{A_i} - \mu^k_{COM}]^2 \\
&= \frac{1}{N}\sum_i E\left[\frac{y_i^k - B_i}{A_i} - \frac{1}{N}\sum_j \frac{\mu_j^k - B_j}{A_j}\right]^2 \\
&= \frac{1}{N}\sum_i E\left[\frac{1}{N}\sum_j \frac{N y_i^k - \mu_j^k}{A_j}\right]^2 \text{(change index from } j \text{ to } i) \\
&= \frac{1}{N}\sum_i \left(\frac{1}{N}\frac{E[\eta_i^k \eta_i^k]}{(A_i)^2}\right) = V^k_{diag}.
\end{aligned} \tag{4.39}$$

From (4.38) and (4.26), it is obvious that divergence is simply:

$$\text{div}^k = -V^k_{ndiag}. \tag{4.40}$$

The negative sign in this term shows that divergence is indeed negatively proportional to the covariance component. Hence, conclusions drawn in Section C.3 also apply here: divergence (negative covariance) is not a sufficient metric for measuring classification error diversity. This explains why a number of heuristics to define classification error diversity have been proposed in the literature [135], all based on zero-one loss function where a threshold has already been applied. What we really want to do is in fact to measure the diversity *without fixing the threshold* in advance. For a specific case in biometric authentication, this can be done via F-ratio as proposed in Section 4.3.4. By doing so, one assumes that the client and impostor scores can be modeled by Gaussian distributions, and that the prior class distributions and cost of two types of errors are equal.

4.6.4 Relation To Bias-Variance-Covariance Decomposition

Ueda and Nakano [144] presented the bias-variance-covariance decomposition while Brown [13] provided the link between this concept and the ambiguity decomposition. However, both discussions were limited to the context of regression, as clearly pointed out by Brown [13, Sec. 3.1.2]. So far, we have not discussed the effect of mismatch between the training and the test conditions. We will show that the concept of bias introduced in [144, 13] is useful but unfortunately not relevant for the classification problem. Section 4.6.5 then a noise mismatch model that is relevant of classification in terms of HTER.

Suppose that the noise model in (4.1) can only be calculated from a training set. During testing, the noise model deviates from the one observed during training, i.e., there is a *mismatch* between training and testing. Suppose that the new noise model now is:

$$y_i^{k'} = \mu_i^k + h_i^k + \eta_i^k, \tag{4.41}$$

where h_i^k is a bias. By using the new noise model, we also assume that the noise term $\eta_i^k|\forall_i$ *do not change* in both training and test sessions. Note that (4.41) is also true for $y^{k'}_{COM}$ as defined in (4.8). Therefore, it is also valid to write:

$$\begin{aligned}
y^{k'}_{COM} &= \frac{1}{N}\sum_i \frac{\left(\mu_i^k + h_i^k + \eta_i^k\right) - B_i}{A_i}, \\
&= \frac{1}{N}\sum_i \frac{y_i^k - B_i}{A_i} + \frac{1}{N}\sum_i \frac{h_i^k}{A_i} + \frac{1}{N}\sum_i \frac{\eta_i^k}{A_i}, \\
&\equiv \underbrace{\mu^k_{COM} + h^k_{COM}} + \eta^k_{COM}, \qquad (4.42) \\
&= \mu^{k'}_{COM} + \eta^k_{COM}, \qquad (4.43)
\end{aligned}$$

whose mean is the underbraced terms resulting in $\mu^{k'}_{COM}$. Using (4.43), the class-dependent Mean-Squared Error (MSE) due to this mismatch can be calculated as follows:

$$\begin{aligned} E\left[\left(y^{k'}_{COM} - \mu^{k'}_{COM}]\right)^2\right] &= E\left[\left(y^{k'}_{COM} - \mu^{k}_{COM} - h^{k}_{COM}\right)^2\right] \\ &= \left(h^{k}_{COM}\right)^2 + E\left[\left(y^{k}_{COM} - \mu^{k}_{COM}\right)^2\right] \\ &= \underbrace{\left(h^{k}_{COM}\right)^2} + \underbrace{V^{k}_{diag} + V^{k}_{ndiag}}. \end{aligned} \tag{4.44}$$

where the first underbraced term is bias^2 and the second underbraced term is variance of the fused score (found in the training set). As defined in (4.26), the second term can be further decomposed into V^k_{diag} (i.e., the average variance of all systems when used separately) and V^k_{ndiag} (i.e., the spread of prediction; negative divergence as found in (4.40)). (4.44) is the so-called *bias-variance-covariance* decomposition. Note that this is a decomposition of MSE. In the context of classification, MSE is not relevant; HTER is and it is defined in (2.7) with the optimal *a posteriori* threshold Δ_{apost} (hence giving EER on the test set). The variance of $y^{k'}_{COM}$ is:

$$\begin{aligned} (\sigma^{k'}_{COM})^2 &\equiv E\left[\left(y^{k'}_{COM} - E\left[y^{k'}_{COM}\right]\right)^2\right] \\ &= E\left[\left((y^{k}_{COM} + h^{k}_{COM}) - (\mu^{k}_{COM} + h^{k}_{COM})\right)^2\right] \\ &= E\left[\left(Y^{k}_{COM} - \mu^{k}_{COM}\right)^2\right] \\ &= (\sigma^{k}_{COM})^2. \end{aligned} \tag{4.45}$$

Under the new noise model, it is interesting to note that the class-conditional variance of the fused score is indeed not affected by the bias, whereas the MSE is. However, Section 4.6.5 will show that the presence of bias can adversely affect the classification error measured by HTER.

4.6.5 A Parametric Score Mismatch Model

Note that a noise mismatch model has been proposed in [76, Chap. 10], but for fusion classifiers in probability using the combination approach discussed in Section 3.4.2. Here, we propose a parametric noise model that is adequate for the fusion classifiers in the LLR space.

When one knows the amount of mismatch (i.e, one has access to the test data), the *a posteriori* F-ratio is:

$$\begin{aligned} \text{F-ratio}_{COM,apost} &= \frac{\mu^{C'}_{COM} - \mu^{I'}_{COM}}{\sigma^{C'}_{COM} + \sigma^{I'}_{COM}} \\ &= \frac{\left((\mu^{C}_{COM} + h^{C}_{COM}) - (\mu^{I,\prime}_{COM} + h^{I}_{COM})\right)}{\sigma^{C}_{COM} + \sigma^{I}_{COM}}. \end{aligned} \tag{4.46}$$

Note that at the *a posteriori* F-ratio and its corresponding *a posteriori* EER, their corresponding threshold is at:

$$\Delta_{apost} = \frac{\left((\mu^{I}_{COM} + h^{I}_{COM})\sigma^{C}_{COM} + (\mu^{C}_{COM} + h^{C}_{COM})\sigma^{I}_{COM}\right)}{\sigma^{I}_{COM} + \sigma^{C}_{COM}}. \tag{4.47}$$

The corresponding HTER will be:

$$\begin{aligned} \text{HTER}_{COM,apost} &\equiv \text{EER}_{COM,apost} \\ &= \frac{1}{2} - \frac{1}{2}\text{erf}\left(\frac{\text{F-ratio}_{COM,apost}}{\sqrt{2}}\right). \end{aligned} \tag{4.48}$$

When one does not know the amount of mismatch, the *a priori* threshold that will be used is the one that is estimated from the training set, i.e.,

$$\Delta_{apri} = \frac{\mu^I_{COM}\sigma^C_{COM} + \mu^C_{COM}\sigma^I_{COM}}{\sigma^I_{COM} + \sigma^C_{COM}}. \tag{4.49}$$

This threshold is then applied to the mismatched test set. As a result, the *a priori* HTER (on the test set) will be:

$$\text{HTER}_{COM,apri} \equiv \text{HTER}_{COM}(\Delta_{apri}) \tag{4.50}$$

where, in a general context, for any given Δ, the corresponding HTER is:

$$\text{HTER}_{COM}(\Delta) = \frac{1}{2}\left(\text{FAR}_{COM}(\Delta) + \text{FRR}_{COM}(\Delta)\right), \tag{4.51}$$

where

$$\text{FAR}_{COM}(\Delta) = \frac{1}{2} - \frac{1}{2}\text{erf}\left(\frac{\Delta - \mu^I_{COM} - h^I_{COM}}{\sigma^I_{COM}\sqrt{2}}\right), \tag{4.52}$$

and

$$\text{FRR}_{COM}(\Delta) = \frac{1}{2} + \frac{1}{2}\text{erf}\left(\frac{\Delta - \mu^C_{COM} - h^C_{COM}}{\sigma^C_{COM}\sqrt{2}}\right). \tag{4.53}$$

It is possible to show that

$$\text{HTER}_{COM,apost} < \text{HTER}_{COM,apri}.$$

This can be done by showing that $\text{HTER}_{COM,apost}$ is the *global* minimum, i.e.,

$$\Delta_{apost} = \arg\min_{\Delta} \text{HTER}_{COM}(\Delta). \tag{4.54}$$

Hence any $\Delta \neq \Delta_{apost}$ will *not be optimal*, including Δ_{apri}. In fact this global minimum happens at EER where FAR=FRR because FRR is an increasing function of the threshold and FAR is a decreasing function of the threshold.

In summary, this section shows that the bias-variance-covariance decomposition (of MSE) is not relevant for classification problems. Specifically, in a two-class problem such as biometric authentication, the concepts of *a priori* and *a posteriori* thresholds play an important role in decision-making because these thresholds directly affect the classification error.

Of course in reality, the mismatch is unknown in advance. One possible solution will be to *estimate* the bias h^k_i, for all i. This estimated bias can then be used to calculate a new threshold using (4.47). This bias-correction at the threshold level is practical, for instance, in a multimodal systems where the participating systems exhibit different degree of bias in different application scenarios.

4.7 Extension of F-ratio to Other Fusion Operators

4.7.1 Motivations and Section Organization

The proposed parametric fusion model discussed until now only applies to the weighted sum fusion classifier/operators (with mean as a special case). The first goal of this section is to generalize the proposed fusion model to other fusion operators. Its second goal is to identify conditions under which a fusion operator is superior or more appropriate for a given fusion task. Prior to our study, several theoretical fusion models have already been proposed, e.g., [66] on the sum and product rules (with $\max$ and $\min$ as special cases), [142] on OS combiners, [73] on several fusion classifiers and the most recent study [123] which takes into consideration correlation and unbalanced system performance. All these studies share the common characteristic that they consider system outputs in probability. Our proposed model is somewhat different because we consider system outputs in the LLR space, where scores can be summarized using first- and second-order statistics. This advantage, not shared by the previous studies [66, 142, 73, 123], allows us to *compare* the performance of different fusion operators using the mentioned statistics.

Table 4.1: Summary of theoretical EER based on the assumption that class-independent scores are normally distributed.

Fusion methods	EER	where
average baseline†[1]	$\text{EER}_{AV} = \text{eer}\left(\frac{\mu^C_{AV}-\mu^I_{AV}}{\sigma^C_{AV}+\sigma^I_{AV}}\right)$	$\mu^k_{AV} = \frac{1}{N}\sum_i \mu^k_i$ $\left(\sigma^k_{AV}\right)^2 = \frac{1}{N}\sum_i \left(\sigma^k_i\right)^2$
single-best classifier	$\text{EER}_{best} = \text{eer}\left(\max_i\left(\frac{\mu^C_i-\mu^I_i}{\sigma^C_i+\sigma^I_i}\right)\right)$	–
mean rule	$\text{EER}_{mean} = \text{eer}\left(\frac{\mu^C_{mean}-\mu^I_{mean}}{\sigma^C_{mean}+\sigma^I_{mean}}\right)$	$\mu^k_{mean} = \frac{1}{N}\sum_i \mu^k_i$ $\left(\sigma^k_{mean}\right)^2 = \frac{1}{N^2}\sum_{i,j} \mathbf{\Sigma}^k_{i,j}$
weighted sum†[2]	$\text{EER}_{wsum} = \text{eer}\left(\frac{\mu^C_{wsum}-\mu^I_{wsum}}{\sigma^C_{wsum}+\sigma^I_{wsum}}\right)$	$\mu^k_{wsum} = \sum_i \omega_i \mu^k_i$ $\left(\sigma^k_{wsum}\right)^2 = \sum_{i,j} \omega_i\omega_j \mathbf{\Sigma}^k_{i,j}$
OS combiners†[3]	$\text{EER}_{OS} = \text{eer}\left(\frac{\mu^C_{OS}-\mu^I_{OS}}{\sigma^C_{OS}+\sigma^I_{OS}}\right)$	$\mu^k_{OS} = \mu^k + \gamma_1\sigma^k$ $\left(\sigma^k_{OS}\right)^2 = \gamma_2\left(\sigma^k\right)^2$

†[1]: This is not a classifier but the average performance of baselines when used independently of each other. By its defintion, scores are assumed independent as classifiers function independently of each other. †[2]: the weighted product (respectively product) takes the same form as weighted sum (respectively sum), except that log-normal distribution is assumed instead. †[3]: OS classifiers assume that scores *across classifiers* are i.i.d. The reduction factor γ is listed in Table 4.2. The mean and weighted sum classifiers *do not* assume that scores are i.i.d.

This Section is divided into four sub-sections. Section 4.7.2 lists the Bayes error of some commonly used fusion operators in the literature. Section 4.7.3 examines the Order Statistics (OS) operators in details, e.g., $\min$, $\max$ and median. Section 4.7.4 compares the performance of different fusion operators with respect to two factors: correlation among system outputs and unbalanced system performance. Lacking the necessary data, the comparison is performed using simulated data according to the class-conditional score Gaussian assumption. The experimental setting in Section 4.7.4 does not allow us to distinguish between $\min$ and $\max$ fusion operators. Section 4.7.5 then explicitly introduces another experimental setting that highlights the differences. This leads to a rarely considered result in previous studies, e.g., [66, 142, 73, 123].

4.7.2 Theoretical EER of Commonly Used Fusion Classifiers

There are more than one ways to extend the proposed parametric fusion model to other fusion operators. One can begin with the Chernoff bound formulation as appeared in (4.6). Note that it is an upper bound of the Bayes error or EER as appeared in (4.14), which is an exact solution. With the Chernoff bound formulation, one can replace $p(\mathbf{Y}|k)$ in (4.6) by $p(y_{COM}|k$, the conditional distribution of the combined score. Because any fusion operator $f_{COM} : \mathbb{R}^N \to \mathbb{R}$ maps from N dimensions to a single dimension, one no longer needs the upper bound parameterized by β so that a direct optimization of WER is possible, i.e.,:

$$\begin{aligned} \text{WER} &= \int \min\left[\alpha p(y_{COM}|I), (1-\alpha)p(y_{COM}|C)\right] dy_{COM}, && (4.55)\\ &= \alpha\text{FAR} + (1-\alpha)\text{FRR}, && (4.56)\end{aligned}$$

recalling that FAR and FRR are integrals of $p(y_{COM}|I)$ and $p(y_{COM}|C)$, respectively. When FAR and FRR are assumed to be integrals of Gaussian and $\alpha = 0.5$, the minimal WER value is EER. As a result, we see that while the Chernoff bound provides an upper bound to the Bayes error, EER provides an *exact* solution. This section will develop the EER of several other combination operators discussed in Section 3.4.2.

Thanks to F-ratio, the analysis of EER can be summarized by the following four parameters: $\{\mu^k, \sigma^k|\forall_{k=\{C,I\}}\}$. The *average baseline* performance of classifiers, considering that each of them works independently of the other, is shown in the first row of Table 4.1. The (class-dependent) average variance,

σ^k_{AV}, is defined as the average over all the variances of classifier. This is in fact not a fusion classifier but the *average performance* of classifiers measured in EER. The single-best classifier in the second row chooses the baseline classifier that maximizes the F-ratio. This is the same as choosing the one with minimum EER because F-ratio is inversely proportional to EER, as implied by the left part of (4.14).

The derivation of EER of weighted sum (as well as mean) fusion can be found in Section D.2.

For the product operator, it is necessary to bound y to be within the range $[0, 1]$, otherwise multiplication is not applicable. Consider the following case: two instances of classifier score can take on any real value. The decision function (3.3) is used with optimal threshold being zero. With an impostor access, both classifier scores will be negative if correctly classified. Their product, on the other hand, will be positive. This is clearly undesirable.

The weighted product (and hence product) at first seems slightly cumbersome to obtain. However, one can apply the following logarithmic transform instead: $\log(y^k_{wprod}) = \sum_i w_i \log(y^k_i)$, for any y^k_i sampled from $p(y^k_i)$. This turns out to take the same form as weighted sum. Assuming that y^k_i is log-normally distributed, we can proceed the analysis in a similar way as the weighted sum case (and hence the mean rule).

4.7.3 On Order Statistic Combiners

To implement fixed rule *order statistics* (OS) such as the maximum, minimum and median combiners, scores must be comparable. This can be done by using F-norm. Unlike the previous section, we further assume here that the scores are i.i.d. (independently and identically distributed). Hence, $p(y_i|k) = p(y_j|k)$ for any $i, j \in [1, \ldots, N]$. Although this assumption seems too constraining, it is at least applicable to fusion with multiple samples which are indeed identically distributed but not independently sampled.

All OS combiners will be collectively studied here. The subscript OS can be replaced by min, max and median. Supposing that $y^k_i \sim Y^k_i$ is an instance of i-th response knowing that the associated access claim belongs to class k. y^i has the following model:

$$y^k_i = \mu^k_i + \omega^k_i,$$

where μ^k_i is a deterministic component and ω^k_i is a noise component. Note that in the previous section ω^k_i is assumed to be normally distributed with zero mean. The fused scores by OS can be written as:

$$y^k_{OS} = \mathrm{OS}(y^k_i) = \mu^k + \mathrm{OS}(\omega^k_i),$$

where i denotes the i-th sample (and not the i-th classifier output as done in the previous section). Note that μ^k is constant across i and it is *not affected* by the OS combiner. The expectation of y^k_{OS} as well as its variance are shown in the last row of Table 4.1, where γ_2 is a reduction factor and γ_1 is a shift factor, such that $\gamma_2(\sigma^k)^2$ is the variance of $\mathrm{OS}(\omega^k_i)$ and $\gamma_1\sigma^k$ is the expected value of $\mathrm{OS}(\omega^k_i)$. Both γ's can be found in tabulated form for various noise distributions [3]. A similar line of analysis can be found in [143] except that class-unconditional noise is assumed, i.e., $\sigma^C = \sigma^I$. The reduction factors of combining the first five samples, assuming Gaussian distribution, are shown in Table 4.2. The smaller γ_2 is, the smaller the associated EER. The fourth column of Table 4.2 shows the reduction factor due to mean (as compared to the second and third columns). It can be seen that mean is overall superior.

4.7.4 Experimental Simulations

Lacking of the necessary data, we performed a set of simulations similar to those mentioned in Section 4.6.2 and following exactly the same assumptions: (i) the two base-systems have the same numerator of F-ratio and that (ii) for each base-system, the variance and covariance of client and impostor distributions are proportional. By doing so, the experimental task can be described by the matrix $(\frac{\sigma_2}{\sigma_1}, \rho)$ (variance ratio and correlation) and the corresponding outcome by β_{min}. The only difference from Section 4.6.2 is that we used min and max and $\prod$ as fusion operators. The results are shown in Figure 4.8.

Comparing Figure 4.7 with Figure 4.8, it can be observed that the mean operator is better than min or max. For all cases except the product operator, low correlation and low variance ratio (unbalanced system performance) are important to guarantee a positive gain. The product rule only has performance as good

Table 4.2: Reduction factor of order statistics.

N	γ_2 values			γ_1 values	
	OS combiners		mean	OS combiners	
	min, max,	median	$(\frac{1}{N})$	min	max
1	1.000	1.000	1.000	0.00	0.00
2	0.682	0.682	0.500	-0.56	0.56
3	0.560	0.449	0.333	-0.85	0.85
4	0.492	0.361	0.250	-1.03	1.03
5	0.448	0.287	0.200	-1.16	1.16

Reduction factor γ_2 of variance (2 for the second moment) with respect to the standard normal distribution due to fusion with $\min$, $\max$ (the second column) and median (third column) OS combiners for the first five samples according to [3]. The fourth column is the *maximum* reduction factor due to mean (at zero correlation), with minimum reduction factor being 1 (at perfect correlation). The fifth and sixth columns show the shift factor γ_1 (for the first moment) as a result of applying $\min$ and $\max$ for the first five samples. These values also exist in tabulated forms but here they are obtained by simulation. For median, γ_1 is relatively small (in the order of 10^{-4}) beyond 2 samples and hence not shown here. It approaches zero as N is large.

as the single-best classifier at variance ratio=1 while does not match the rest of the fusion classifiers. Its performance does not evolve with correlation. One plausible explanation of such suboptimal performance comes from [66], which states that the the product rule is more sensitive to error as compared to the sum (or mean) rule.

4.7.5 Conditions Favoring A Fusion Operator

In this Section, we would like to investigate conditions which favor a given fusion operator, e.g., $\min$, $\max$, etc. Due to the assumptions $\sigma^C = \sigma^I$ and $\mu^C = \mu^I$, the simulations in Figure 4.8 could not distinguish between the two operators. The F-ratio of OS combiners can be written as:

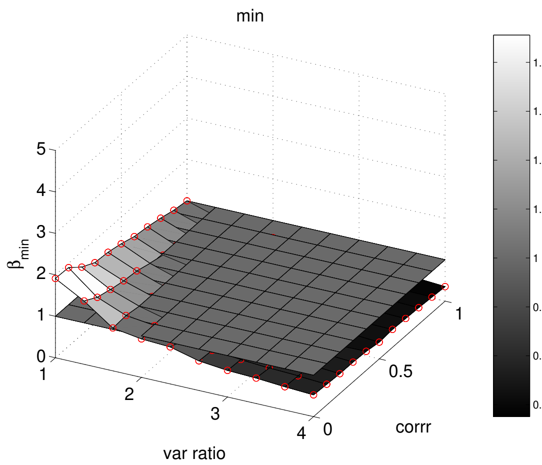
$$\text{F-ratio}_{OS} = \frac{\mu^C_{OS} - \mu^I_{OS}}{\sigma^C_{OS} + \sigma^I_{OS}} = \frac{\mu^C - \mu^I + \underbrace{\gamma_1(\sigma^C - \sigma^I)}}{\sqrt{\gamma_2}(\sigma^C + \sigma^I)}, \tag{4.57}$$

for both $OS \in \{\min, \max\}$. The underbraced term is critical in that it is different for $\min$ and $\max$ whereas the rest of the terms remain the same for both operators. In order that this quantity is positive (to ensure gain $\beta_{min} > 1$), there are two possibilities:

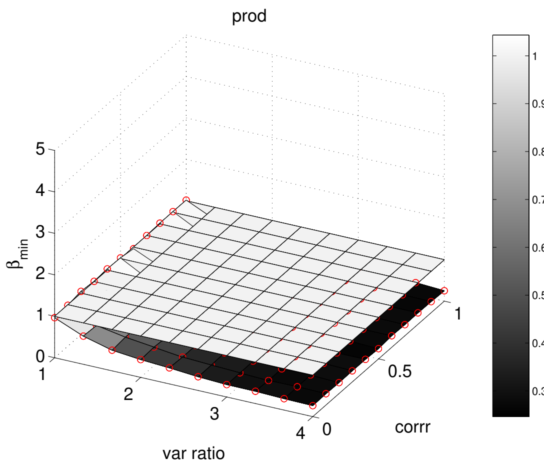
- $\gamma_1 > 0$ and $\sigma^C > \sigma^I$ – in which case $\max$ is better.
- $\gamma_1 < 0$ and $\sigma^C < \sigma^I$ – in which case $\min$ is better.

As can be observed, the magnitude of σ^k for $k = \{C, I\}$ determines largely which operator is more suitable. We performed a simulation using the experimental settings as before but this time, we varied the variance ratio $\frac{\sigma^C}{\sigma^I}$ whereas $\rho = 0$. The results are shown in Figure 4.9. As can be observed, when $\sigma^C = \sigma^I$, so that the ratio is 1, $\min$ and $\max$ are equivalent. However, as dictated by the constraint of (4.57), $\max$ is better when $\sigma^C > \sigma^I$ and vice versa for $\min$. It is interesting to observe that when $\frac{\sigma^C}{\sigma^I} > 1.6$, $\max$ is even better than mean or weighted sum. This shows that contrary to what one may expect, in some situations, $\max$ may be better than weighted sum.

Finally, we also carried out some empirical evaluations to verify the findings so far using the XM2VTS score-level fusion benchmark datasets with 32 two-system fusion tasks. Each system output is first pre-processing such that $y'_i \equiv f_Z(f_{LLR}(y_i))$ for any system i. The empirical results [107] show that (EPC curves not shown):



(a) min or max



(b) product

Figure 4.8: Comparison between $\min$ or $\max$ and the product operator using synthetic data. Performance gain β_{min}, (the Z-axis) across different variance ratios (of two systems) from 1 to 4 (the X-axis) and different correlation values from 0 to 1 (the Y-axis), as a result of fusing synthetic scores of two system outputs using (a) $\min$ or $\max$ (both produce identical results) and (b) product fusion operators.

- $\max_i y'_i$ is better than $\min_i y'_i$. An analysis shows that $Var[y'_i|C] > Var[y'_i|I]$ is true for most system outputs y'_i, for any i in this data set.
- Weighted sum fusion operator, $\mathbf{w}'\mathbf{y}$ (whose weights are optimized by minimizing EER on the development set), is better than $\min$, $\max$ or mean rule. This indicates that trainable fusion classifiers are optimal for the 32 fusion tasks.

4.8 Summary of Contributions

While estimating Bayes error is a classical problem in machine learning, e.g., the Chernoff bound, we demonstrate that the Bayes error in our fusion setting, which is equivalent to EER in our case, can be estimated *exactly* (hence *tighter* estimate). The underlying assumption is that the class-conditional scores are normally distributed. Even though this assumption seems to be restrictive, by using more than a thousand biometric experiments, we show that the estimated EER is acceptable in practice. Thanks to the fusion model, we can:

1. **Justify why fusion is better than the average performance of its sub-systems empirically and theoretically:** Although this subject has been investigated elsewhere, e.g., [57, 67, 68, 133], our justification is unique in the sense that it is *directly* related to the reduction of classification error in terms of EER. The empirical justification shown in Section 4.2 was summarized from our paper [97] whereas the theoretical justification shown in Section 4.4 was extended from our paper [103].
2. **Predict fusion performance:** To the best of our knowledge, prior studies on classifier combination, e.g., [67, 68, 123], have not dealt with the subject of performance prediction since they deal with system outputs in probability. However, by working on the alternative LLR space, we show that performance prediction is not only feasible, but also that the predicted performance is sufficiently accurate to be used in classifier selection. This study that was discussed in Section 4.5 was summarized from our work [102].
3. **Understand the effects of unbalanced classifier performance and correlation:** These two factors have been studied in [123] by considering weighted sum fusion in probability. Our parametric

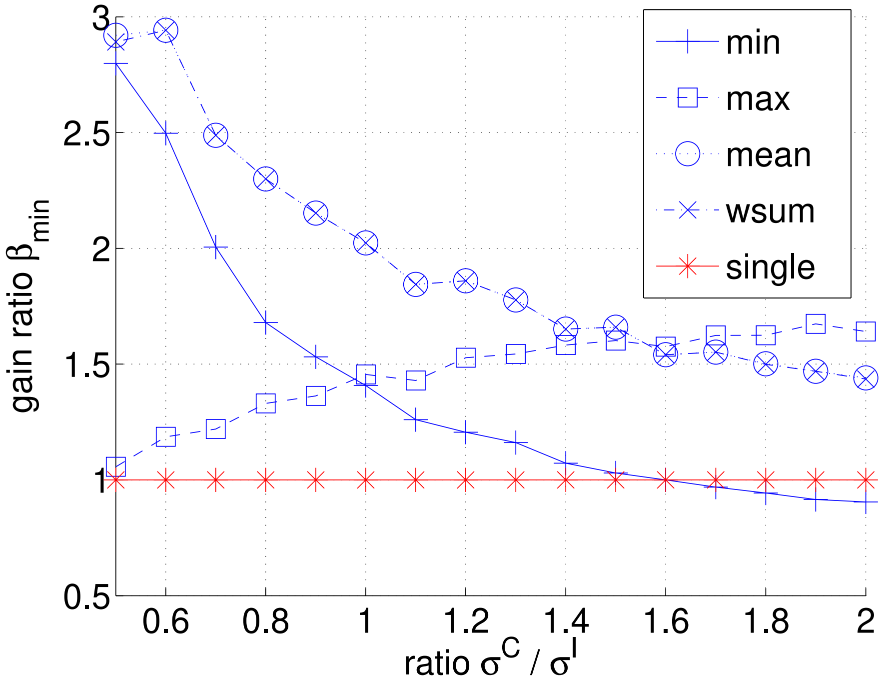


Figure 4.9: Performance gain β_{min} (with respect to the best underlying system) versus conditional variance ratio $\frac{\sigma^C}{\sigma^I}$ of different fusion operators.

approach models these two factors in the LLR space. Although both studies lead to the *same* conclusion, the approach based on LLR is undoubtedly *much simpler* compared to [123] thanks to the ability to summarize data in the LLR space (using the first- and second-order moments). For instance, the pairwise correlation can naturally be described by the covariance matrix in LLR but this is not obvious in probability. Therefore, our proposed model provides an *alternative understanding* of fusion with respect to the two effects mentioned. Our study as described in Section 4.6) was taken from our published work [109].

4. **Study the adverse effect of bias on fusion:** The study of score-level mismatch between training and test sets was examined in [76, Chap. 10] for the case of fusion using simple operator in probability. It further makes the system output independence assumption. Different from [76, Chap. 10], our study uses weighted sum as a fusion classifier and considers the system output dependency *explicitly* in the LLR space. Due to using LLR, our approach is more advantageous because it allows one to correct the bias while such remedial procedure is non-obvious with probability. This study was also taken from our published paper [109].

5. **Identify conditions which favor a particular fusion operator:** Thanks to the parametric model, these conditions, described using class conditional Gaussian parameters, can be identified. By using many experimental simulations, we found two observations interesting and somewhat surprising. Firstly, contrary to popular beliefs, there exists conditions under which $\max$ and $\min$ operators are better than weighted sum (or mean as a special case). In practice, however, these conditions occur rarely. Secondly, there exists conditions under which $\min$ is better than $\max$, and vice-versa, in the context of fusion. Prior to our study, e.g., [67, 68, 123, 76], these conditions were not well understood. This study as described in Section 4.7 has not been published yet.

In brief, we have shown that working in LLR is more advantageous than in probability since we can summarize and analyze the *same* fusion problem (in both cases) more easily thanks to the Gaussian distribution.

Part II

User-Specific Processing

Chapter 5

A Survey on User-Specific Processing

5.1 Introduction

While Part I of this thesis is about user-independent fusion, Part II is about user-specific fusion. In theory, extending the user-independent parametric fusion model to a user-specific one is straightforward, e.g., replacing the statistics $\boldsymbol{\mu}^k$ and $\boldsymbol{\Sigma}^k$ by user-specific statistics $\boldsymbol{\mu}_j^k$ and $\boldsymbol{\Sigma}_j^k$ for a given user index j. In practice, however, due to limited amount of user-specific data, the reliability of user-specific statistics are greatly reduced. We will survey in this chapter all techniques that rely on using data specific to a user. We call this family of techniques *user-specific processing*. Examples of user-specific processing are user-specific feature extraction, user-specific model/template, user-specific fusion classifier, user-specific score normalization and user-specific threshold.

There are at least two motivations to apply user-specific processing in biometric authentication. Firstly, it has been observed that in a database acquired in similar conditions, a fraction of users are more difficult to recognize than the rest [33]. It is, in fact, possible to rank users in a database according to an index of ease of recognition (Section 7.4). Secondly, it is common knowledge that human beings recognize people by their salient traits. These traits are best seen in human caricature characters where remarkable traits of a person are exaggerated.

User-specific processing is a challenging problem because very often, extremely few samples are available per user. This is even more so for newly enrolled users. For instance, it has been shown in [40] that at least six genuine samples are needed before its proposed user-specific procedure can outperform the baseline system. Ten samples were reported in [139] and five in [50]. Such a large number of samples can be inconvenient if one considers that conventional non-automatic biometric applications use only one sample, e.g., a single mug-shot photo for traveling documents. Therefore, an important challenge to overcome in user-specific processing is to reduce the required number of genuine training samples, i.e., learning with small sample size. This is a non-trivial machine-learning problem. Chapters 6 and 7 provide two alternatives of applying user-specific processing that can work with a single genuine training sample.

To the best of our knowledge, one of the earliest applications of user-specific processing is user-specific score normalization [48]. Since then, such family of methods is extended to user-specific threshold, e.g., [92], and user-specific fusion, e.g., [61, 139, 40]. These studies show that exploiting user-specific information can effectively improve the system accuracy. This chapter provides a survey as well as a thorough analysis of this subject. To the best of our knowledge, despite its importance, this is the first survey written on the subject.

Chapter Organization

This chapter is organized as follows: Section 5.2 introduces the terminology in user-specific processing and motivates user-specific decision making. Section 5.3 will give an overview of user-specific processing techniques from an architectural perspective. Sections 5.4–5.6 present user-specific fusion, user-specific score normalization and user-specific threshold. Section 5.7 analyzes the relationship between user-specific normalization and threshold. Finally, Section 5.8 concludes the chapter.

5.2 Terminology and Notations

5.2.1 Terminology Referring to User-specific Information

Due to the evolving nature of this field, several terms have been introduced by different authors, e.g. [43, 139]. To avoid confusion, in this thesis, we will adopt the following terms:

- **User-specific**/client-dependent/local: (adjective) on a per client basis.
- **User-independent**/client-independent/global/common: (adjective) indifferent to each client.
- **User-adapted**: (adjective) that makes use of *both* user-specific and user-independent statistics..
- **Client-centric**/target-centric: (adjective) that makes use of user-specific client accesses only.
- **Impostor-centric**: (adjective) that makes use of user-specific impostor accesses only.
- **Client-impostor centric**/target-impostor centric: (adjective) that makes use of both user-specific client and impostor accesses.

Note that the bold terms are used in this thesis whereas the rest of the terms separated by "/" are synonyms.

5.2.2 Towards User-Specific Decision

Let $j \in \{1, \dots, J\}$ be the identity being claimed when making an access request and there are J users. The user-specific decision will necessary take the index j into consideration. In contrast to user-independent decision based on Log-Posterior Ratio (LPR) as defined in (3.1), the user-specific decision function, which considers the identity claim j, using LPR, can be written as:

$$\begin{aligned}
\text{LPR}_j &\equiv \log\left(\frac{P(C,j|\text{person})}{P(I,j|\text{person})}\right) \\
&= \underbrace{\log\frac{P(\text{person}|C,j)}{P(\text{person}|I,j)}} - \underbrace{\log\frac{P(I,j)}{P(C,j)}}
\end{aligned} \tag{5.1}$$

Instead of considering at the "person" level (the composite of digitized biometric signals), the LPR test is also valid at the feature level (by replacing "person" with the feature vector $\mathbf{x}$) or at the system level (by replacing "person" with a vector of system outputs $\mathbf{y} = [y_1, \dots, y_N]'$ with N elements). By considering N systems, our framework generalizes to a single system output where $N = 1$.

To illustrate the usefulness of user-specific decision, we will focus on LPR_j at the system output level. Therefore, (5.1) can be written as:

$$\text{LPR}_j = \underbrace{\log\frac{P(\mathbf{y}|C,j)}{P(\mathbf{y}|I,j)}} - \underbrace{\log\frac{P(I,j)}{P(C,j)}} \equiv \Psi_j(\mathbf{y}) - \Delta_j, \tag{5.2}$$

where one can recognize that $\Psi_j : \mathbb{R}^N \to \mathbb{R}$ is a user-specific fusion function and Δ_j is its corresponding user-specific threshold. When $N = 1$, the function $\Psi_j : \mathbb{R} \to \mathbb{R}$ is called a user-specific *score normalization*. Following a similar discussion as in Section 3.2.2, the decision function of (5.2) can be written as:

$$\text{decision}(\mathbf{y}) = \begin{cases} accept & \text{if } \Psi_j(\mathbf{y}) > \Delta_j \\ reject & \text{otherwise.} \end{cases} \tag{5.3}$$

This decision function is impractical for two reasons. Firstly, the user-specific threshold Δ_j is difficult to estimate due to lack of genuine samples associated to identity j. Secondly, the user-specific fusion function (or score normalization for $N = 1$) is also difficult to estimate for the same reason. Despite the difficulties, this form of solution was examined in [139], where as many as ten samples were used – demonstrating the drawback of this approach.

In order to be robust to few user-specific training samples, (5.3) can be approximated by the following ways:

1. **Using only user-specific function:** This results in the following decision function:

$$\text{decision}(\mathbf{y}) = \begin{cases} accept & \text{if } \Psi_j(\mathbf{y}) > \Delta \\ reject & \text{otherwise,} \end{cases} \tag{5.4}$$

   In this case, the threshold Δ is common to all users. The function Ψ_j in (5.4) is a *user-specific fusion* for $N > 1$ and is called a *user-specific score normalization procedure* for $N = 1$.

2. **Using only user-specific threshold:**

$$\text{decision}(y) = \begin{cases} accept & \text{if } y > \Psi'_j(\Delta) \\ reject & \text{otherwise,} \end{cases} \tag{5.5}$$

   where $\Psi'_j : \mathbb{R} \to \mathbb{R}$ is a *user-specific threshold* (Ψ_j with $'$). In this case, the fusion function is common to all users. This form was examined by [43] for instance.

3. **Using neither one:** In this case, no user-specific information is used. This results in the user-independent decision function shown in (3.3) where $\Psi(\Delta) = \Delta$. This is the *de facto* approach.

Note that (5.4) and (5.5) are closely related. Their relationship will be shown in Section 5.7. This section is original because to the best of our knowledge, such relationship has not been shown in the literature. The dual relationship is useful because it indicates that it is always possible to find an equivalence of user-specific threshold from user-specific score normalization but not necessarily the other way round (depending on whether the common threshold Δ is considered or not). In other words, user-specific score normalization generalizes over user-specific threshold. For this reason, we choose to focus on user-specific score normalization.

Our contributions to be discussed in Chapters 6 and 7 will be based on (5.4) in the context of fusion (for $N > 1$) and that of score normalization (for $N = 1$).

5.3 Levels of User-Specific Processing

User-specific processing can be applied at the following three architectural levels:

1. **Feature level – User-specific feature set**. At this level, different feature representations are used for different user or group of users. For instance, for users whose fingerprint minutiae cannot be extracted reliably, the textual information may be more useful. In [22], it was shown that the performance of a speaker verification task can be enhanced by using a subset of features for each user. These features are chosen using a feature selection technique.

2. **Model level – User-specific model**. This is a standard approach whereby a biometric authentication system builds a model on a *per user* basis. For instance, it is common to train an MLP classifier to separate the face of a user from the rest of the users [81]. This strategy is called the one-against-all classification strategy. The-state-of-the-art approach in speaker verification, which is based on a user-adapted model [122] from a general speaker independent model, is also based on the same strategy. Recent techniques in face verification also follow the same trend, e.g., [16] using local features (which are classified with a user-adapted model) and [151] using user-specific Fisher's projection.

3. **Score level** – which can be further divided into:

   - **User-specific score normalization.** The most representative example is called Z-norm and first proposed by [48], which relies on user-specific impostor scores to carry out the normalization. In [126], a similar version of Z-Norm but using only user-specific genuine scores was reported. However, this technique requires much more user-specific genuine accesses. The authors' experiments were based on 5 accesses per user. Since the first work by [48], the form of normalization has not been changed much although the context of application is extended beyond that of mitigating user-induced score variations, such as T-Norm [4], (aiming at extenuating the mismatch during test), H-norm [53], (aiming at extenuating the mismatch due to the use of different handsets) or and D-Norm [6] (aiming at reducing model-induced variations and is specific to GMMs). Other normalization techniques employing both user-specific client and impostor information (i.e., client-impostor centric) include EER-norm [43] and the proposed F-norm in Chapter 7.

- **User-specific fusion**. This technique was proposed by [61] using a linear weighing scheme to weigh the outputs of several multimodal systems while a non-linear version, achieved via Multi-Layer Perceptron (MLP) was reported in [71]. In [40], a Support Vector Machine (SVM) classifier was used to construct a user-specific fusion function while in [41], a Bayesian classifier was used.
- **User-specific threshold**. This class of techniques is commonly applied to speaker verification tasks for instance [48, 126, 75, 93, 18].

The literature cited here is certainly not exhaustive but it represents the state-of-the-art in user-specific processing.

Often, the score-level techniques are used together with the feature-level techniques. For instance, the state-of-the-art speaker verification technique based on adapted Gaussian Mixture Model [122, 4] uses both user-specific model and user-specific score normalization. The same adapted GMM architecture was employed successfully to signature verification [43] and to face verification [130]. In [79], another possible combination was proposed, i.e., between user-specific score normalization (based on Z-norm) and user-independent fusion.

A recent study [139] proposed a new paradigm consisting of two dichotomies: user-specific/user-independent fusion (called “local/global learning” by the author) and user-specific/user-independent threshold (called “local/global decision”). These two dichotomies thus give four categories of methods to incorporate user-specific information, at the score level. Rather than just looking at these dichotomies, one should investigate the possibility of applying user-specific strategies at *all* possible levels listed here.

A detailed discussion on user-specific fusion can be found in Section 5.4, score normalization in Section 5.5 and threshold in Section 5.6. Section 5.7 shows the duality between user-specific score normalization and threshold normalization.

5.4 User-Specific Fusion

The user-specific fusion, Ψ_j, can be constructed based on the following methodologies:

1. A classifier with N inputs but one for each user.
2. A classifier receiving $N + 1$ inputs, i.e., N system outputs to be combined and an identity label.
3. A classifier with N inputs, based on a common model, but its parameters change according to the score statistics of each user.

In the first case, one does not make use of the data of the rest of the users. Therefore, it is inefficient in terms of data usage. In the second case, due to parameter sharing, the use of data is more efficient. In the third case, the possible sets of solution is restricted but with the right model, its generalization performance may be superior over the first two cases. The following five types of user-specific fusion classifiers are found to be relevant:

- **Brute-Force User-Specific Weight Sum:** The first work that exploited user-specific fusion can be attributed to [61], whereby a linear combination of the form $\sum_i w_{i,j} f_{prob}(y_i)$ is used, with the constraint that the weights sum up to one and that the solution with equal weights is preferred. The function f_{prob} converts the output to probability (see Algorithm 1). The weight $w_{i,j}$ for a given user j and system i is tuned directly to minimize the population EER criterion from the data. A potential problem with this technique is that if there are J users and N systems, then there are a total of $N \times J$

weight parameters to solve. Given the high degree of freedom, the solution is unlikely to generalize well.

- **D-prime Based User Specific Weighted Sum:** An improved version of user-specific weighting scheme over [61] was proposed in [137]. The improved scheme uses $w_{i,j} \propto d_j'^{-1}$ where d_j' is user-specific d-prime as defined in (4.17) except that the statistics are derived uniquely from user-specific data (scores). Although this solution is expected to be more robust than the direct weight estimation approach, the user-specific statistics inherent in d-prime can be very unreliable. As a result, such strategy may not generalize well (see Section 7.2).

- **User-Specific SVM:** In [40], a standard SVM was used in a somewhat novel way, i.e., an SVM was constructed using a user-independent set of scores plus a user-specific set of scores. Each of these sets of scores contain both client and impostor classes of scores. This strategy was called "adapted user-dependent fusion" by the author. This is to be distinguished from "user-independent fusion" whereby no user-specific data is used, or "user-dependent fusion" whereby only user-specific (client and impostor) scores are used (while ignoring the existence of user-independent client and impostor scores). The mentioned novelty in the said study is the use of the C parameter in SVM [146]. This parameter rates the *relative influence* of each example. When included in the support vectors (i.e., examples falling in the margin), the relatively high C parameters of these examples can change the decision boundary drastically. In [40], two C values are assigned to two sets of scores, i.e., one for the user-specific scores and one for the user-independent scores. In order for the adapted fusion to be effective a *greater* C value has to be associated to the precious user-specific scores as compared to the C value of the user-independent scores. It was demonstrated empirically that when C was tuned *a posteriori* on the test set (due to lack of available data for tuning the C parameter), the adapted fusion was potentially beneficial as compared to either user-independent or user-specific fusion. Since the additional free parameter C was tuned *a posteriori*, hence providing an additional degree of freedom to fit the data, the experimental results are thus *biased* towards the adapted fusion strategy.

- **User-Specific Gaussian Classifier:** Another similar idea using Bayesian adaptation (instead of using SVM) was reported by the same author in [41], also using the same multimodal database. The architecture employed is similar to the Gaussian Mixture Model (GMM) with Maximum *A Posteriori* (MAP) adaptation, the current state-of-the-art system in speaker verification [122]. However, a single Gaussian component with a diagonal covariance matrix was used[1]. According to our understanding, the justification for using a single Gaussian component is that there are just too few user-specific client scores to adapt (from two to three, depending on bootstrap samples). Similar to the C parameter in SVM, the GMM-MAP algorithm also has a free parameter called a "relevance factor" (to be discussed in Section 6.3.2). This factor is crucial in that it balances the right mix between the user-specific and user-independent information. In other words, both C and relevance factor play the same role in this context. Again, the relevance factor was tuned *a posteriori* and thus inevitably reporting *biased* performance towards the GMM-MAP algorithm. Ideally, any free parameter should be tuned on a separate validation set.

- **Identity-based MLP Fusion:** In [71], an MLP was employed to combine the vector of system outputs $\mathbf{y}$ together with the user-identity index j. Hence, the MLP has $N + 1$ inputs. It was shown that employing the identity claim as an additional feature can improve the performance, albeit insignificantly.

These user-specific classifiers shows that it is important/useful to:

- Use explicitly user-specific score statistics, e.g., [137].
- Share parameters and/or training data among different user-specific classifiers, e.g., [41, 40].

[1] In the context of speaker verification, the use of GMM with a diagonal matrix per Gaussian component is fine since a full covariance matrix does not necessarily provide better performance. On the other hand, in the context of score-level fusion, a single Gaussian component with a full covariance matrix may be more appropriate, if the covariance information is *believed to be* valuable. Unfortunately, no comparative study was reported in this regard.

- Restrict the possible solution space by choosing a model, e.g., [137, 41].

However, none of the above studies possess all these characteristics. We will propose in Chapters 6 and 7 two designs of user-specific classifiers that consider *all* these characteristics which are extremely important in order to reduce the size of user-specific training samples to one. This is a significant savings considering that the studies presented here rely on at least five training samples before the classifier outperforms a *de facto* fusion classifier which does not consider the user label.

5.5 User-Specific Score Normalization

User-specific score normalization can be categorized into two families:

- **Z-norm Based Normalization:** The desired effect is that the distribution of normalized impostor score is aligned. These methods are impostor-centric.
- **EER-norm Based Normalization:** The *sign* of the normalized score is indicative of the class label. These methods are client-impostor centric.

We will introduce another class of methods based on F-norm in Chapter 7. F-norm belongs to a different family because the expected values of the normalized client and impostor scores are *simultaneously* aligned.

The Ideal User-specific Normalization Procedure

If one considers user-specific LLR score as in (5.2) and assumes the class-conditional Gaussian distribution, $\Psi_j(y)$ can be written as:

$$\Psi_j(y) \quad = \quad -\frac{1}{2(\sigma_j^C)^2}\left((y-\mu_j^C)^2\right)+\frac{1}{2(\sigma_j^I)^2}\left((y-\mu_j^I)^2\right)-\log\frac{\sqrt{2\pi(\sigma_j^C)^2}}{\sqrt{2\pi(\sigma_j^I)^2}}, \tag{5.6}$$

where μ_j^k and σ_j^k are the class conditional mean and standard deviations of user j for $k=\{C,I\}$. We call these statistics *user-specific statistics*.

Being an LLR, such a user-specific normalization procedure is optimal (i.e., results in the lowest Bayes error) when

1. The parameters μ_j^k, σ_j^k for $k \in \{C,I\}$ and for all j are estimated correctly.
2. The class-conditional scores can be described by the first and second order statistics.

The second condition can be fulfilled by converting any score type to LLR using Algorithm 2). The first condition is unlikely to be fulfilled in practice because one is always lack of user-specific training data. As a result, in its original form, (5.6) is not a practical solution.

Z-norm Based Normalization

Since μ_j^C and σ_j^C cannot be reliably estimated, the following constraints may be applied to (5.6): $\sigma_j^C=\sigma_j^I$ and $\mu_j^C=y$ (the score itself). As a result, (5.6) becomes:

$$\Psi_j(y) \quad = \quad \frac{(y-\mu_j^I)^2}{2(\sigma_j^I)^2},$$

which is proportional to the square of Z-norm [48] having the form.:

$$y_j^Z \quad = \quad \frac{y-\mu_j^I}{\sigma_j^I}. \tag{5.7}$$

A more involved discussion of score normalization of this form can be found in [63]. If one further imposes the constraint $\sigma_j^I =$ *a constant* because it is non-informative, one obtains:

$$y_j^{Z'} \quad = \quad y - \mu_j^I. \tag{5.8}$$

We call this expression *Z-shift*. Note that the constant can be discarded as the common threshold in the decision function of (5.4) can be adjusted accordingly.

EER-norm Based Normalization

Note that Z-norm is *impostor* centric because it relies only on the impostor distribution. A *Client-impostor* centric normalization was also studied in [43] and has two variants:

$$y^{TI1} \quad = \quad y - \Delta'_j \tag{5.9}$$
$$y^{TI2} \quad = \quad y - \Delta_j \tag{5.10}$$

where Δ'_j is a threshold found by assuming that the class-conditional distribution is Gaussian and Δ_j is found empirically. Δ'_j takes the form of (4.13) with the difference that all the user-independent terms are replaced by the user-specific terms, i.e., $\frac{\mu_j^I \sigma_j^C + \mu_j^C \sigma_j^I}{\sigma_j^I + \sigma_j^C}$. In reality, the empirical version (5.10) cannot be used when only one or two user-specific genuine scores are available.

Another study conducted in [139] used a rather heuristic approach to estimate the user-specific threshold. This normalization is defined as (the rest of the approaches can be seen as an approximation to this one):

$$y^{mid} \quad = \quad y - \underbrace{\frac{\mu_j^I + \mu_j^C}{2}} \tag{5.11}$$

The under-braced term is consistent with the term Δ'_j in (5.9) when one assumes that $\sigma_j^C = \sigma_j^I = 1$.

Common characteristics of User-Specific Score Normalization Procedures

All the procedures presented here are linear with respect to the score, i.e., $y^m = \frac{y - B_j}{A_j}$ where the scaling factor and bias, (A_j, B_j) are dependent on the statistics of user-specific distribution. This characteristic also generalizes to the F-norm to be discussed in Chapter 7.

5.6 User-Specific Threshold

Considering the vast amount of works on user-specific threshold procedures, we will provide a brief survey here. They are summarized in Table 5.1. These procedures are categorized by their type (i.e., client, impostor or client-impostor centric), the biometric modality applied to and whether they use a global threshold or not. The inclusion of a global threshold (e.g., rows 3, 6 and 9 of Table 5.1) is important for association with user-specific score normalization (see Section 5.7) and for providing an added degree of flexible or refinement to the local threshold.

Admittedly, most works are reported in the speaker verification community and few come from other biometric domains. This is because there are conditions (notably the fact that the client and impostor sets of scores each follows approximately a normal distribution) that make threshold normalization procedures *more effective* in the state-of-the-art systems used in speaker verification (mostly based on Gaussian Mixture Models or the like) than other systems[2].

[2]Our experimental outcome to be presented in Section 7.3.3 (in particular Figure 7.5(b)) suggests that score normalization procedures are more effective when applied to GMM-based systems than when applied to other systems.

Table 5.1: A survey of user-specific threshold methods applied to biometric authentication tasks.

No.	Equations	authors	type (-centric)	modality applied	use global threshold
1	$\Psi'_j(\Delta) = \alpha\left((\mu^I(j) + \sigma^I(j)\right) + \beta$	Furui [48]	impostor	speech	no †
2	$\Psi'_j(\Delta) = \alpha\mu^I(j)\sigma^I(j) + \beta\mu^I(j) + \gamma\sigma^I(j)$	Pierrot [92]	impostor	speech	no †
3	$\Psi'_j(\Delta) = \Delta - \underbrace{\left(\alpha\mu^I(j)\sigma^I(j) + \beta\mu^I(j) + \gamma\sigma^I(j)\right)}_{b}$	Genoud [52]	impostor	speech	yes ‡
4	$\Psi'_j(\Delta) = \mu^I(j) + \alpha\left(\sigma^I(j)\right)^2$	Lindburg *et al* [75]	impostor	speech	no †
5	$\Psi'_j(\Delta) = \alpha\mu^I(j) + (1-\alpha)\mu^C(j)$		client-impostor	speech	no †
6	$\Psi'_j(\Delta) = \Delta + \underbrace{\alpha\left(\mu^C(j) - \mu^I(j)\right)}_{b}$		client-impostor	speech	yes ‡
7	$\Psi'_j(\Delta) = \alpha(\mu^I(j) + \beta\sigma^I(j) + (1-\alpha)\mu^C(j)$	Chen [18]	client-impostor	speech	no †
8	$\Psi'_j(\Delta) = \mu^C(j) - \alpha\sigma^C(j)$	Saete *et al* [126]	client	speech	no †
9	$\Psi'_j(\Delta) = \underbrace{\mu^I(j)}_{b} + \underbrace{\sigma^I(j)}_{a}\Delta$	Jonsson *et al* [64]	impostor	face	yes ∗

Parameters a and b correspond to those found in (5.14). †: For these equations (which use a global threshold), the b term corresponds to the right hand-side of the respective equation and $a = 0$. ‡: For these equations, $a = 1$. ∗: Although went unnoticed by the author, this is *exactly* the dual form of Z-norm and was applied to a correlation-based matcher.

5.7 Relationship Between User-Specific Threshold and Score Normalization

The user-specific score normalization in Section 5.5 and user-specific threshold normalization in Section 5.6 are strongly related. Taking the right-hand sides of (5.4) and (5.5), we have:

$$\Psi_j(y) > \Delta, \tag{5.12}$$

$$y > \Psi'_j(\Delta). \tag{5.13}$$

Note that the threshold Δ refers to the threshold found *after* applying a respective user-specific score normalization procedure and *not before* (i.e., not directly on the scores prior to normalization).

To show that they are dual, we will re-express (5.13) into the form of (5.12). To do so, it is necessary to assume that $\Psi'_j(\Delta)$ takes the following form, as a function of Δ:

$$\Psi'_j(\Delta) = a\Delta + b. \tag{5.14}$$

Note that all equations in Table 5.1 can be expressed by (5.14) using different a and b. In particular, for those which do not contain a global threshold, b corresponds to the right hand-sides of the equations. For those using a global threshold, any multiplicative factor to the global threshold will be represented by a and the rest of the terms are represented by b. Replacing (5.14) into (5.13), and after rearrangement, we obtain:

$$\frac{y-b}{a} > \Delta \tag{5.15}$$

From (5.12) and (5.15), we see that:

$$\Psi_j(y) = \frac{y-b}{a}, \tag{5.16}$$

For equations whose $a = 0$, we have:

$$\Psi_j(y) = y - b, \tag{5.17}$$

As a result, manipulating the threshold or the score y has *exactly the same* effect. Hence, the threshold refinement procedure (row thee of Table 5.1) is just another score normalization technique. The additional advantage of score normalization over threshold normalization is the additional flexibility provided by the global threshold which can still be adjusted to different operating costs of false acceptance and false rejection.

5.8 Summary

In this chapter, we survey user-specific processing, i.e., a family of techniques that considers the user claimed user index. These techniques can be categorized into three types, according to the level of information dealt with, i.e., feature level, model level, and score level. User-specific score-level processing can further be divided into three types: user-specific fusion, user-specific score normalization and user-specific threshold procedure. Although user-specific processing is extremely useful and has been shown by numerous authors, this is the first survey written on the subject.

There are two somewhat original ideas in this chapter. Firstly, by analyzing the decision function using LLR, we unify the three types of user-specific score-level processing in a single framework. Thanks to the framework, user-specific score normalization can be seen as a special case of user-specific fusion having only a single system. This observation has a significant influence in our work because user-specific fusion techniques can suddenly be used as user-specific score normalization techniques, e.g., Chapter 6, and vice-versa, e.g., Chapter 7.

Secondly, we show that, in theory, user-specific score normalization and user-specific threshold procedure are equivalent. In practice, however, one may not obtain exactly the same result depending on the optimization criterion used and on whether or not the global threshold is considered for decision making. Between these two, user-specific score normalization is more advantageous due to an added degree of flexibility – the global threshold which can still be tuned after the normalization. We will therefore focus only on user-specific score normalization. This survey has not been published yet.

Thanks to the survey, we identify our contributions in user-specific processing as follows:

- An original compensation scheme that combines both user-specific and user-independent fusion classifiers consisting of N participating systems (Chapter 6). This framework generalizes to the case of $N = 1$ which can be considered as a novel user-specific score normalization.

- A user-specific score normalization called the "F-norm" and a user-specific fusion classifier called the "OR-switcher". (Chapter 7).

Chapter 6

Compensating User-Specific with User-Independent Information

6.1 Introduction

While prior works on user-specific fusion require many user-specific genuine samples (apart from those used to train the base-systems) in order to outperform the conventional fusion classifiers, e.g., as many as ten in [139] and six in [41], our goal in this chapter is to reduce the number of required user-specific genuine training samples to one or two.

This chapter contains two original ideas. The first idea is on the design of a user-specific fusion classifier that is in fact a Gaussian classifier with highly constrained Bayesian adaptation. Our novelty lies on the introduction of a set of useful constraints representing the domain knowledge. The second idea is referred to as a *compensation scheme* since one combines both the outputs of a user-specific fusion classifier (based on the first idea) and a user-independent (conventional) fusion classifier. The scheme is advantageous for three reasons. Firstly, it compensates for the possibly unreliable (due to lack of training data) but useful user-specific fusion classifier. Secondly, both the underlying fusion classifiers can be trained independently of each other. Thirdly, both the fusion classifiers are likely to be independent of each other thanks to the "phenomenon of large number of users". This phenomenon is based on our observation that when the number of users is large, the class-conditional score likelihood of a population is independent of that of a given user (who can be a member of the population). The scheme is in fact very general because it extends to the case where only a single system is involved; hence resulting in a compensated user-specific score normalization procedure.

Chapter Organization

Section 6.2 analyzes the effect of large number of users. The two original ideas – a compensation scheme and a user-specific classifier – are discussed in Section 6.3. The scheme is then empirically evaluated in Section 6.4. Finally, Section 6.5 draws the conclusions.

6.2 The Phenomenon of Large Number of Users

The idea of user-specific versus user-independent information is deeply related to the phenomenon of *large number of users*. To show this property, let the class-conditional score distribution be $p(\mathbf{y}|k)$ for $k = \{C, I\}$, where $\mathbf{y} = [y_1, \ldots, y_N]'$ is the vector of N system outputs to be combined. Note that the vector $\mathbf{y}$ generalizes to the case of a single system, i.e., $N = 1$. The likelihood of the user-independent $\mathbf{y}$

is thus a result of *accumulated* user-specific likelihood $\mathbf{y}$ of all identities $j \in \mathcal{J}$, i.e.,

$$\begin{aligned} p(\mathbf{y}|k) &= \frac{1}{P(k)} \sum_{j \in \mathcal{J}} p(\mathbf{y}, k, \mathrm{ID} = j) \\ &= \frac{1}{P(k)} \sum_{j \in \mathcal{J}} p(\mathbf{y}|k, \mathrm{ID} = j) P(k, \mathrm{ID} = j), \end{aligned} \tag{6.1}$$

where $P(k, \mathrm{ID} = j)$ denotes the prior probability of an impostor claiming identity j, i.e., $P(I, \mathrm{ID} = j)$, or the prior probability of user j making an identity claim, i.e., $P(C, \mathrm{ID} = j)$.[1]. We will now single out a particular user $j_* \in \mathcal{J}$ from the rest of the users.

$$p(\mathbf{y}|k) = \frac{1}{P(k)} \Big(p(\mathbf{y}|k, \mathrm{ID} = j_*) P(k, \mathrm{ID} = j_*) + p(\mathbf{y}|k, \mathrm{ID} \neq j_*) P(k, \mathrm{ID} \neq j_*) \Big) \tag{6.2}$$

Assuming the independence $P(k, \mathrm{ID}) = P(k)P(\mathrm{ID})$ and equal priors, i.e., $P(\mathrm{ID} = j) = \frac{1}{J}$ for all $j \in \mathcal{J}$, $P(\mathrm{ID} = j_*) = \frac{1}{J}$ and $P(\mathrm{ID} \neq j_*) = 1 - \frac{1}{J}$. As a result, (6.2) can be written as:

$$\begin{aligned} p(\mathbf{y}|k) &= \frac{1}{P(k)} \Big(p(\mathbf{y}|k, \mathrm{ID} = j_*) P(k) \frac{1}{J} + p(\mathbf{y}|k, \mathrm{ID} \neq j_*) P(k) (1 - \frac{1}{J}) \Big) \\ &= p(\mathbf{y}|k, \mathrm{ID} = j_*) \frac{1}{J} + p(\mathbf{y}|k, \mathrm{ID} \neq j_*) (1 - \frac{1}{J}) \\ &\approx p(\mathbf{y}|k, \mathrm{ID} \neq j_*) \text{ when } J \to \infty. \end{aligned} \tag{6.3}$$

We observe that when the number of users, J, is large, the user-specific likelihood, $p(\mathbf{y}|k, \mathrm{ID} = j_*)$, cannot contribute significantly to the overall population likelihood, $p(\mathbf{y}|k)$. Because of this phenomenon, one can model $p(\mathbf{y}|k)$ by a mixture of user-independent (and hidden) components. Let the n-th user-independent component be denoted by c_n and there are N^k_{cmp} components for each class k. The user-independent likelihood can be estimated by:

$$p(\mathbf{y}|k) \equiv \sum_{j=1}^{J} P(\mathrm{ID} = j) p(\mathbf{y}|k, \mathrm{ID} = j) \tag{6.4}$$

$$\approx \sum_{n=1}^{N^k_{cmp}} P(c_n) p(\mathbf{y}|k, c_n) \tag{6.5}$$

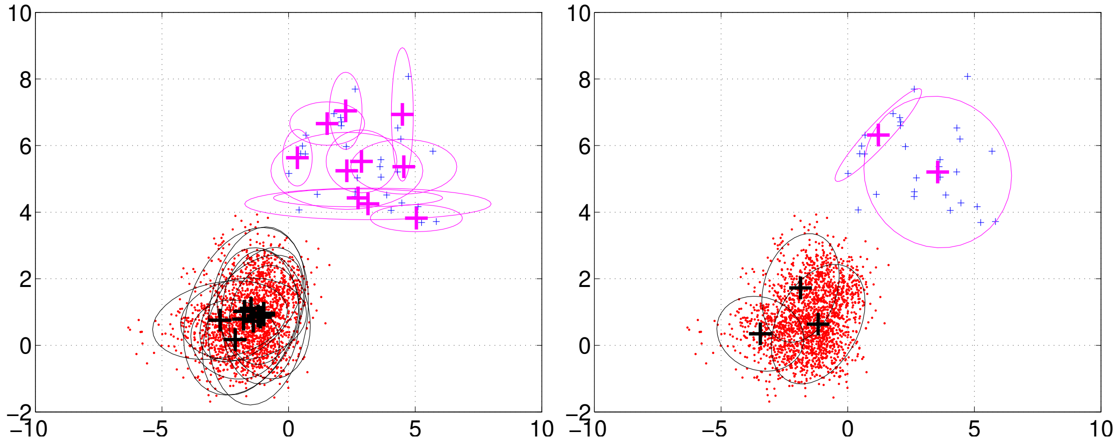
where both $p(\mathbf{y}|k, c_n)$ and $p(\mathbf{y}|k, \mathrm{ID})$ are each modeled by a Gaussian distribution. The difference, however, is that the number of Gaussian components is much fewer than the number of users available, i.e.,

$$N^k_{cmp} \ll J. \tag{6.6}$$

An Illustration

This phenomenon is illustrated in Figure 6.1. We randomly chose 10 users out of 200 for one of the XM2VTS fusion tasks. In Figure 6.1(a), the user-specific class conditional score density, $p(\mathbf{y}|k, j)$ is represented by a single Gaussian, for each $k\{C, I\}$ and each $j = \{1, \ldots, 10\}$. In Figure 6.1(b), by ignoring the claimed identity, the density $p(\mathbf{y}|C)$ requires only two mixture of Gaussian components whereas $p(\mathbf{y}|I)$ requires only three. The number of Gaussian components was tuned by cross-validation. In both cases, the number of Gaussians in the mixture is always smaller than the number of users. Therefore, (6.6) is always true.

[1] In real application, the user with high probability of being imposed will have high $P(I, \mathrm{ID} = j)$ and the user who uses more frequently the system than the rest will also have high $P(C, \mathrm{ID} = j)$.



(a) User-specific (b) User-independent

Figure 6.1: An illustrative example of the independence between user-specific and user-independent information. For both figures, the X- and Y-axes are the output score-space of a face and speech systems, respectively. The upper right clusters are client accesses whereas the lower left clusters are impostor accesses. In (a), the user-specific class conditional score distribution is represented by a single Gaussian distribution. Note that these distributions are very different from each other, especially for the client class. In (b), by not using the claimed identity, the user-independent class-conditional distribution requires a significantly lesser number of Gaussian mixtures.

6.3 An LLR Compensation Scheme

Section 6.3.1 proposes the compensation framework between user-specific and user-independent classifiers and two of its possible forms of realization, i.e., a fusion and a score normalization procedure. Section 6.3.2 discusses the design issue related to the user-specific classifier which requires a special attention due to few training samples.

6.3.1 Fusion of User-Specific and User-Independent Classifiers

In Chapter 5, we have motivated the use of the following form of user-specific decision:

$$\Psi_j(\mathbf{y}) > \Delta,$$

whereby a user-specific fusion classifier, $\Psi_j(\mathbf{y})$ is used in conjunction with a common (user-independent) threshold Δ where $\mathbf{y} = [y_1, \ldots, y^N]'$ is a vector of system scores to be combined (see (5.4)). However, considering the fact that $\Psi_j(\mathbf{y})$ is potentially unreliable, we could consider the following form instead:

$$\gamma\Psi_j(\mathbf{y}) + (1-\gamma)\Psi(\mathbf{y}) > \Delta,$$

where $\Psi(\mathbf{y})$ is a user-independent fusion classifier and $\gamma \in [0, 1]$ adjusts the contribution of the two classifier outputs. We will consider the user-specific and user-independent fusion classifier below:

$$\Psi_j(y) = \log \frac{p(\mathbf{y}|C, \mathrm{ID} = j)}{p(\mathbf{y}|I, \mathrm{ID} = j)} \tag{6.7}$$

and

$$\Psi(y) = \log \frac{p(\mathbf{y}|C)}{p(\mathbf{y}|I)}, \tag{6.8}$$

respectively. There are three advantages using the above form because of:

- **Mutual compensation:** The solution compensates for the potentially unreliable user-specific classifier but at the same time, enhances the user-independent classifier with a user-specific one.
- **Hybrid learning algorithms:** Both classifiers can be trained independent of each other. For instance, in practice, $\Psi_j(\mathbf{y})$ is restricted to Gaussian classifier due to the lack of training data whereas $\Psi(\mathbf{y})$ can be implemented using any general purpose fusion classifier described in Section 3.4. This is perfectly logical since there is no reason to restrict $\Psi(\mathbf{y})$ to be a Gaussian classifier.
- **Independence of information:** Following the justification in Section 6.1 that $p(\mathbf{y}|k$ is independent of $p(\mathbf{y}|k, \text{ID} = j$ when J is large, it is reasonable to expect that $\Psi(\mathbf{y})$ and $\Psi_j(\mathbf{y})$ are also likely to be independent. This is highly desirable because combining independent outputs will lead to improved generalization performance.

An Overview of Compensation Scheme

Consistent with our discussion in Part I, we will now restrict the classifiers Ψ and Ψ_j to those that output LLR scores. We will also consider two specific cases in which the proposed compensation scheme can be realized: a single-modal system where $N = 1$ and a multimodal system where $N > 1$. The realization for both cases are:

$$y_{com} \quad = \quad f_{adjust}\big(\Psi_j(y), \Psi(y)\big) \tag{6.9}$$

and

$$y_{com} = f_{adjust}\big(\Psi_j(\mathbf{y}), \Psi(\mathbf{y})\big), \tag{6.10}$$

respectively, where:

1. $f_{adjust} : \mathbb{R}^2 \rightarrow \mathbb{R}$ is a linear combination function of two LLRs. In theory, any trainable linear classifier discussed in Sections 3.4.2 and 3.4.4 can be used. We choose two techniques: one is trainable via SVM and the other one is a fixed rule using the mean operator such that $\gamma = \frac{1}{2}$.
2. $\Psi : \mathbb{R}^N \rightarrow \mathbb{R}$ is a fusion classifier that outputs LLR scores. While we choose a GMM classifier for this purpose, any classifier discussed in Sections 3.4.3 and 3.4.4 can be used. In the case where $N = 1$, Ψ reduces to a user-independent/system-level score normalization procedure, i.e., f_{LLR} as described using Algorithm 2 in Section 3.3.
3. $\Psi_j : \mathbb{R}^N \rightarrow \mathbb{R}$ is a user-specific fusion classifier. Due to lacking user-specific data, a careful treatment is required. This is discussed in Section 6.3.2. Note that in the case $N = 1$, Ψ_j reduces to a user-specific score normalization procedure. In theory, the ideal form of solution is given by (5.6). In practice, however, approximated solutions using Z-,F- and EER-norms are simpler to implement (see Section 5.5). We will deal with Ψ_j in the context of fusion and generalizes the result to the case $N = 1$. The approximated solutions will not be dealt with here.

As will be shown, Step 1 is crucial to guarantee the success of the scheme, especially when relying on Ψ_j alone can fail. Step 3 is particularly difficult to design because the problem is N-dimensional (corresponding to combining N system outputs). Consider the solution using a multivariate Gaussian. In this case, the covariance matrix must be estimated from at least $N + 1$ samples in order to ensure a non-singular covariance matrix. In most cases, this condition cannot be fulfilled unless one assumes a diagonal covariance matrix (in which case one cannot model the correlation among system outputs). Furthermore, due to the small training size, the obtained statistics may not be reliable. Section 6.3.2 deals with the design issue of user-specific fusion classifier.

6.3.2 User-Specific Fusion Procedure Using LLR Test

Approximating user-specific LLR is more difficult than approximating user-independent LLR since few user-specific data points are available, especially the genuine scores. The same difficulty does not apply to the user-specific impostor scores because these scores can be generated by using an external database. We tackle the lack of training data using the following rules:

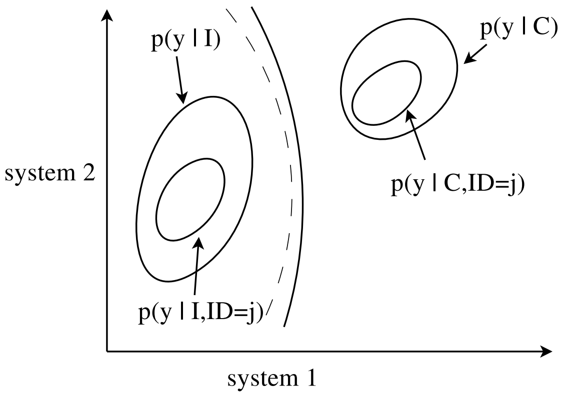


Figure 6.2: An illustration user-specific versus user-independent fusion of two system outputs. $p(\mathbf{y}|k, \text{ID} = j)$ is the j-th user's (hence user-specific) distribution whereas $p(\mathbf{y}|k)$ is a user-independent distribution, for $k = \{C, I\}$. The user-independent (global) decision boundary is drawn with a continuous line whereas the user-specific (local) decision boundary, for user j, is drawn with a dashed line. Each oval shape, as illustrated here, is a bivariate Gaussian.

1. Use simple classifier model (with low degree of freedom)
2. Estimate parameters using reliable data only
3. Rely on some prior knowledge such as user-independent distribution.

Because of few user-specific data points, the best one can do is to assume that each class of user-specific scores is normally distributed. The first rule implies that using more than one Gaussian components as in the user-specific case will probably result in overfitting. We present here two classifiers based on the concept of Maximum *A posteriori* (MAP) adaptation.

User-Specific Gaussian Classifier

The idea of user-specific fusion classifier, implemented as a Gaussian classifier, is illustrated in Figure 6.2. There are essentially two decision boundaries, one is user-independent (the classical solution) and the other is user-specific. Although using only user-specific information seems to be the best approach, in practice, one has extremely few samples to estimate the user-specific parameters reliably. The optimal solution is therefore found somewhere between the two decision boundaries. A good and proven solution is to use Bayesian adaptation which has been successfully deployed in speaker verification [122]. A simplified framework using a single multivariate Gaussian (with a diagonal covariance matrix) was used in [41]. The user-specific classifier, in its most general form, is shown in (6.7). The solution proposed in the literature on speaker verification is the so-called Maximum *A posteriori* (MAP) adaptation. In our context, this classifier can be written as:

$$\Psi_j^{qda}(y) = \log \frac{\mathcal{N}(\mathbf{y}|\boldsymbol{\mu}_{adapt,j}^C, \boldsymbol{\Sigma}_{adapt,j}^C)}{\mathcal{N}(\mathbf{y}|\boldsymbol{\mu}_{adapt,j}^I, \boldsymbol{\Sigma}_{adapt,j}^I)} \tag{6.11}$$

where $\boldsymbol{\mu}_{adapt,j}^k$ and $\boldsymbol{\Sigma}_{adapt,j}^k$ are the *adapted* class-conditional mean and covariance as respectively, for $k = \{C, I\}$ and for user j. The adapted parameters are defined by

$$\boldsymbol{\mu}_{adapt,j}^k = \boldsymbol{\mu}_j^k \gamma_1^k + \boldsymbol{\mu}^k (1 - \gamma_1^k) \tag{6.12}$$

and

$$\boldsymbol{\Sigma}_{adapt,j}^k = \boldsymbol{\Sigma}_j^k \gamma_2^k + \boldsymbol{\Sigma}^k (1 - \gamma_2^k), \tag{6.13}$$

respectively. Both parameters γ_1^k and γ_2^k (for the first and second moments) are within the range $[0, 1]$. This form of adaptation can be found in [51] and is called Maximum A Posteriori (MAP) adaptation by the authors. They balance between the user-specific estimate and the user-independent estimate of the two Gaussian parameters.

One can recognize that the Gaussian classifier Ψ_j shown in (6.11) is a Quadratic Discriminant Analysis (QDA) classifier when $\mathbf{\Sigma}^C_{adapt,j} \neq \mathbf{\Sigma}^I_{adapt,j}$ and as a special case, a Linear Discriminant Analysis (LDA) classifier when $\mathbf{\Sigma}^C_{adapt,j} = \mathbf{\Sigma}^I_{adapt,j}$. The only difference between the usual MAP adaptation as implemented in speaker verification is that only a single Gaussian component is used here as opposed to a mixture of Gaussians.

Due to few genuine samples, the determination of the four γ_i^k parameters for $k \in \{C, I\}$ and $i = \{1, 2\}$ is unfortunately problematic in practice since one cannot use cross-validation. This subject is somewhat involved and will be discussed in Section 6.3.3.

User-Specific GMM Classifier

Note that (6.11) imposes the constraint that the user-independent distribution ($p(\mathbf{y}|k)$) is also a Gaussian distribution. In reality, it must be a mixture of Gaussian distributions since it contains many different users. To take this fact into consideration, we use the following user-specific classifier:

$$\Psi_j^{gmm}(y) = \log \frac{\gamma^C p(\mathbf{y}|C, \mathrm{ID} = j) + (1 - \gamma^C) p(\mathbf{y}|C, \mathrm{ID} \neq j)}{\gamma^I p(\mathbf{y}|I, \mathrm{ID} = j) + (1 - \gamma^I) p(\mathbf{y}|I, \mathrm{ID} \neq j)} \tag{6.14}$$

where $p(\mathbf{y}|k, \mathrm{ID} = j)$ is a Gaussian distribution of the form $\mathcal{N}(\mathbf{y}|\boldsymbol{\mu}_j^k, \mathbf{\Sigma}_j^k)$ and $p(\mathbf{y}|k, \mathrm{ID} \neq j)$ is a mixture of Gaussian distributions of the rest of the users, i.e.,:

$$p(\mathbf{y}|k, \mathrm{ID} \neq j) = \sum_{j' \in \mathcal{J} - j} p(\mathbf{y}|k, \mathrm{ID} = j') \tag{6.15}$$

Note that γ^k can be interpreted as a prior probability $P(k, \mathrm{ID} = j)$ and $1 - \gamma^k$ as the prior probability of $P(k, \mathrm{ID} \neq j)$. We use $\gamma \equiv \gamma^C = \gamma^I$. The use of γ^k again in reminiscent of MAP adpatation in the user-specific Gaussian classifier. The difference is that, in (6.14), γ^k weighs LLRs instead of Gaussian parameters. (6.14) is different from the standard GMM used in speaker verification because in our case, the Gaussian component is not hidden but is conditioned on the *observed* identity claim. For this reason, (6.14) is called a user-specific GMM classifier.

Similar to the user-specific Gaussian classifier, determining γ^k is again problematic because one is always lack of user-specific genuine training scores. In our experiments, a non-informative prior of these values are used, i.e., $\gamma^C = \gamma^I = 0.5$.

6.3.3 Determining the Hyper-Parameters of a User-Specific Gaussian Classifier

This Section deals with setting the hyper-parameters γ_1^k, γ_2^k for $k \in \{C, I\}$, as appeared in (6.12) and (6.13), respectively. At first sight, having the four free parameters γ_i^k to tune is too many if one considers that there are about a hundred user-specific impostor scores and about two user-specific client scores. One strategy is to parameterize γ_i^k via a relevance factor. This solution was reported in speaker verification [122]. We will propose another solution by pre-fixing some of the parameters, which is better suited to the problem of fusion. Both approaches are described below:

- **Relevance Factor:** A "relevance factor", r, parameterizes γ_i^k for all $i \in \{1, 2\}$ and $k \in \{C, I\}$ as a function of the number of available user-specific class-conditional samples N_j^k. The resultant γ_i^k is:

$$\gamma_i^k \equiv \frac{N_j^k}{N_j^k + r}, \tag{6.16}$$

  Note that the *relevance factor*, r, takes only positive values. In biometric authentication, where $N_j^I \gg N_j^C$, r will give more weight to the user-specific impostor Gaussian parameters than their

Table 6.1: Proposed pre-fixed values for γ_i^k

	class k	
i	C	I
1	tune	1
2	0	1

client counterparts. The use of relevance factor in a user-specific Gaussian classifier for fusion was reported in [41].

- **Pre-fixed Parameter:** Based on the observation that $N_j^I \gg N_j^C$, whereby a very large set of simulated impostors is available, we propose to fix $\gamma_1^I = \gamma_2^I = 1$, hence putting full confidence on the user-specific impostor estimates. Furthermore, we can also set $\gamma_2^C = 0$, hence, putting zero confidence on the user-specific client covariance estimate since it is likely to be unreliable due to the small size of training samples. These constraints effectively limit the degree of freedom tighter than the relevance factor. The result is that we are left with a single parameter $\gamma_1^C \in [0, 1]$ to tune. The pre-determined parameters will be justified by experiments in Section 7.2.

Differences with the User-Specific Gaussian Classifier Proposed in [41]

The proposed user-specific classifier here is undoubted very similar to that proposed in [41]. There are, however, two differences:

- The relevance factor was used in [41] while we use pre-defined values γ_i^k, which are shown in Table 6.1.

- A diagonal covariance matrix was used in [41] while we use a full covariance matrix which is capable of capturing the possible correlation among the system outputs.

It should be recognized that relevance factor is also a form of constraint. Otherwise, a different r for each k or for each i would have meant that one has still to tune the four parameters. In our case, we fixed these parameters *a priori* to further constrain the model fitting.

6.4 Experimental Validation of the Compensation Scheme

Choice of Database

For the purpose of experimental validation, we could not use the BANCA database because the BANCA protocols are defined as such that the development and evaluation sets consist of two different population sets of genuine of users. The XM2VTS database, on the other hand, satisfies our need[2] and will be used. The fusion tasks can be found in Section 2.1.1.

Section 6.4.1 first examines the compensation scheme in multimodal fusion whereas Section 6.4.2 reports a more detailed analysis on the experiments done.

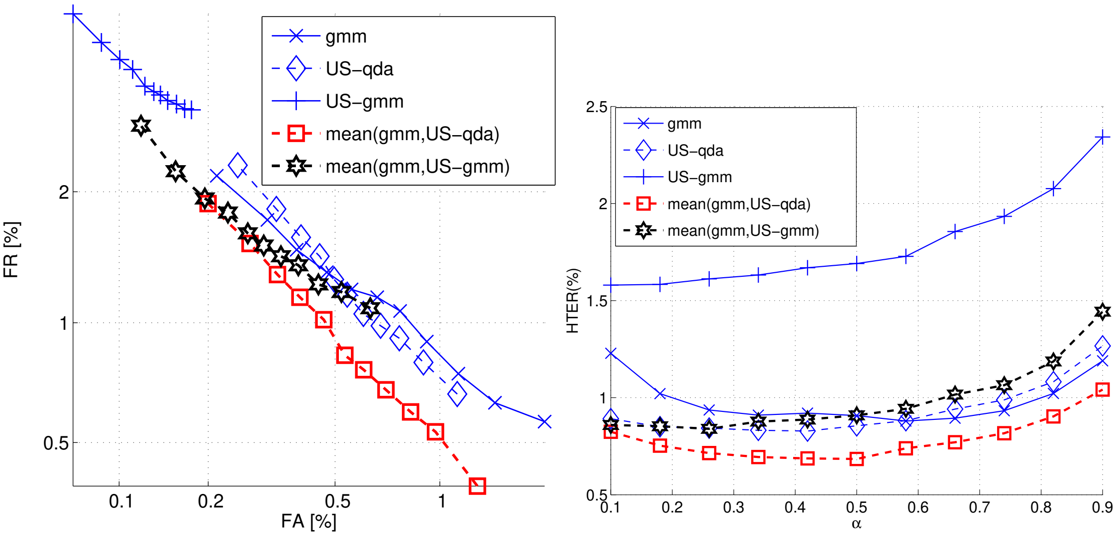
6.4.1 Pooled Fusion Experiments

For the multimodal fusion experiments, the following classifiers are used:

1. gmm – a user-independent GMM

2. US-gmm – a user-specific GMM as shown in (6.14)

[2]To be precise, the genuine users are found in both the development and evaluation sets but not the impostors.



(a) DET

(b) EPC

Figure 6.3: Experimental results validating the effectiveness the proposed compensation scheme between user-specific and user-independent fusion classifier on the 15 XM2VTS multimodal fusion tasks shown using (a) pooled DET curves and (b) EPC curves. “gmm” is user-independent fusion classifier, “US-qda” is a user-specific Gaussian-based fusion classifier, “US-gmm” is a user-specific GMM-based fusion classifier and the last two are two compensated classifiers combining the two classifiers using the mean operator.

3. US-qda – a user-specific Gaussian (QDA) classifier as shown in (6.11). The default γ parameters used are shown in Table 6.1 with $\gamma_1^C = 0.5$.

4. mean(gmm, US-qda) – a combination of gmm and US-qda using the mean operator

5. mean(gmm, US-gmm) – a combination of gmm and US-gmm using the mean operator

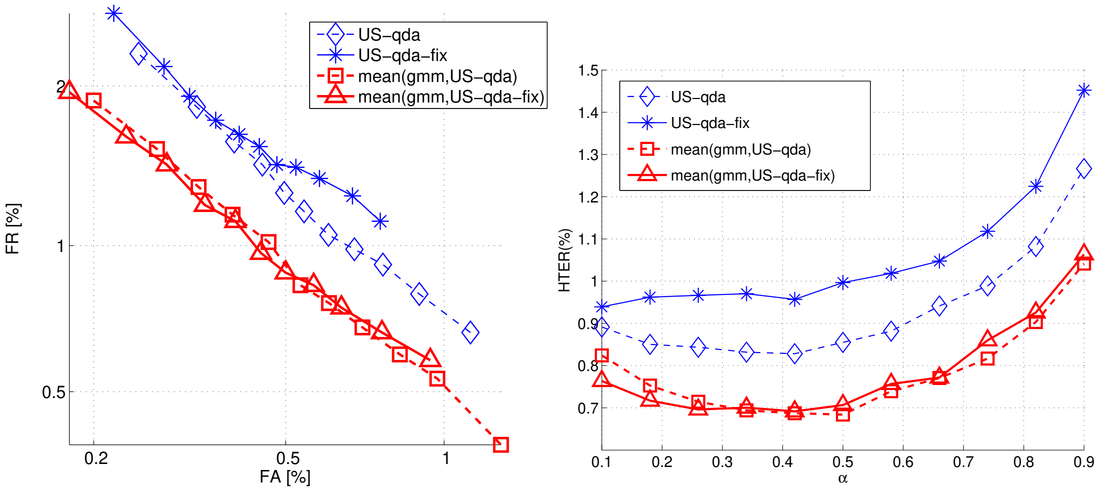
The results are shown in Figure 6.3. As can be observed the compensation scheme, particular mean(gmm, US-qda), results in the best generalization performance. The classifier US-gmm did not achieve the expected result because the classifier overfits the training data. Since this *biased* training data is used to tune the *a priori* chosen threshold, the resultant performance on the test set is thus sub-optimal[3]. This shows that using a full mixture of Gaussians, where each Gaussian represents the score density of a user, is not a suitable model since its capacity or degree-of-freedom is more than necessary. On the contrary, US-qda which highly restricts the model is an adequate choice.

6.4.2 Experimental Analysis

In this Section, we examine several factors that could influence the performance of the proposed compensation scheme, i.e.,:

- **Sensitivity to the γ parameter:** One of the difficulties related to constructing a user-specific fusion classifier is its instability and sensitivity to any hyper-parameters, i.e., parameters than control other parameters. In our case, these parameters are γ_i^k's. While a pre-determined set of γ_i^k values have been proposed in Table 6.1, it is still unclear how γ_1^C should be tuned. In the previous experiments, a non-informative prior (since it can be seen as a probability) of 0.5 was used throughout the experiments.

[3]When we plot a pool DET curve, the WER criterion was used so that each DET curve is aligned thanks to the α parameter of the WER criterion. To evaluate the WER criterion, two sets of (combined) scores are needed: one from the development and the other from the evaluation sets. The so-called development set of combined scores for US-gmm in this case is the output of US-gmm itself. Although a procedure such as cross-validation as described in Section A could have been used, for the purpose of algorithmic comparison, it was not used in all algorithms considered here.



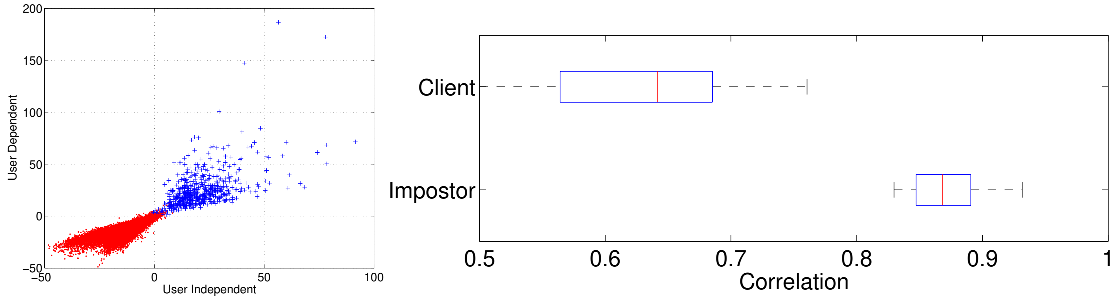
(a) DET (b) EPC

Figure 6.4: Multimodal fusion experimental results shown using (a) a DET and (b) EPC curve verifying the sensitivity of the compensation scheme with respect to the γ parameter of the user-specific fusion classifier. “US-qda” is the Gaussian classifier with the default $\gamma = 0.5$ and “US-qda-fix” is the same classifier with $\gamma = 1$. In both cases, the same user-independent fusion classifier compensation scheme and the fusion between the user-specific and user-independent classifier is a mean operator. SVM was used in place of the mean operator and this resulted in slightly degraded performance because it relies on the *biased* training which are outputs associated to the data its two base classifiers used to train on.

We repeated the experiments with $\gamma_1^C = 1$ and measured the generalization performance of the resultant compensated classifier. The results are shown in Figure 6.4. As can be observed, although setting $\gamma_1^C = 1$ degrades the performance of the user-specific Gaussian-based fusion classifier, its influence on the compensated classifier is insignificant on the resultant compensated classifier.

- **On the use of trainable fusion classifier in place of the mean operator to combine user-specific and user-independent classifier:** We replaced the mean operator with a logistic regression (LR) and found that the generalization performance degrades. Although in theory LR is better than the mean operator, in this case, the training data is *biased* since the data was used to construct the user-specific and user-independent fusion classifiers.

- **Correlation between the output of a user-specific and a user-independent fusion classifier:** Since our justification in Section 6.2 shows that the estimate of the class-conditional likelihood of the user-specific classifier and that of the user-independent classifier will be different when the number of users is large, it is natural to verify to what extent two LLR-based fusion classifiers carry complementary information. For this purpose, we measured the correlation between the class-conditional outputs of the two fusion classifiers. An example of the LLR scores are shown in Figure 6.5(a). In this case, two correlation values can be measured, each conditioned the client and impostor classes. We measured the correlation across all the 15 fusion experiments and their distributions are shown in Figure 6.5(b) as boxplots. As can be observed, the client LLR scores has lower correlation – indicating that the two classifiers are *more complementary* on the client accesses than on the impostor accesses. From Chapter 4, we know that lower correlation contributes to higher F-ratio of the combined scores. Hence, this shows that the compensation scheme is largely responsible for the *statistically significant* improvement of generalization performance.



(a) An Example

(b) Correlation

Figure 6.5: Correlation between user-independent and user-specific fusion classifier output. Figure (a) is an example of the scatter plot of the LLR scores. In this case, two correlation values can be calculated for each of the client and impostor sets of scores. Figure (b) is a vertical boxplot that shows the extent of the correlation values over the 15 fusion experiments.

6.5 Conclusions

This chapter proposes an *alternative* scheme to implement user-specific processing at the score level, a subject which has been investigated in [40, 41, 122, 139, 61, 71, 137]. The proposed scheme capitalizes on the use of user-specific and user-independent information sources. By representing both information sources as two Log-Likelihood Ratios (LLRs), e.g., one due to a user-specific fusion classifier and the other due to a user-independent one, the scheme proposes to linearly combine the output of these two fusion classifiers. Therefore, we call this scheme "fusion of fusion".

This proposed scheme has the following benefits:

- **Mutual compensation:** The solution compensates for the potentially unreliable user-specific classifier but at the same time, enhances the user-independent classifier with a user-specific one.
- **Hybrid learning algorithms:** Both classifiers can be trained independently of each other. This is an advantage since the user-specific classifier is limited to a Gaussian classifier, the user-independent one is not. The compensation scheme therefore *relaxes* the Gaussian assumption.
- **Independence of information:** Following the justification in Section 6.1, both classifiers are likely to complement each other when the number of users is large.

The compensation scheme compares favorably with [40, 41, 139, 71, 61] principally because it is the only one that can learn from *very few* user-specific genuine samples, which is a non-trivial machine-learning problem. A second advantage is that the *domain knowledge*, in the form of pre-fixed adaptation parameters (as in Table 6.1), is exploited in the user-specific classifier that we proposed. The class of solutions is therefore so highly constrained that the only free parameter, γ_1^C, has no strong influence on the overall system performance. This is the main difference between our proposed user-specific classifier and that reported in [41]. Our proposed scheme also compares favorably with [139] whereby due to the same problem, noise is injected to increase the number of user-specific client scores. Due to the Bayesian scheme, our approach handles such an uncertainty in a natural way.

Apart from those experiments reported here, in [105], we also considered the compensation scheme with a single system where $N = 1$, i.e., a user-specific score normalization procedure. Although the data sets and experimental settings are somewhat different, the conclusions remain the same.

Chapter 7

Incorporating User-Specific Information via F-norm

7.1 Introduction

This chapter offers an *alternative approach* to applying user-specific processing at the score level. In particular, four distinctive but related topics are analyzed. Firstly, we evaluate the robustness of class-conditional user-specific score statistics, i.e., the degree of invariance with respect to different train/test conditions of μ_j^k and σ_j^k for $k = \{C, I\}$ for each user j. Secondly, we investigate a new user-specific score normalization procedure that aims to *reduce the user-induced variability* and that possesses a list of desired characteristics, e.g., robustness to deviation from the class-conditional Gaussian assumption, to few user-specific genuine samples and to mismatch between train/test conditions. Thirdly, we design a criterion that is robust and that can rank users according to their ease of recognition after reducing the user-induced variability. Finally, we design a fusion classifier that selectively combines a subset of systems on a per person basis. This fusion classifier is a proof-of-concept of the effectiveness of the first three ideas since we literally put all the above findings into a single working algorithm.

Motivations

We describe below the motivations of investigating the four mentioned topics:

- **On the robustness of class-conditional user-specific score statistics:** Although user-specific statistics have been used extensively in user-specific score normalization (Section 5.5) and user-specific threshold (Section 5.6) procedures, to the best of our knowledge, no *systematic* study has been conducted to examine the *robustness* of these statistics. A user-specific score statistic is considered *robust* if it is *invariant* to different biometric samples, possibly separated over a fixed duration, of the same person for the client class, and of *different* persons for the impostor class. We expect the user-specific impostor statistics to be more robust than their client counterparts because there are simply more simulated impostor data[1] than client data. We are also motivated by the empirical findings by Doddington *et al* [33], which suggest that the user-specific statistics are different from one user to another. An important difference between our approach taken here and that of Doddington *et al*'s is that the authors did not consider the concept of robustness of statistics (to mismatch between training and test conditions). By considering the robustness of statistics, our aim is to devise algorithms that exploit only robust statistics for user-specific processing. The next three topics are examples of such processing.

- **On reducing user-induced variability:** Given that the user-specific score statistics are predictable to some extent, our next investigation is to design a user-specific score normalization procedure of

[1]Note that in reality, professional impostors should be used. Unfortunately, few databases today have such a data.

the form $\Psi_j : \mathbb{R} \rightarrow \mathbb{R}$ – taking a score as input and outputting a normalized score – that have *reduced* user variability. Two categories of score normalization have been surveyed in Section 5.5; they are Z-norm based methods ($\Psi_j^Z(y) = \frac{y-\mu_j^I}{\sigma_j^I}$) and EER-norm based methods ($\Psi_j^{eer}(y) = y - \Delta_j$ where $\Delta_j = \frac{\mu_j^I \sigma_j^C + \mu_j^C \sigma_j^I}{\sigma_j^I + \sigma_j^C}$). These two categories of techniques have their own short-comings. For instance, Z-norm does not consider client statistics and EER-norm relies heavily on the class-conditional Gaussian assumption due to its extensive use of second-order statistics. These two short-comings motivate us to investigate a new category of normalization that we call "F-norm".

- **On ranking users according to their ease of recognition:** In [33], Doddington *et al* showed that a minority of users are particularly difficult to be recognized – the so-called goats, some are easy to imitate – the lambs, and some are particularly successful at imitating others – the wolves. Although identifying these groups of users is important, there is no direct way to rank users according to their ease of recognition. In order to rank users, one has to *simultaneously* consider the user-specific client and impostor scores. A natural candidate to rank users is the F-ratio proposed in Chapter 4 except that it is applied on a per user basis. Directly applying user-specific F-ratio may fail because not all the user-specific statistics are equally robust. Therefore, this motivates us to design a robust equivalent of F-ratio with the possibility of reducing the user-induced variability.

- **On designing a selective user-specific fusion classifier:** Motivated by the fact that we have at our disposal a criterion to rank users given a system, we attempt to modify the criterion so that it can rank a subset of systems to combine, on a per person basis. Such a criterion can be used in a multi-modal biometric fusion whereby based on the criterion, a fusion operator decides an optimal subset of systems to combine, based on a validation data set. This fusion classifier is unique in its category because it is both *user-specific* and *selective*. It has at least two advantages. Firstly, the selective strategy means hardware cost saving for personal devices since an under-performing biometric system does not have to be built in the first place. Secondly, the authentication can be performed faster since not all biometric modalities are considered. The novel fusion technique is called the OR-switcher. Our experimental results suggest that, without using the selective strategy, the OR-switcher *always* outperforms the state-of-the-art fusion techniques. When the selective strategy is used, the performance of the OR-switcher can still outperform the state-of-the-art fusion techniques in some experimental settings. The added advantage, however, is that not all the participating systems need to be operational. Such a flexibility mimics our human ability where a person can still be recognized with only some partial evidences.

Chapter Organization

This chapter is organized as follows: Section 7.2 reports our experiments that objectively quantify the robustness of user-specific statistics. Section 7.3 proposes and evaluates the new user-specific F-norm. Section 7.4 designs a criterion to rank user. Section 7.5 presents the OR-switcher. Finally, Section 7.6 summarizes the original contributions presented in this chapter.

7.2 An Empirical Study of User-Specific Statistics

We have motivated the use of class-conditional Gaussian assumption when surveying user-specific score normalization in Section 5.4. One important concern is whether or not the user-specific statistics, μ_j^k or σ_j^k, are robust to the unseen data which may be different from the training conditions.

Choice of Data Set and Preparation

In order to answer this question, we analyzed the scores of the 13 systems in XM2VTS (Section 2.1). First, the score sets are divided into two subsets: a development set and an evaluation set, such that the same clients must be found in both sets of scores. The impostors, however, may be from two different sets of populations. The XM2VTS score data sets satisfy the requirement but not the BANCA score data sets.

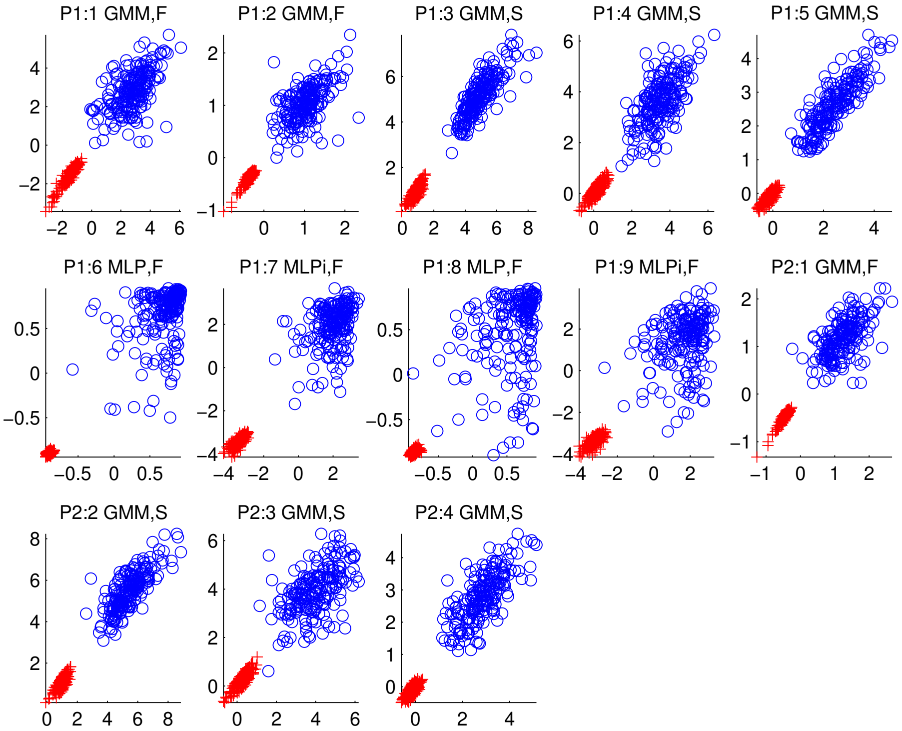


Figure 7.1: An initial study on the robustness of the user-specific mean statistic. User-specific conditional score mean of development set (Y-axis) versus that of evaluation set (X-axis), i.e., $\mu_j^k|dev$ versus $\mu_j^k|eva$, for $k = \{C, I\}$, of the 13 XM2VTS systems. There are 200 data points for each statistic which correspond to 200 users. Blue circles are genuine means whereas red plus signs are impostor mean.

This is because the g1 and g2 data sets in BANCA contain different population of clients. Note that the XM2VTS fusion protocols (see Section 2.1.1) have already defined both the development and evaluation sets. Whenever a system output is an MLP with sigmoid or hyperbolic tangent activation function, we convert the scores into LLR using Algorithm 2 (Section 3.3) to ensure that the scores follow a normal distribution. Both the original and the converted score data sets are used in the experiments. The original data set is labelled "MLP" whereas the converted one is labelled "MLPi" ('i' for probabilistic inversion). We kept these two data sets in order to study the effect of non-conformity of scores to the Gaussian assumption – a fundamental assumption of our proposed techniques.

Experimental Results

For each set of scores (development or evaluation), each class $k \in \{C, I\}$ and each user $j \in \mathcal{J}$, we computed the class-conditional (genuine and impostor) first and second-order moments (μ_j^k and σ_j^k). The statistics are then compared as follows:

- $\mu_j^k|dev$ versus $\mu_j^k|eva$ (see Figure 7.1)
- $\sigma_j^k|dev$ versus $\sigma_j^k|eva$ (see Figure 7.2)

for both classes $k \in \{C, I\}$ and all $J = 200$ users (hence 200 data points for each μ_j^C, σ_j^C, μ_j^I and σ_j^C).

One way to measure the degree of generalization or "agreement" is by computing correlation ρ_t^k between the statistic $t \in \{\mu, \sigma\}$ estimated on the development set and the one estimated on the evaluation set, for each class $k = \{C, I\}$. We summarize ρ_t^k of the 13 systems in Figure 7.3 as a box plot. Each box indicates the bounds of the upper and the lower quantiles. The two horizontal lines at the top and the bottom of a box covers the 95% confidence bound. Any data sample (correlation in this case) beyond this

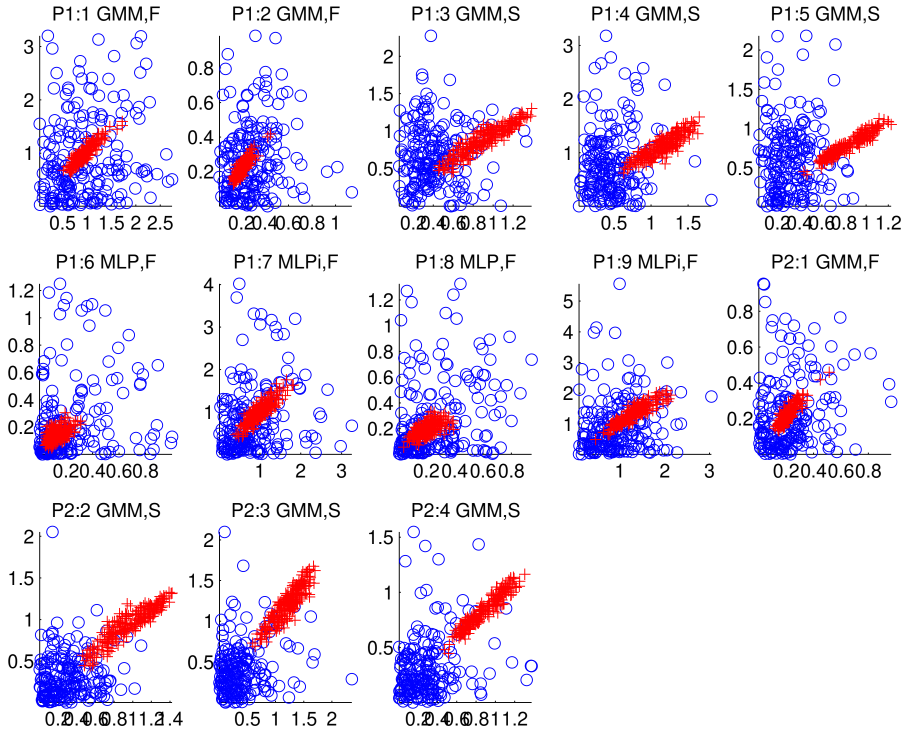


Figure 7.2: As per Figure 7.1, except that σ_j^C and σ_j^I are used in place of μ_j^C and μ_j^I. The X-axis is $\sigma_j^k|eva$ and the Y-axis is $\sigma_j^k|eva$

bound is denoted with a plus sign and is considered an outlier. Each bar contains 13 data samples. The higher the correlation, the more robust the statistic is. As can be observed and as expected, the user-specific impostor statistics are likely to be more robust than that of genuine, independent of the underlying systems. Note that there are two or three samples (depending on LP1 or LP2 protocol) to estimate the user-specific genuine statistics. Despite this fact, μ_j^C is still informative. On the other hand, σ_j^C is not at all informative, judging from its relatively low correlation (whose median is 0.2).

Note that the outliers (with very low correlation values; indicated by plus signs) are due to the MLP systems prior to converting the scores into LLR using Algorithm 2 (as discussed in Chapter 3). This is expected since the MLP user-specific class-conditional output scores are not normally distributed but are known to have a skewed distribution due to the nature of the non-linear activation function. As a result, their associated user-specific statistics generalize poorly compared to the rest of the systems. This shows that Algorithm 2 is *effective* in mitigating this systematic and undesirable effect.

7.3 User-Specific F-norm

This Section is divided into five sub-sections. Section 7.3.1 proposes the user-specific F-norm. The user-specific F-norm is then compared to other user-specific score normalization procedures in Section 7.3.2 theoretically and in Section 7.3.3 empirically. Section 7.3.4 improves the way F-norm is parameterized so that the number of user-specific genuine samples can automatically be taken into account. Finally, Section 7.3.5 illustrates the usefulness of F-norm in the context of fusion.

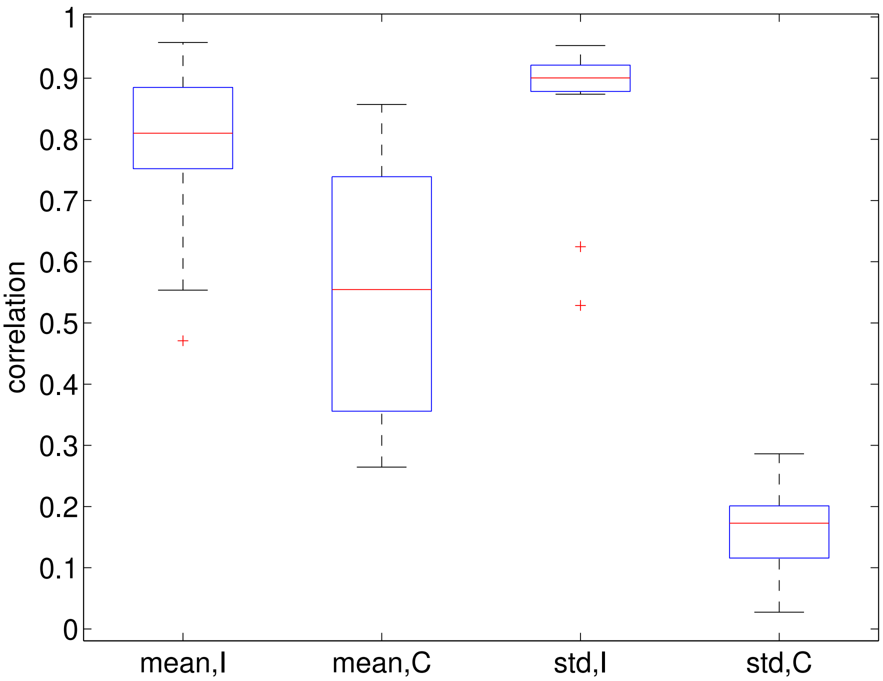


Figure 7.3: A summary of the robustness of user-specific statistics. Box plot of the conditional correlation $\rho^k, \forall_k$ of the four parameters, μ_j^I, μ_j^C, σ_j^I and σ_j^C of the 13 face and speech systems XM2VTS. Each correlation value is measured on 200 users. The two outliers (with plus signs) in σ_j^I are due to (MLP,F) of P1:6 and P1:8, respectively. Similarly the outlier in μ_j^I is due to (MLP,F) of P1:6.

7.3.1 Construction of User-Specific F-norm

The user-independent F-norm was derived in Section 4.4.3 and is given by (4.22). In the user specific context, one can simply replace the system output index i by the user-specific index j, hence giving:

$$y_j^F = \frac{y - \mu_j^I}{\mu_j^C - \mu_j^I}. \tag{7.1}$$

Directly using (7.1) can lead to a complete failure since μ_j^C cannot be estimated reliably. To account for such unreliability, a Bayesian solution is to compensate the user-specific statistic μ_j^C with the user-independent statistics μ^C via an adjustable parameter $\gamma \in [0, 1]$, i.e.,

$$\gamma\mu_j^C + (1-\gamma)\mu^C.$$

We have seen this solution in Chapter 6. Although this Bayesian solution is classical (and therefore not a heuristic), e.g., [56, Chap. 4], surprisingly, it has not been introduced to the user-specific score normalization or user-specific threshold procedures surveyed in Chapter 5. Thanks to the Bayesian solution, (7.1) can be rewritten as[2]:

$$y_j^F = \frac{y - \mu_j^I}{\gamma \mu_j^C + (1-\gamma)\mu^C - \mu_j^I}. \tag{7.2}$$

where γ has to be tuned. Two sensible default values are 0 when μ_j^C cannot be estimated because no data exists and at least 0.5 when there is only a single user-specific sample. γ thus accounts for the degree of reliability of μ_j^C and should be close to 1 when abundant genuine samples are available.

7.3.2 Theoretical Comparison of F-norm with Z-norm and EER-norm

In Section 5.5, two groups of user-specific score normalization procedures were surveyed, i.e.,

- **Z-norm based methods:** Two examples are Z-norm itself and Z-shift. For Z-norm, the user-specific statistics *after transformation* have the following characteristics: $\mu_j^I = 0$ and $\sigma_j^I = 1$ for all $j \in \mathcal{J}$. For Z-shift, only the constraint $\mu_j^I = 0$ is satisfied. The advantage of these methods are that μ_j^I and σ_j^I are robust statistics and can generalize across different impostor sets. Their weakness, however, is that they do not consider the user-specific client statistics.

- **EER-norm based methods:** These methods are based on EER-norm and its variants. The user-specific threshold, Δ_j, after applying these methods, becomes common to all users, i.e., $\Delta_j = \Delta_s = 0$ for all $j, s \in \mathcal{J}$. These methods, as represented by (5.9)–(5.11), differ only in their assumptions. The least assumption made among the three, (5.9), requires many more user-specific client data and hence is impractical. (5.10) makes the class-conditional Gaussian assumption but is unlikely robust due to the inclusion of σ_j^C which is uninformative when few user-specific genuine samples are available. Finally, the mid-point solution of (5.11) includes the μ_j^C statistic which may not be robust.

In comparison with these two families of score normalization techniques, the user-specific F-norm is another family of techniques. This is because the user-specific statistics *after applying F-norm* satisfy another set of constraints: $\mu_j^C = \mu_s^C$ and $\sigma_j^I = \sigma_s^I$ for all $j, s \in \mathcal{J}$ for the general case as proposed in our published paper [108] or $\mu_j^C = 1$ and $\sigma_j^I = 0$, for all $j \in \mathcal{J}$, for the F-norm proposed in (7.2)[3]. The advantage of (user-specific) F-norm over Z-norm is that F-norm considers the user-specific client statistic (μ_j^C). Hence, F-norm is client-impostor centric. F-norm's advantage over EER-norm is that it does not consider the non-robust second order σ_j^C statistic. Although F-norm uses the possibly non-robust μ_j^C, its γ parameter compensates for its unreliability. Figure 7.4 illustrates the differences among Z-, F- and EER-norms with respect to a list of characteristics just discussed. Table 7.1 summarizes the differences of Z-, F- and EER-norms.

Z-norm and F-norm share the common denominator but have different numerators. In Z-norm, the numerator is $\sigma_j^I = \sqrt{E\left[(y-\mu_j^I)^2\right]}$; and in F-norm, this term is $\mu_j^C - \mu_j^I$ by setting γ to 1. Both terms quantify some kinds of "score difference" in different ways but are in the same unit scale (domain). While Z-norm is impostor centric, F-norm can be seen as its improved version by incorporating the user-specific client information, making F-norm client-impostor centric. As a result, if F-norm can make use of the client information *reliably*, it can be superior over Z-norm.

In summary, F-norm possesses many interesting characteristics:

[2]The original form of F-norm was proposed in [108] and has the following form:

$$y_j^{F'} = \frac{y - \mu_j^I}{\gamma \underbrace{(\mu_j^C - \mu_j^I)} + (1-\gamma)\underbrace{(\mu^C - \mu^I)}}.$$

This version is superseded by (7.2) due to our finding in Section 7.2. Note that by setting $\gamma = 1$, both F-norm and its variant converge to the same solution. Their difference is thus rather subtle. Preliminary experiments on the XM2VTS fusion benchmark dataset show their generalization performance is not significantly different.

[3]The general case of user-specific F-norm and the special case proposed here are both theoretically and empirically equivalent, i.e., they result in exactly the same generalization performance. For this reason, we opted to present only the special (but also the simpler) case.

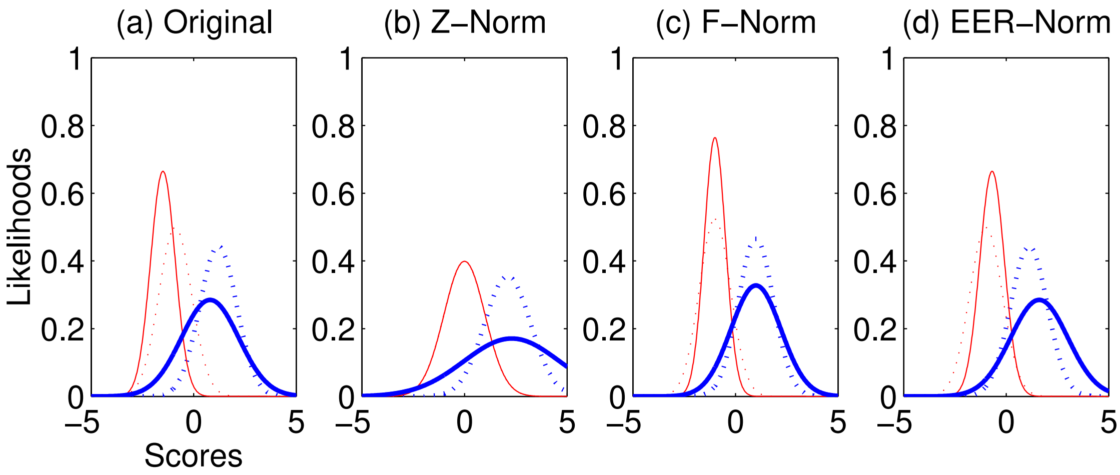


Figure 7.4: Comparison of the effects of Z-, F- and EER-norms. (a) The original distributions containing 2 user models (each represented by continuous and dotted lines; The genuine score distributions are plotted with thick lines and impostor score distributions with thin lines). A global threshold may not be optimal. (b) After applying Z-norm, the impostor distributions become normal whereas the client distributions vary. (c) After applying F-norm, all the client and impostor distributions are aligned so that a global threshold can be found easily. (d) After applying EER-norm, all the client and impostor distributions are aligned at their corresponding EER.

- It is more robust to departure from the Gaussian assumption since it does not rely on second-order statistics (an observation also remarked by Lindberg *et al* in [75]) in comparison with EER-norm.
- It is client-impostor centric as opposed to Z-norm which is only impostor-centric.
- It is more robust to few user-specific genuine training samples in comparison with EER-norm, since F-norm relies on user-independent information.

As a result, F-norm can be expected to perform better than Z-norm or EER-norm. Having compared the procedures *qualitatively*, the next Section will compare them quantitatively.

Table 7.1: Qualitative comparison between different user-specific normalization procedures.

Characteristics	Z-norm	F-norm	EER-norm
Formula	$\frac{y-\mu_j^I}{\sigma_j^I}$	$\alpha'(y-\mu_j^I)$ where $\alpha' = \mu_j^C\gamma + (1-\gamma)\mu^C - \mu_j^I$	$y - \frac{\mu_j^I\sigma_j^C + \mu_j^C\sigma_j^I}{\sigma_j^I + \sigma_j^C}$
Use second-order user-specific statistic	yes	no	yes
centric type	impostor	client-impostor	client-impostor
Rely on global information	no	yes	no
Robustness to few user-specific accesses	moderate	high with $\beta = 0.5$ low with $\beta = 1$	low

7.3.3 Empirical Comparison of F-norm with Z-norm and EER-norm

In this section, we designed several experiments to validate the following hypotheses in comparison with Z- and EER-norms:

(1) F-norm works with *fewer* samples.

(2) F-norm improves *faster* with increasing training genuine samples.

(3) F-norm is *more robust* to deviation from the class-conditional Gaussian assumption.

The NIST2005 database is used to test these hypotheses because it has abundant user-specific genuine accesses[4]. To test hypotheses (1) and (2), we chose a subset of users all having at least 7 accesses. The experiment is conducted for each user until all the users are processed. For each user, 7 partitions of equal size are created such that each partition contains exactly one genuine score (but can have many more impostor scores). One of the partitions is reserved as a test set whereas the other 6 partitions are used as training sets. 6 training sets are created by adding one partition of data at a time, such that the first training set is a subset of the second training set, the second is a subset of the third, and so on. These 6 training sets simulate the scenario where more data is available in an incremental manner. Although having 6 training sets, there is only one and *common* test set. All 6 normalization procedures, i.e., the baseline without normalization, EER-norm, Z-norm, Z-shift, F-norm with $\gamma = 1$ and F-norm with $\gamma = 0.5$, are tested on all the 24 systems and all the 6 training sets. This experimental setting results in

$$6 \text{ training sets} \times 6 \text{ normalization procedures} \times 24 \text{ systems} = 864 \text{ EPC/DET curves.}$$

Due to the large amount of data, we chose to evaluate only the point $\alpha = 0.5$ on the EPC. The results are shown in Figure 7.5(a). Note that each curve is calculated from the *pooled* HTER of all 24 systems. Based on the experiments, we conclude that:

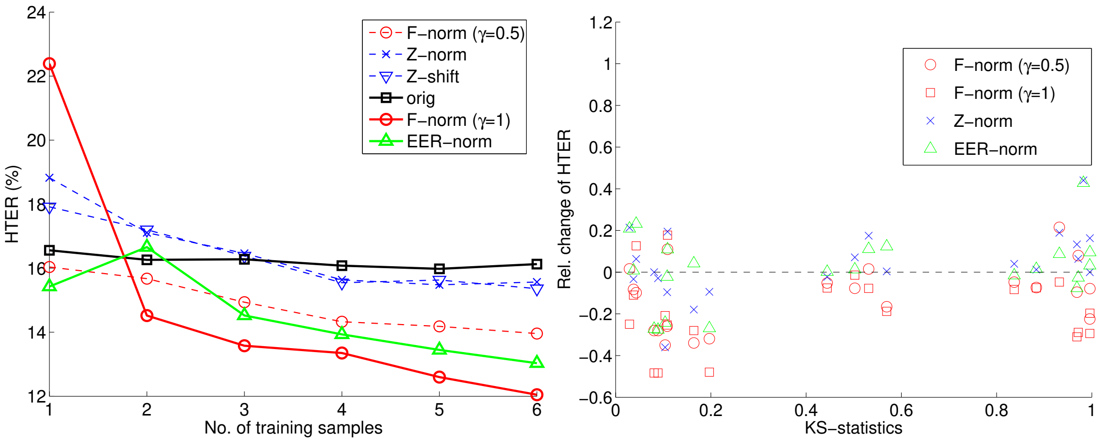
- Increasing training samples can improve the generalization performance of user-specific score normalization;
- Client-impostor centric procedures i.e., F-norm and EER-norm, are generally better than the classical impostor centric procedures, i.e., Z-norm and Z-shift.
- Large γ value of F-norm is favorable with increasing training sample size.
- F-norm with $\gamma = 0.5$ can improve over the baseline systems (without normalization) even with a single genuine sample.

As for hypothesis (3), it is necessary to measure the degree of deviation from Gaussian. We used the KS-statistic for this purpose and it is calculated on the scores prior to applying any user-specific score normalization procedure. It is calculated as

$$\max |\hat{\Psi}(y|I) - \Psi(y|\mu^I, (\sigma^I)^2)|,$$

where $\hat{\Psi}(y|I)$ is the estimated *cdf* of the impostor scores and $\Psi(y|\mu^I, (\sigma^I)^2)$ is the *cdf* of the impostor scores assuming that the scores are normally distributed. Note that the same statistic but for the genuine scores are not used because the statistic is less robust due to much fewer samples. We then plotted the relative change of HTER, i.e., $(\text{HTER}_{norm} - \text{HTER}_{orig})/\text{HTER}_{orig}$, due to different normalization schemes. Negative change implies better performance. For a realistic scenario, we considered the normalization procedures trained with two partitions of data. The results are plotted as relative change of HTER versus KS-statistic as in Figure 7.5(b). As can be observed, F-norm performs almost always the best across different KS-statistics. When the KS-statistic is more than 0.4, Z-norm almost always degrades in performance (with respect to the original system).

[4]The XM2VTS has also been used and the results are somewhat consistent with the results reported here [99]. We will therefore not report the results carried out on XM2VTS here.



(a) Increasing sample size (b) Deviation from Gaussian

Figure 7.5: Comparison of the effects of different normalization techniques. The comparison is done with respect to (a) the sample size and (b) deviation from class-conditional Gaussian distribution of scores. For (b), larger KS-statistic implies larger deviation from Gaussian.

7.3.4 Improvement of Estimation of γ

As a final note, given the observation that β scales with the number of examples, it is possible to define a function which fulfills the following constraints:

- $\gamma = 0$ when the number of user-specific client accesses is zero, i.e., $N_j^C = 0$ (for client j).
- $\gamma = 0.5$ when $N_j^C = 1$.
- $0.5 < \gamma \leq 1$ when $N_j^C \geq 1$.

This function is:

$$\gamma = \frac{\left(N_j^C\right)^r}{\left(N_j^C\right)^r + 1} \tag{7.3}$$

where $r \geq 1$ parameterizes γ according to the available training data. This function is shown in Figure 7.6.

Note that (7.3) is somewhat similar to the "relevance factor" proposed in [122] having the form

$$\gamma^k = \frac{N_j^k}{N_j^k + r}$$

(also appeared in (6.16)) with N_j^k representing the number of user-specific accesses for any $k \in \{C, I\}$. Note that the role of relevance factor r in both cases are different in that (7.3) is exponential while the relevance factor of (6.16) is additive.

7.3.5 The Role of F-norm in Fusion

This Section examines the effectiveness of F-norm in minimizing the effect of user-variability in the context of fusion. For this purpose, we used the 15 XM2VTS face and speech fusion tasks described in Section 2.1.1. We randomly chose ten users from one of the 15 fusion tasks. The scores of each user as well as the class-conditional Gaussian fit (whose mean is represented by a plus sign and whose covariance is represented by an ellipse) are shown in Figure 7.7(a) prior to applying F-norm and in Figure 7.7(b) after applying F-norm. Since there are ten users and two classes, there are 20 ellipses in each figure. As can be observed, the user-specific impostor distributions are all centered at the origin whereas the user-specific client distributions are scattered very close to the point $(1, 1)$. This is expected due to the two F-norm's

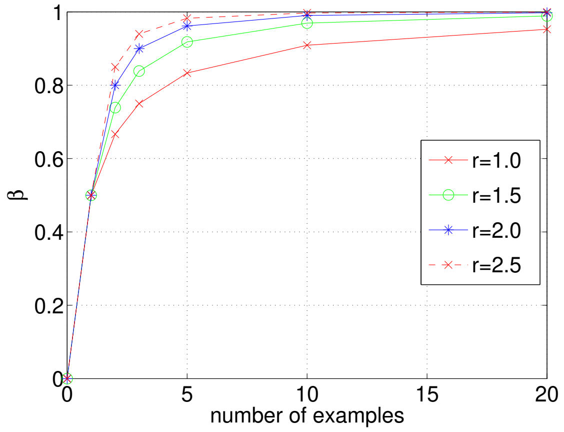


Figure 7.6: Parameterizing γ in F-norm with different relevance factor r's after taking into account the number of user-specific client accesses available.

properties: $\mu_j^{F,I} = 0$ and $\mu_j^{F,C} = 1$ where $\mu_j^{F,k} = E[y^F|k,j]$, i.e., the expectation of the user-specific class-conditional F-normalized scores. Note that the parameters of the F-norm were learned from the development score set and the figures shown here are plotted using the evaluation score set. For the example shown here, the F-norm's γ parameter was set to the default 0.5 so that the user-specific Gaussians cannot be perfectly aligned as in the impostor case. This choice is reasonable because μ_j^C cannot be estimated reliably due to too few user-specific genuine scores (two in this case). Forcing $\gamma = 1$ will result in overfitting.

We then used GMM to combine the 2D scores for both the data sets before and after applying F-norm. Their corresponding DET curves plotted using the evaluation score set are shown in Figure 7.7(a). In this case, we obtained a reduction *a posteriori* error from 0.57% EER to 0.25% EER, or a *relative* reduction of EER of 56%. Considering the already highly accurate systems, this error reduction is thus important. In order to ensure that this improvement is systematic, we compared the *a posteriori* EER before and after applying F-norm across all the 15 fusion tasks. These pair of EERs are plotted in Figure 7.7(b). As can be observed, the EER due to F-norm is systematically smaller than the EER prior to applying F-norm.

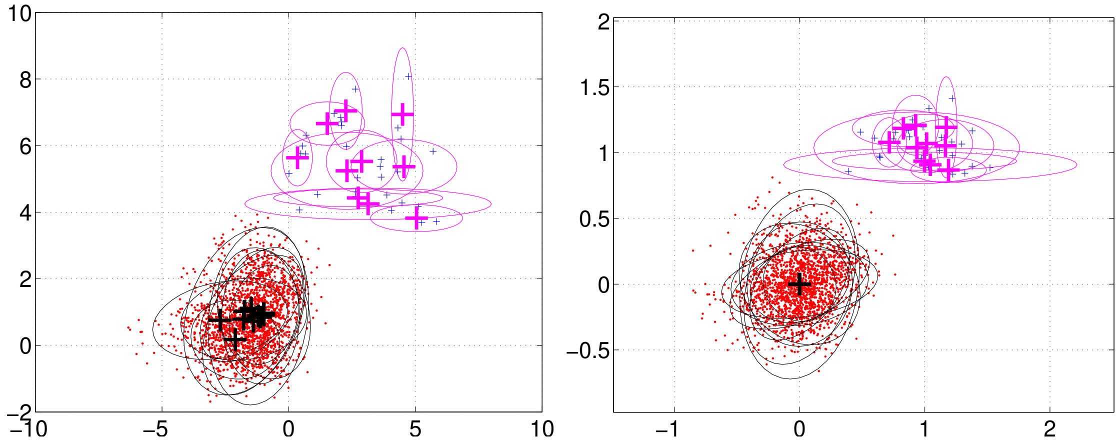
We then repeated the experiments but this time with *a priori* evaluation where the thresholds are optimized on the development set. The results depicted using the pooled DET curves calculated on the evaluation set are shown in Figure 7.9. The following observations can be made:

- Applying F-norm to the output of the speech systems can improve the baseline system (without normalization) significantly.

- Applying F-norm to the output of the face systems, on the contrary, does not improve the baseline system significantly.

- The combined systems due to F-norm is statistically significantly better than the baseline combined systems.

The degree of user-induced variability is obviously different for different biometric modalities. As a result, the effectiveness of F-norm is also different. In this case, the speech systems contain more variability than the face systems. Given that only the scores are available, and that the user-induced variability is an observed phenomenon, the reason why the face systems have lower user-induced variability is not exactly known. One possible reason is that, from the system point of view, face is much more homogeneous than speech. Measuring the degree of user-induced variability across different biometric modalities and systems will be a future subject of research.



(a) Before F-norm (b) After F-norm

Figure 7.7: An example of the effect of F-norm For both figures, the X- and Y-axes are the output score-space of a face and speech systems, respectively. The upper right clusters are client accesses whereas the lower left clusters are impostor accesses. In (a), before the score normalization the user-induced variability is high. In (b), after applying F-norm, the user-specific distributions are better aligned and separated as well.

7.4 In Search of a Robust User-Specific Criterion

Since the user-specific statistics are variable, the performance associated to each user must be different. The goal of this Section is to rank users given their associated user-specific statistics. To the best of our knowledge, this is the first study that attempts to rank users according to their performance. Having a criterion to rank users is useful in practical biometric applications. For example, Immediately after a new user has just been introduced to the system, it is important to know if the reference data (template) just registered is of reasonable quality. The quality in this case is taken as the estimated user-specific performance in terms of EER. If the EER is too high, remedial procedures can then be taken, e.g., acquiring more registration data to ensure a better modeling of the biometric features, using a different feature extraction algorithm or classifier, using different biometric traits, etc.

A good user-specific criterion should:

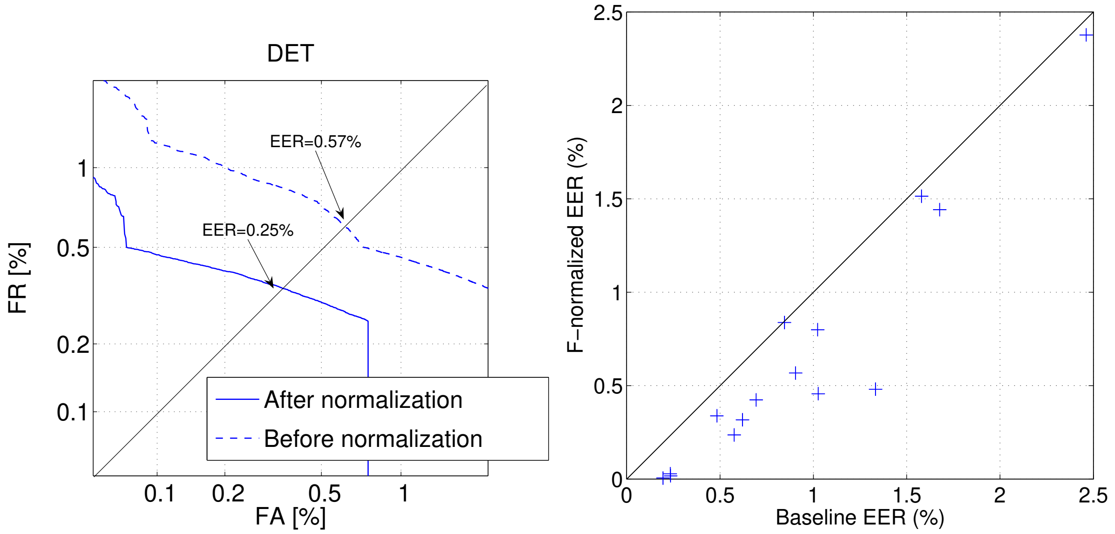
- Be robust to mismatch between the training and test data sets
- Be estimated based on as few samples as possible
- Necessarily contain the four (or less) user-specific statistics: $\mu_j^k, \sigma_j^k | k = \{C, I\}$ for each user j. From Section 7.2, we know that σ_j^C can be ignored since it is not informative.

Because the criterion must be related to performance, the user-specific F-ratio (from (4.15)) can be a good candidate, i.e.,

$$\text{F-ratio}_j = \frac{\mu_j^C - \mu_j^I}{\sigma_j^C + \sigma_j^I}. \tag{7.4}$$

Other similar measures are the d-prime statistic used in [28] and the two-class Fisher-ratio [11]. However, the user-specific F-ratio is preferred because it is functionally related to EER by (4.14) in a closed form.

Using the same datasets as those in Section 7.2, we compared the user-specific F-ratio of the 13 XM2VTS systems given the development set versus its evaluation set counterpart and the results are shown in Figure 7.10. In this case, 13 correlation values can be measured. As can be seen, using the original form as given, this quantity is very noisy and does not generalize well. Therefore, the user-specific F-ratio (similarly d-prime and two-class Fisher ratio) is not a good criterion because it is not robust.



(a) An example of improvement

(b) *A posteriori EER*

Figure 7.8: Improvement of class-separability due to applying F-norm prior to fusion. (a) An example of improvement due to F-norm visualized using a DET curve. (b) *A posteriori* EER of the baseline systems versus that due to F-norm for all the 15 fusion tasks.

Ideally, we would like to maximize the user-specific F-ratio. However, in this study, the user-specific model (which constitutes the baseline biometric classifier) has already been built and therefore its parameters cannot be modified. Our primary goal here is to make the user-specific F-ratio more *robust*, especially to mismatch between the training and test sets. One way to do so is by dropping the term σ_j^C since following the findings in Section 7.2, σ_j^C is not robust. The resultant *constrained* user-specific F-ratio thus becomes:

$$\text{F-ratio}_j = \frac{\mu_j^C - \mu_j^I}{\sigma_j^I}. \tag{7.5}$$

One important assumption when using (7.4) and (7.5) is that the optimal user-specific threshold is known. In this case, one implicitly assumes that the decision function as in (5.3) can be used. In practice, however, a user-independent threshold is more appropriate. In this case, the more practical decision function as appeared in (5.4) is used. The choice of user-specific score normalization procedure Ψ_j (where $N = 1$) can be F-norm or Z-norm. The advantage of applying user-specific score normalization prior to ranking the users is that the user-induced variability is effectively reduced even before the ranking takes place. The resultant F-ratio and its constrained counterpart for both the original, F-normalized and Z-normalized scores are summarized in Table 7.2. A figure similar to Figure 7.10 is not shown here for the rest of the five versions of user-specific F-ratios. However, without loss of generality, the goodness of prediction can still be objectively quantified by the following two measures:

- The correlation between $\text{F-ratio}_j|dev$ and $\text{F-ratio}_j|eva$ over all observed $j \in \mathcal{J}$
- The arithmetic difference between a given criterion estimated on a development set and its counterpart estimated on an evaluation set over all users $j \in \mathcal{J}$, i.e.,:

$$\text{bias} \equiv E_j[\text{F-ratio}_j|dev - \text{F-ratio}_j|eva].$$

Figure 7.11 summarizes the robustness of the original user-specific F-ratio and its five variants using two box-plots which correspond to the two measures just explained. As can be observed, the constrained F-norm ratio, i.e.,

$$\text{CFNR}_j = \frac{1}{\sigma_j^{F,I}}, \tag{7.6}$$

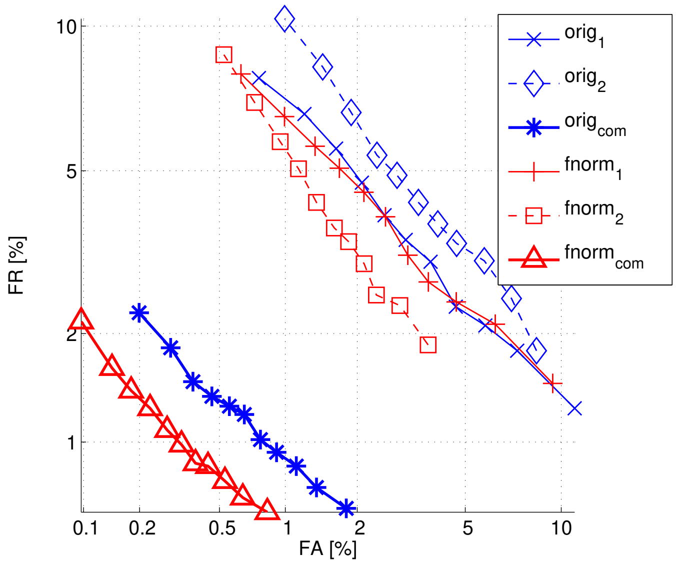


Figure 7.9: An empirical comparison of F-norm-based fusion the conventional fusion classifiers. The fusion performance. Each DET is pooled over the 15 fusion experiments. $orig_1$ contains face systems, $orig_2$ contains speech systems, $orig_{com}$ are combined $orig_1$ and $orig_2$ systems using GMM, $fnorm_1$ and $fnorm_2$ and the normalized $orig_1$ and $orig_2$ systems after applying F-norm; $fnorm_{COM}$ are combined $fnorm_1$ and $fnorm_2$ systems using GMM.

has the highest correlation while having an acceptable level of bias whose median is centered at zero.

Before concluding this section, we evaluated the goodness of the Constrained F-norm Ratio (CFNR) as shown in Table 7.2, i.e, by filtering away the N worst performing users where $N = \{200, 180, \ldots, 20\}$. The data sets used are the same 13 XM2VTS systems used in the previous sections. In order to ensure unbiased user ranks, the users were ranked according to the development set and this same user rank was applied to the evaluation set. The results of 8 of the 13 filtered system performances are shown in Figure 7.12. As can be observed, by removing the under-performing users, the system performance gradually improves. While the trend is more obvious in the *a posteriori* DET curves on which CFNR was calculated (see Figures 7.12(a–b)), this trend is somewhat reasonable on the evaluation set (see Figures 7.12(c–d)). The other five systems which were not shown behavior similarly.

Table 7.2: User-specific F-ratio and its constrained counterpart

Score normalization procedure	F-ratio	constrained F-ratio	Remarks
None	$\frac{\mu_j^C-\mu_j^I}{\sigma_j^C+\sigma_j^I}$	$\frac{\mu_j^C-\mu_j^I}{\sigma_j^I}$	σ_j^C is not robust
Z-norm	$\frac{\mu_j^{Z,C}}{\sigma_j^{Z,C}}$	$\mu_j^{Z,C}$	$\mu_j^{Z,I}=0$ and $\sigma_j^{Z,I}=1$
F-norm	$\frac{1}{\sigma_j^{F,C}+\sigma_j^{F,I}}$	$\frac{1}{\sigma_j^{F,I}}$	$\mu_j^{F,C}=1$ and $\mu_j^{F,I}=0$

Note: In the second and third rows, $\sigma_j^{Z,C}$ and $\sigma_j^{F,C}$ are omitted for computation in the corresponding *constrained* F-ratio because they are functionally dependent on σ_j^C which is not robust. The superscripts F and Z denotes statistics derived from F- and Z-norms, .e.g, $\sigma_j^{F,k} \equiv var[y_j^F|k]$ and $\mu_j^{F,k} \equiv E[y_j^F|k]$. The statistics for Z-norm is calculated in a similar manner.

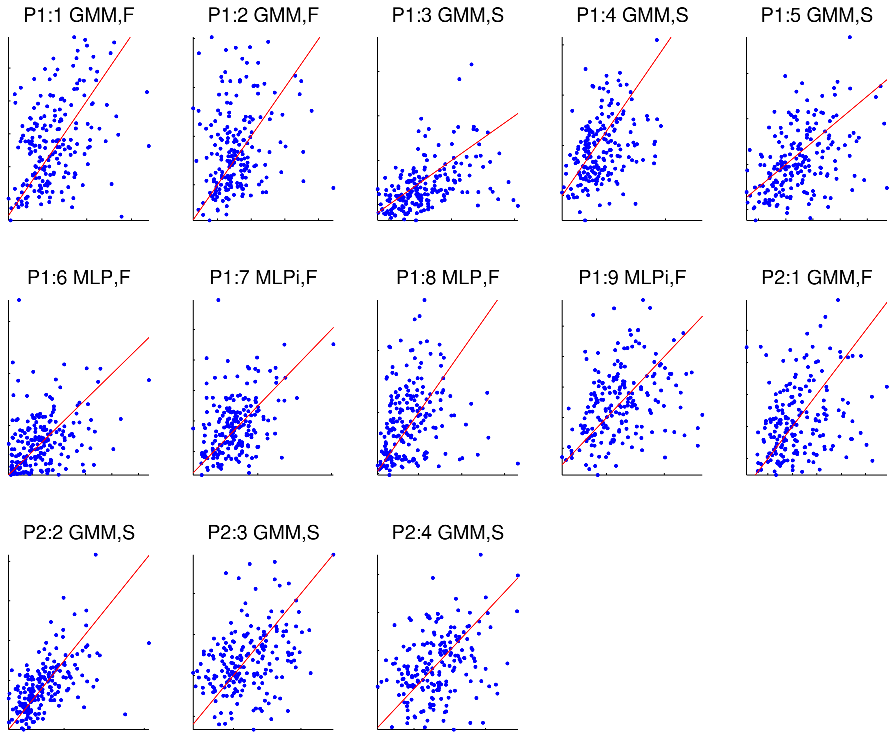


Figure 7.10: User-specific F-ratio as in (4.15) of development set versus that of evaluation set of the 13 face and speech based XM2VTS systems .

Discussions

User-ranking is a difficult problem for two reasons. Firstly, one is always lack of user-specific genuine data. Secondly, for this particular database, the simulated impostors are totally different from those used in the development set. This is a realistic scenario. We therefore conclude that user-ranking based on the proposed CFNR criterion is feasible, although there are definitely rooms for improvements. We will consider below some practical examples of how CFNR can be used:

- **As a diagnostic tool:** Immediately after a new user has just been introduced to the system, CFNR can be used to determine the quality of the reference data (template) just registered. To proceed, we can acquire one or two trial access requests. This gives us one or two genuine scores. The biometric samples of an arbitrary large set of simulated impostors can be used to generate some impostor scores. The CFNR criterion can then be evaluated given these two sets of scores. Two indications can be used to decide if the reference data is of poor quality. Firstly, the absolute CFNR is not high enough (say, by comparing to an *a priori* minimal CFNR value). Secondly, one can determine the rank of the newly registered user. When the CFNR value or the user's rank according to CFNR is too low, a warning will be issued. To the best of our knowledge, such a mechanism has not been previously proposed in the literature.

- **As a criterion for *selective* fusion:** While the existing combined system uses *all* systems by default, the CFNR criterion can be used to determine if fusion is indeed needed at all if the user is not among the worst performing users. For biometric applications where convenience are more important than security, having the option of not using all the biometric systems but tailored to a particular user's need can be important. Furthermore, in an application involving personal devices, the low-performing biometric sensor associated to a particular user does not need to be built into the device. Consequently, the hardware and software costs can be further reduced. Of course, it is expected that the systems may degrade in performance with respect to the case where all the available biometric

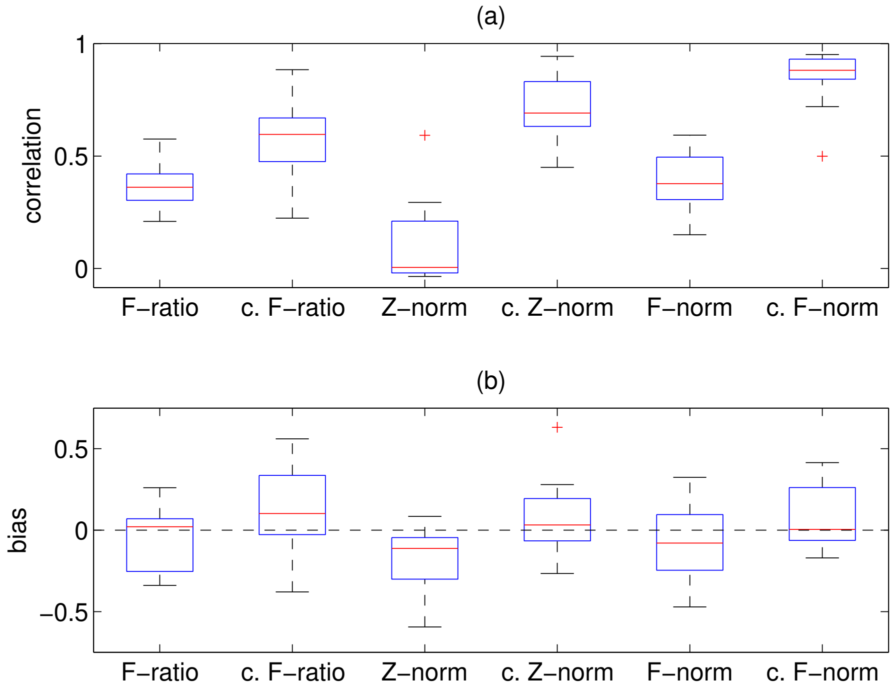


Figure 7.11: Comparison of the proposed six user-specific F-ratio as listed in Table 7.2, i.e., F-ratio, constrained F-ratio, Z-norm's F-ratio, constrained Z-norm's F-ratio, F-norm's F-ratio and constrained F-norm's F-ratio (or the constrained F-norm ratio) using (a) correlation and (b) bias between a given criterion of the development and that of the evaluation sets of the 13 XM2VTS face and speech systems. Each bar thus contains 13 (correlation or bias) statistics. Higher correlation and bias around zero are desirable properties. Note that the bias values of the constrained Z-norm's F-ratio (third column in (b)) were divided by 100 since they are originally in the range of $[10, -60]$.

systems are used. This subject of *selective fusion* will be investigated in Section 7.5.

7.5 A Novel OR-Switcher

7.5.1 Motivation

As far as fusion in the context of biometric authentication is concerned, the usual approach is to combine *all* the available system outputs. While this is certainly easier to design, all the participating biometric systems have to be operational. Despite the fact that the system is designed with the redundancy of having multiple biometric systems (devices), the verification cannot proceed if one of the sub-systems (devices) fail. For this reason, we investigate the possibility of *selective* fusion, where a multimodal (and multi-algorithmic) system will be capable of giving an output score even when one of the sub-systems fails or determines that its acquired sample is unreliable. This selective fusion strategy in a way mimics biological perception in the nature. For instance, human is capable of recognizing a person by just having a partial evidence, e.g., speech, gait or occluded face. Very often, only salient features are needed. One prominent example is human caricature[5]. Our preliminary findings here suggest that the user-specific and selective fusion strategy can indeed be better than the state-of-the-art fusion techniques to some extent.

[5]Test your skill at http://www.magixl.com

Our Proposal

The fusion operator to be proposed here is different from the state-of-the-art fusion techniques in two aspects:

- **User-specific**: it must take the user specific performance into consideration. The CFNR criterion can readily be used for this purpose because from the previous experiments, it has been shown to be robust and can be computed using only a few user-specific genuine samples.

- **Selective:** It must be able to handle "missing values", where some underlying biometric systems cannot output scores. If the classifier is based on LLR, for instance using GMM to estimate the class-conditional score distribution, handling missing values becomes integrating the distribution with respect to the missing values. This subject will be further discussed in Section 7.5.4.

We call the novel fusion operator the "OR-switcher". To the best of our knowledge, because of the two properties just mentioned, the OR-switcher is a unique fusion operator.

Section Organization

Note that while CFNR can indicate a user's performance, it does not indicate which combination of system subset will give a theoretically optimal fusion performance. This subject will be dealt with in Section 7.5.2. Section 7.5.3 then gives an overview of the OR-switcher. Section 7.5.4 deals with the problem of conciliating the output due to missing scores. Section 7.5.5 proposes two metrics to evaluate the OR-switcher. These metrics do not deal with the generalization performance but with the adequecy of the choice of the system subset and computational saving. Finally, Section 7.5.6 compares the performance of the OR-switcher with two other baseline classifiers

7.5.2 Extension to the Constrained F-norm Ratio Criterion

This section aims to extend CFNR$_j$ to take into account the performance due to a system subset p, i.e., CFNR$_{j,p}$. If there are 3 systems (hence $N = 3$), p will be one of the possible power set of $\{1, 2, 3\}$, excluding the empty set. In our notation, we write:

$$p \in \mathcal{P}(\{1,2,3\}) - \emptyset \equiv \big\{\{1\},\{2\},\{3\},\{1,2\},\{1,3\},\{2,3\},\{1,2,3\}\big\}.$$

We also denote the default fusion mode that uses all the systems as $com \equiv \{1, 2, 3\}$.

In order to calculate CFNR$_{j,p}$, we first need to prepare the combined score set due to using the system subset p, i.e.,$\{y_p^F|j\}$. A good candidate to use is the mean operator:

$$y_{j,p}^F = \text{mean}_{i \in p}\, y_{i,j}^F. \tag{7.7}$$

Since $y_{i,j}^F$ can be interpreted as an LLR, taking the sum (or mean in this case) corresponds to making the independence assumption of the system outputs $i \in p$. Using the labeled development set $\{y_p^F|j,k\}$ for $k \in \{C, I\}$, we can effectively assess CFNR$_{j,p}$ as in (7.6).

7.5.3 An Overview of the OR-Switcher

We will consider here the case of combining two biometric systems. The extension to N systems is straightforward. We will discuss here an overview of the proposed strategy. It should be noted that there are two data sets: development and evaluation sets. The development set is served to derive all the training parameters, e.g., F-norm's parameters, the user-specific CFNR criterion and the optimal decision threshold. The evaluation set is served uniquely as a test set.

1. Apply F-norm to each participating biometric system independently. Note that the F-norm parameters must be derived from the development set.

2. Train a GMM fusion classifier of the form $y^F_{com} = \log \frac{p(\mathbf{y}^F|C)}{p(\mathbf{y}^F|I)}$ by estimating the class-conditional score distribution $p(\mathbf{y}^F|k)$ for each $k = \{C, I\}$ separately (see Section 3.4.3).
3. For each user $j \in \mathcal{J}$ and each possible subset combination p, assess the CFNR$_{j,p}$ criterion given the labeled combined scores $\{y^F_p|k, j\}$ based on the development set.
4. Sort the users in descending order based on CFNR$_{com}$ (the default mode where all the systems are considered). For the $r \times 100\%$ top portion of users, we determine that fusion is not necessary. In this case, we decide the next best alternative of system subset p. In the case of $N = 2$ systems, $p \in \big\{\{1, \}, \{2\}\big\}$, we choose the better of the two systems, i.e.,
$$p^*_j = \arg\max_p \text{CFNR}_{j,p}.$$
5. During the operational phase, the combined LLR score is calculated as $y_{OR} = \log \frac{p(\mathbf{y}^F|C,p^*_j)}{p(\mathbf{y}^F|I,p^*_j)}$ where $p(\mathbf{y}^F|k, p^*_j)$ is a marginalized distribution of $p(\mathbf{y}^F|k)$ with respect to the systems *not in* p.

There are two points to note regarding the strategy presented here. Firstly, the fusion classifier of the form $y^F_{com} = \log \frac{p(\mathbf{y}^F|C)}{p(\mathbf{y}^F|I)}$ is the F-norm based classifier presented in Section 7.3.5 (denoted as $\mathsf{fnorm_{com}}$). In this case, steps 3 and 4 can be omitted and in step 5, p^*_j is replaced by the default p^{com}_j (which uses all the systems). By setting the fraction $r = 0$, the OR-switcher converges to $\mathsf{fnorm_{com}}$. We expect that when r increases, the performance will degrade since less and less information is considered. In other words, the OR-switcher will be inferior to $\mathsf{fnorm_{com}}$. However, the question we are interested in is, to what extent r can take such that the performance of the OR-switcher is as good as the standard fusion classifier based on GMM, i.e., $y_{com} = \log \frac{p(\mathbf{y}|C)}{p(\mathbf{y}|I)}$. In our experience, other standard fusion classifiers, e.g., SVM and logistic regression, give similar results [101]. This is expected since they all rely on the same training data and none exploit special knowledge, e.g., the user-specific information.

Secondly, there is an elegant way to convert from the default likelihood $p(\mathbf{y}^F|k)$ – where all the systems are used – to $p(\mathbf{y}^F|k, p^*_j)$ – where only the system subset p^*_j is used when $p(\mathbf{y}^F|k)$ is approximated using a mixture of Gaussian components. This is discussed in Section 7.5.4.

7.5.4 Conciliating Different Modes of Fusion

Let $\mathbf{y}^{F,k} = [y^{F,k}_1, \ldots, y^{F,k}_N]'$ be a vector of the class-conditional scores to be combined *after* applying F-norm. Let us approximate the joint conditional distribution of $\mathbf{y}^{F,k}$, $p(\mathbf{y}^{F,k})$ by a mixture of Gaussian components of the form:
$$p(\mathbf{y}^{F,k}) = \sum_{c=1}^{N_c} w_c \mathcal{N}\big(\mathbf{y}|\boldsymbol{\mu}^{F,k}_c, \boldsymbol{\Sigma}^{F,k}_c\big), \tag{7.8}$$
where w_c is the prior of the c-th Gaussian component whose parameters are $\boldsymbol{\mu}^{F,k}_c$ and $\boldsymbol{\Sigma}^{F,k}_c$, for $k = \{C, I\}$. Note that this classification is user-independent but receives input from user-specific normalized scores obtained via F-norm.

Given the *joint* distribution described by the mixture of Gaussian parameters $\{w_c, \boldsymbol{\mu}^{F,k}_c, \boldsymbol{\Sigma}^{F,k}_c|\forall_c\}$, our goal is to find the marginal distribution spanned only by the subset (or subspace) $p \subseteq \{1, \ldots, N\}$. One way is to marginalize the conditional joint distribution $p(\mathbf{y}^{F,k})$ with respect to the output of the systems not considered. Using a mixture of Gaussian parameters, this can be done in a rather straight-forward manner. First, let us drop the parameters F, k and c from $\boldsymbol{\mu}^{F,k}_c, \boldsymbol{\Sigma}^{F,k}_c$ since the discussion that follows will always be dealing with $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in the F-norm domain, applying to each k and each c Gaussian component individually. Then, the marginalized parameters due to using the subset p can be written as $\boldsymbol{\mu}_p$ and $\boldsymbol{\Sigma}_p$. The matrices before and after marginalization are related by:
$$\boldsymbol{\mu} = [\boldsymbol{\mu}_p, \boldsymbol{\mu}_{\bar{p}}]'$$
$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_p & \boldsymbol{\Sigma}_q \\ \boldsymbol{\Sigma}'_q & \boldsymbol{\Sigma}_r \end{bmatrix}$$

where $\boldsymbol{\mu}_{\bar{p}}$ is the mean vector whose elements are systems *not* in the set p and $\boldsymbol{\Sigma}_t|t \in \{q, r\}$ are the rest of the sub-covariance matrices which contains the elements not in p. The above marginalization procedure for GMM can be found in [87], for instance, and is used for noisy band-limited speech recognition. Let us take an example of $N = 3$ systems. Suppose the optimal subset is $p = \{1, 2\}$ and the excluded system set is $\bar{p} = \{3\}$. Consequently,

$$\boldsymbol{\mu}_p = [\mu_1, \mu_2]', \boldsymbol{\mu}_{\bar{p}} = [\mu_3]', \boldsymbol{\Sigma}_p = \begin{bmatrix} e_{1,1} & e_{1,2} \\ e_{2,1} & e_{2,2} \end{bmatrix}, \boldsymbol{\Sigma}_q = \begin{bmatrix} e_{1,3} \\ e_{2,3} \end{bmatrix}, \boldsymbol{\Sigma}_r = [e_{3,3}],$$

where $e_{m,n}$ is the m-th row and n-th column element of the covariance matrix $\boldsymbol{\Sigma}$ and $e_{m,n} = e_{n,m}$ (since a covariance matrix is reflexive).

7.5.5 Evaluating the Quality of Selective Fusion

Two types of evaluation are considered here, i.e., by agreement and by computational saving.

Evaluation by Agreement

Note that p_j^* contains the subset of systems that are considered optimal, in the F-norm domain, for a user j according to the *development* set. One could equally evaluate the same parameter for the *evaluation* set. A useful way to evaluate if $p_j^*|dev$ is optimal or not is by comparing the same parameter derived from the evaluation set $p_j^*|eva$ – which is considered the ground truth. Let $I(m, n)$ be an indicator function that outputs 1 if the sets m and n are identical and zero otherwise. The probability of choosing the "right" mode of fusion, within the population of users considered, in the OR-switcher, can be defined as:

$$d = \frac{\sum_j I\big(p_j^*|dev, p_j^*|eva\big)}{J}$$

Higher d is thus clearly desired.

Evaluation by Computational Saving

One can also evaluate the computational savings due to not using some of the biometric systems. It can be quantified by:

$$\text{computational saving} = 1 - \frac{\sum_{j \in \mathcal{J}} \sum_{i=1}^{N} I(\text{system}_{i,j})}{2 \times J},$$

where $I(\text{system}_{i,j})$ is an indicator function that gives 1 if the i-th biometric system of user j is used and zero otherwise and there are J users. In the case of a conventional fusion classifier where all the systems are used, the computational saving is simply zero. In our case, when two systems are considered using the OR-switcher, the fraction r as presented in Section 7.5.3 is directly related to the computational saving in the following way:

$$\text{computational saving} = (1 - r)/2 \times 100\%.$$

7.5.6 Experimental Validation

Fusion Experiments

We set up a fusion protocol in the following ways: (i) for LP1, we combined exhaustively the face systems { P1:1, P1:2, P1:7, P1:9 } with the speech systems { P1:3, P1:4, P1:5 }; (ii) for LP2, we combined exhaustively the face system { P2:1 } with the speech systems { P2:2, P2:3, P2:4 }. LP1 (resp. LP2) has 12 (resp. 3) multimodal fusion tasks.

Using our proposed criterion, the percentage of correctness d is measured to be 88.5% with minimum and maximum being 80% and 97.5%, respectively, across all 15 fusion tasks.

We then compared the OR-switcher with two baseline systems, as follows:

Table 7.3: Comparison of the OR-switcher and the conventional fusion classifier using *a posteriori* EER evaluated on the evaluation set of 15 face and speech XM2VTS fusion benchmark database.

No.	*a posteriori* EER on the eva. set (%)					
	OR-Switcher's r values					baseline
	0.6	0.7	0.8	0.9	1.0	
1	0.87	**0.57**	**0.46**	**0.34**	**0.32**	0.62
2	2.07	2.00	1.87	1.73	**1.51**	1.58
3	1.36	**0.82**	**0.72**	**0.52**	**0.48**	1.33
4	**0.46**	**0.39**	**0.34**	**0.29**	**0.24**	0.58
5	**0.96**	**1.00**	**0.94**	**0.88**	**0.80**	1.02
6	**0.74**	**0.72**	**0.69**	**0.63**	**0.57**	0.91
7	1.15	0.93	0.79	0.64	**0.34**	0.48
8	1.46	1.49	1.39	1.11	**0.84**	0.85
9	1.33	**1.02**	**0.78**	**0.73**	**0.46**	1.03
10	1.64	1.39	1.05	0.83	**0.42**	0.69
11	4.16	4.08	3.74	2.94	**2.38**	2.46
12	3.40	3.14	2.59	2.02	**1.44**	1.68
13	0.43	0.39	0.35	**0.05**	**0.01**	0.19
14	0.50	0.47	0.28	**0.03**	**0.03**	0.23
15	**0.21**	**0.19**	**0.05**	**0.03**	**0.02**	0.24

Note: The EER values in bold indicate that the respective OR-switcher has an EER lower than that of the baseline classifier. The data in the last two columns were plotted in Figure 7.8(b). When $r = 1$, the OR-switcher is equivalent to combining F-normalized scores. All the classifiers evaluated here are GMM classifiers. The DET curves of experiments 15, 3 14 and 10 (in this order) are shown in Figures 7.13(a)–(d).

- **The *de facto* fusion classifier based on GMM:** In this case, the scores $y_i \forall_i$ are used. [6]
- **The user-specific GMM based on F-normalized scores:** In this case, the GMM classifier was trained with F-normalized scores, i.e., $y_i^F \forall_i$

The OR-switcher behaves different for a given set of the fraction values $r = \{0.6, 0.7, 0.8, 0.9\}$. The system performances are plotted using only DET curves and are shown in Figures 7.13. Since we could not plot all the DET curves which behave very differently from each other, we listed the *a posteriori* EER performance evaluation in Table 7.3. We can identified four types of experimental outcomes:

- **Ideal:** where no lost is observed at $r = 0.6$.
- **Potential:** where no lost is observed at $r = 0.7$
- **Satisfactory:** where no lost is observed at $r = 0.9$
- **No gain:** where no lost is observed at $r = 1.0$

According to this categorization, at EER, 4 systems are considered ideal, 3 are potential, 2 are satisfactory and 5 has no gain. The DET curves of an example in each category is shown in Figure 7.13.

Discussion

The experimental outcomes suggest that it is still possible to make decisions based on *incomplete* information. The proposed OR-switcher is really a proof of this concept. While having less information (depending

[6] From our previous study [101], the GMM fusion classifier performs as well as the logistic regression and Support Vector Machines with a linear kernel. Since all these classifiers rely on the same training sets with carefully tuned hyper-parameters, their generalization performances cannot be *significantly* different.

on the pruning rate r), the OR-switcher is at least as good as the conventional fusion classifier, if not better. However, by using lower r (higher pruning), the system is expected to degrade steadily in accuracy. However, at least, the OR-switcher does not fail completely as would the conventional fusion classifier because the OR-switcher can capitalize on the inherent system redundancy. Furthermore, one of its advantage over the conventional fusion classifier is that the OR-switcher makes use of the user-specific information.

7.6 Summary of Contributions

This Chapter contains the following novelties:

- **Empirical investigation of the robustness of user-specific statistics:** Although the user-specific statistics, i.e., μ_j^k and σ_j^k, have been used in user-specific score normalization and threshold procedures (Chapter 5), no systematic study has been made regarding the robustness (the ability to generalize to unseen data) of the mentioned statistics. Our experiments in Section 7.2 show that σ_j^C is not robust and hence should not be considered. This has significant influence on the design of user-specific procedures. This Section appears in our published paper [115].

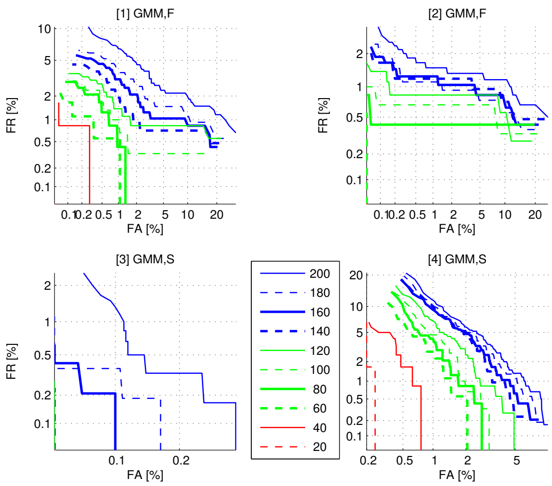
- **User-specific score normalization based on F-ratio (F-norm):** Our study in Section 7.3 shows that F-norm belongs to a new family of user-specific score normalization besides Z-norm and EER-norm. Our empirical and theoretical analysis show that in comparison to Z- and EER-norms, F-norm has the following advantages:

  - F-norm is more robust to deviation from Gaussian since it does not use the second-order user-specific statistics.
  - F-norm can work with fewer training samples since it does not use the second-order user-specific statistics and it relies on Bayesian adaptation.
  - Empirically, its generalization performance increases faster in proportion to the number of genuine samples since it is client-impostor centric.

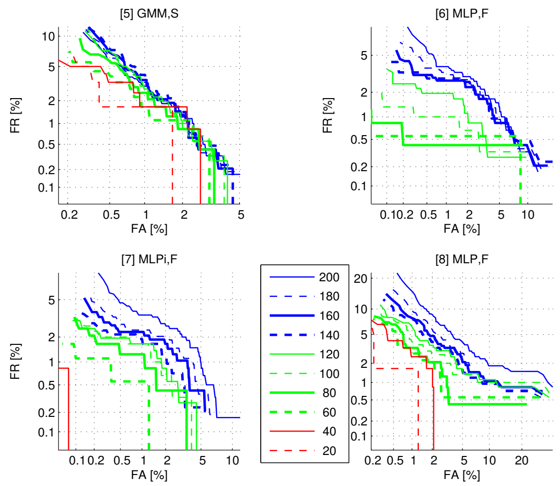
  This Section appears in our published paper [108].

- **Criterion to rank users:** Although Doddington *et al* [33] were the first to develop techniques to categorize different types of users in a biometric database according to their score statistics, they did not provide a technique to rank users according to their ease of recognition. Furthermore, the statistical techniques developed by Doddington *et al* were not designed with statistical robustness as a primary concern. In Section 7.4, we found out that such a criterion is best evaluated using a constrained F-ratio with scores transformed into F-norm. This criterion is called Constrained F-norm Ratio (CFNR). Due to working in the F-norm domain, user-induced variability is effectively reduced before the ranking takes place. This is an advantage because this variability can adversely affect the user ranking. Again, CFNR is designed with maximal robustness and this property was verified using 13 face and speech biometric systems on XM2VTS. This Section appears in our published paper [115].

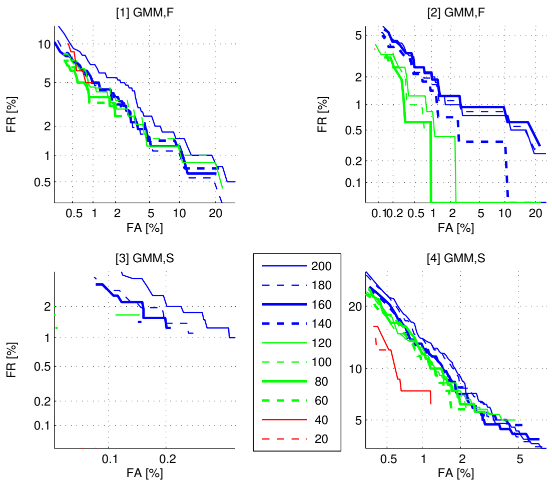
- **User-specific fusion via the OR-switcher:** The ability to rank users based on CFNR has a practical application in the context of multimodal biometric fusion. We illustrated the usefulness of CFNR to selectively combine systems on a per user basis. We called this novel fusion operator the OR-switcher. The performance of the OR-switcher is as good as the fusion system that combines *all* the system outputs with user-specific F-normalized scores. However, because the OR-switcher does not use all the biometric systems, it can reduce the computational cost. For instance, in our experimental setting with 15 fusion tasks *on average*, the OR-switcher can reduce the computational resources up to a quarter of that with a conventional fusion classifier (that uses all the sub-systems). This is achieved *without* significant reduction in performance with respect to the one with full-fledged systems which is also based on F-norm. We also compared the performance of the OR-switcher with the state-of-the-art technique which uses trainable user-independent fusion classifiers. We used GMM in this case but SVM gave also similar performance as reported in [101]. Since the OR-switcher exploits the user-specific information, its performance is *statistically significantly* better than the state-of-the-art fusion classifiers; and this is achieved by reducing the overall software/hardware resources. This advantage becomes more apparent for multimodal authentication using a personalized device because an under-performing biometric hardware with respect to a given user can be removed from the device. This Section appears in [112] and is under peer-review.



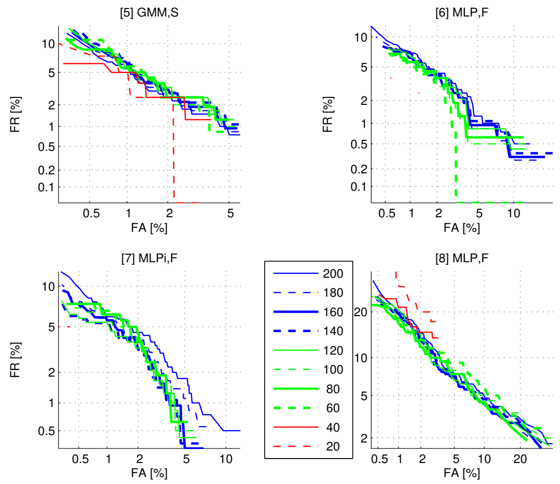
(a) Systems 1–4 (dev)



(b) Systems 5–8 (dev)

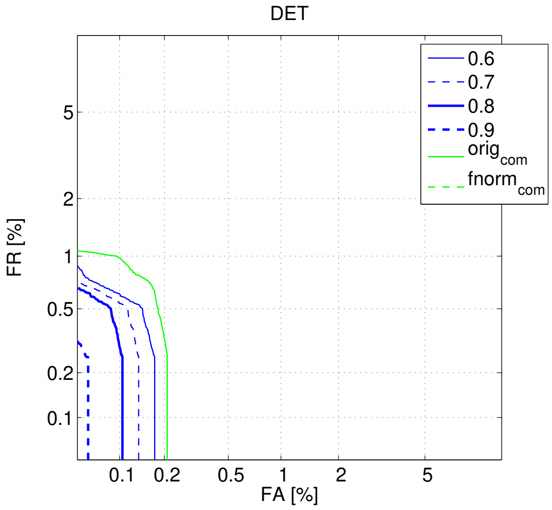


(c) Systems 1–4 (eva)

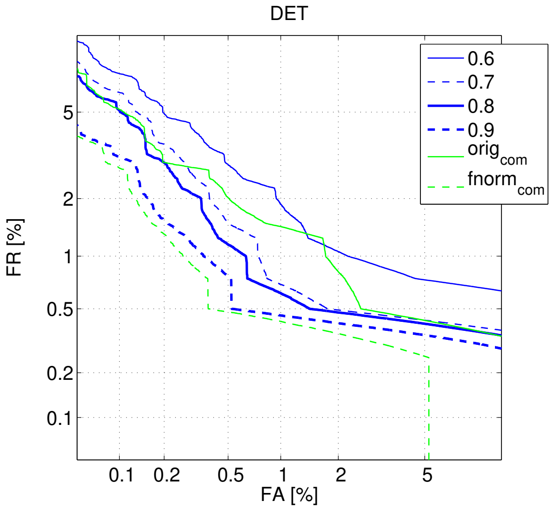


(d) System 5–8 (eva)

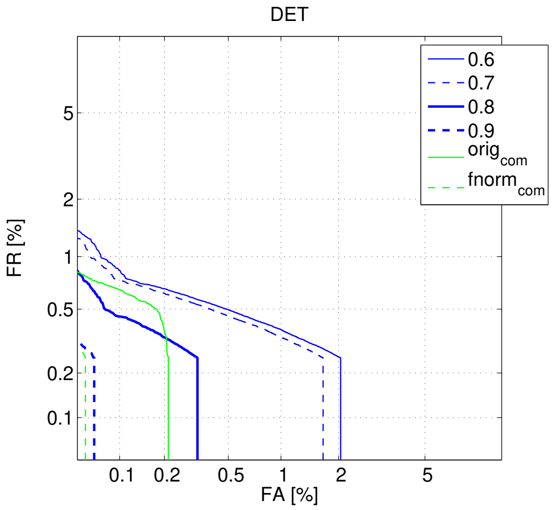
Figure 7.12: Results of filtering away under-performing users for each of the first 8 XM2VTS systems shown using DET curves. The users were ranked according to the constrained F-norm ratio (CFNR, or $(\sigma_j^{F,k})^{-1}$) based on the data of the development set. The $N \in \{200, 180, \ldots, 20\}$ lowest performing users are filtered at each stage. Figures (a) and (b) show the *a posteriori* filtered DET curves of the development score set on which CFNR was calculated and Figures (c) and (d) show the *a priori* filtered DET curves evaluated on the *evaluation* score set. Some DET curves cannot be plotted because no error was observed.



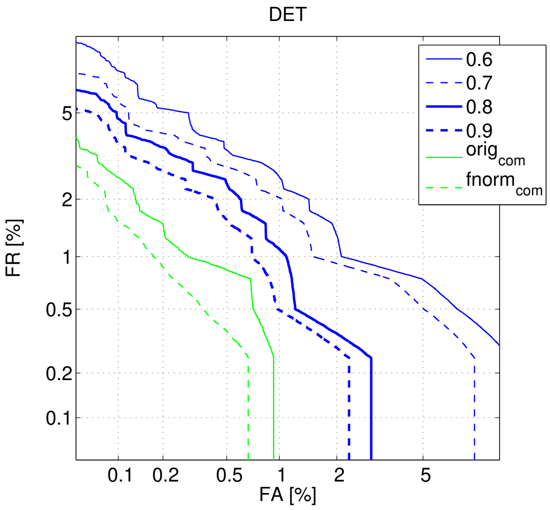
(a) Ideal situation



(b) Potential



(c) Satisfactory



(d) No gain

Figure 7.13: An empirical comparison of user-specific classifier, OR-switcher and the conventional fusion classifier. The fusion performance depicted by DET curves. An example of each of the four types of experimental outcomes were observed: (a) ideal, where the OR-switcher achieves 20% computational savings (whose cutting rate is 0.6) without remarkable lost of performance compared to the baseline system (4 fusion experiments in this category); (b) potential, where 15% computational savings (cutting rate = 0.7 was achieved (3 experiments); (c) satisfactory, where 10% computational savings (cutting rate = 0.8) was achieved; and (d) no gain, where 5% computational savings (cutting rate = 0.9) was achieved (6 experiments).

Chapter 8

Conclusions and Future Work

8.1 Conclusions

Benefits of Using LLR in Fusion

In the literature, fusion is dominated by techniques that convert participating system outputs to probability prior to combining them using simple fixed rules [66]. Score conversion is important because different systems output different score types. We proposed a unifying framework that converts different score types to probability or LLR (see Section 3.3.2). Deviating from the mainstream, we showed that converting scores into the LLR space is *more useful* than into probability because the underlying statistics can better be described using the first- and second-order moments. Thanks to this advantage, we could:

- Analyze fusion via a parametric fusion model (Part I)
- Better exploit the user-specific information (Part II).

These two parts are closely related in that the parametric fusion model can be extended to user-specific processing by conditioning the model to a particular user, i.e., using user-specific statistics, $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ instead of user-independent $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

Parametric Fusion Model

With a parametric fusion model (Chapter 3), we could:

- Explain why fusion works
- Predict fusion performance
- Identify conditions which favor fusion with a particular fusion operator
- Study the joint phenomena of combining classifiers with different degree of strength and correlation
- Reduce the adverse effect of bias (or score-level mismatch between training and test sets) on fusion

An interesting statistic from the proposed parametric fusion model is called the F-ratio. It characterizes the separability between the genuine scores and the impostor scores. Although relying on class-conditional score distribution, we showed that the F-ratio (as well as other related performance measures such as FAR, FRR and EER) is robust to deviation from the assumption in the context of classification (see Sections C.1–C.2).

An application of performance prediction using the F-ratio is to select an optimal subset of (possibly correlated) systems to combine (see Section 4.5). In this context, one is ready to trade-off insignificant performance gain with less computation. F-ratio is more useful than the empirically calculated EER because F-ratio is more robust to different population composition. Furthermore, the system selection using F-ratio has a complexity that is independent of the data available since only the F-ratio criterion has to be evaluated for each possible combination.

User-Specific Processing

There are three original contributions to the state-of-the-art in user-specific processing.

1. **Survey on user-specific processing:** We analyzed some of the desired characteristics from existing literature in order to exploit user-specific processing. In particular, we generalized user-specific fusion and user-specific score-normalization techniques in the following form for making the accept/reject decision:
$$\Psi_j(\mathbf{y}) > \Delta$$
for $\mathbf{y} = [y_1, \ldots, y_N]'$. When $N > 1$, Ψ_j is a user-specific fusion classifier. when $N = 1$, Ψ_j is a user-specific score normalization procedure. We also showed that the user-specific threshold technique is a special case of the user-specific score normalization technique. The survey was reported in Chapter 5.

2. **A compensation scheme:** In Chapter 6, we proposed the following alternative framework for decision making:
$$\gamma\Psi_j(\mathbf{y}) + (1-\gamma)\Psi(\mathbf{y}) > \Delta,$$
where Ψ is a user-independent function (fusion classifier or score-normalization procedure) and γ adjusts the contribution between the user-specific and user-independent functions. This form has the following benefits:
   - **Mutual compensation:** The solution compensates for the unlikely robust user-specific classifier but at the same time, enhances the user-independent classifier with a user-specific one.
   - **Hybrid learning algorithms:** Both classifiers can be trained independently of each other.
   - **Independence of information:** Following the justification in Section 6.1, both classifiers are likely to produce independent outputs when the number of users is large.

   The compensation framework compares favorably with [40, 41, 139, 71, 61] principally because it is the only one that can learn from very few user-specific genuine samples, which is a non-trivial machine-learning problem.

3. **User-Specific F-ratio based techniques:** We extended the system level F-ratio used in the parametric fusion model to the user-specific F-ratio (Chapter 7). The usefulness of the user-specific F-ratio is shown in the following applications:
   - **F-norm:** F-norm is a user-specific score normalization technique that aims to reduce the user-induced variability. F-norm is superior to existing normalization techniques, e.g., Z-norm, EER-norm and their variants, due to its following properties:
     - Robustness to the Gaussian assumption
     - Robustness to extremely few genuine training samples thanks to Bayesian adaptation – an advantage not shared by existing methods in user-specific score/threshold normalization, e.g. [18, 48, 52, 64, 75, 92, 126]
     - Client-impostor centric – making use of both the genuine and impostor scores
   - **Criterion to rank users:** Although the user-induced variability has been studied [33], there exists no criterion that ranks users according to their ease of recognition. Such a criterion is important to decide the usability and suitability of a biometric system on a per person basis. We proposed a criterion based on F-ratio, called constrained F-norm ratio (CFNR), which is *robust* (able to generalize to unseen data), is *unbiased* to mismatch between training and test sets and can be *reliably estimated* from few samples.
   - **The OR-switcher:** The OR-switcher is a user-specific selective fusion whereby only a subset of systems are used. It strongly relies on the CFNR criterion after taking into account all possible combinations of system subsets. The OR-switcher is better than the conventional fusion classifiers proposed in the literature because it makes the resultant multimodal system faster (less processing), cheaper (less hardware component in applications with personal devices) and better (more accurate by exploiting user-specific information).

Other Contributions in User-Specific Processing

We summarize here the results of two related topics which are original contributions but could not fit exactly in the two major themes chosen in this thesis.

1. **A discriminative framework to combine user-specific and quality information:** While studies have been conducted on incorporating user-specific and quality information *separately*, we considered fusion of these two information sources *simultaneously*. The discriminative framework is useful for two reasons:

   - **Ease of implementation:** The framework can be implemented using any existing discriminative classifier whose properties are well studied rather than using specialized classifiers for this purpose, e.g., support vector machines, multi-layer Perceptrons, etc, linear or non-linear.
   - **Ease of integration with user-specific information:** User-specific information can be integrated into the framework by means of *any* user-specific score normalization whose effectiveness can be evaluated independently from the framework.

   We showed that combining both information sources is better than using either one, or using none of them. This paper was published in [111] and was the winner of the best student poster award in the 5th Int'l. Conf. Audio- and Video-Based Biometric Person Authentication (AVBPA 2005) for contribution on "biometric fusion".

2. **User-specific performance trend analysis:** The goal of this study was to assess whether or not the performance of *individual users* as well as that of the *overall* system changes when a biometric authentication system is operational on a regular basis. While a pilot study in [46] attempted to assess the overall system performance, there was no study that makes the assessment at the individual level. The trend is useful to decide when a user's template or model should be updated. We proposed to model the user-specific trend in two steps. Firstly, one models the user-specific client and impostor sequences of scores over time using a regression algorithm. The output of regression is a series of *time-dependent user-specific statistics* in terms of mean and variance, i.e., $\mu^k_{j,t}$ and $\sigma^k_{j,t}$ over time index t for a given user j and class $k = \{C, I\}$. By assuming the class-conditional Gaussian assumption, the instantaneous user-specific performance (e.g., user-specific F-ratio, EER) can be traced. The conventional approach uses a sliding window, which defines the set of scores inside a limited period, to calculate a time dependent performance [147].

   There are two disadvantages with the conventional approach compared to our proposed one:

   - **The trade-off between time precision and reliability of performance estimate:** A large window reduces the time precision but increases the reliability of performance estimate whereas a small window increases the time precision but decreases the reliability of the performance estimate.
   - **Limitation to user-independence analysis:** The sliding window approach cannot be used to estimate the user-specific trend because user-specific genuine scores are extremely limited.

   Because of this trade-off, deciding on the window size is also a difficult problem. Our proposed algorithm uses standard regression tools whose parameters can be tuned elegantly. Furthermore, the model can estimate the trend to an *arbitrarily smoothed precision*.

The devised algorithm to estimate both the user-specific and user-independent (system level) trend estimation is an important proof of concept that user-specific processing is extremely powerful in biometric authentication as well as identification. Our finding suggests that only a quarter of users degrade significantly in performance over time. Furthermore, the initial template, and not the user identity, is responsible for the trend. This study can be found in [96].

To the best our our knowledge, at the time of writing, this thesis represents the state-of-the-art of *user-specific processing* in biometric authentication.

8.2 Future Work

This Section provides a non-exhaustive list of future work related to biometric person authentication. Some of these issues were encountered during the research for this thesis but could not be fully addressed.

- **Composite DET/EPC curve.** Visualizing a composite DET/EPC curve becomes a necessity for algorithmic evaluation when several data sets are involved. This is done by establishing a global coordinate among different DET curves. To the best of our knowledge, three types of coordinate exist, namely, DET angle [2], LLR unique to each DET [54] and the α value used in the WER criterion (see (2.5)). The merits of each approach should be examined.

- **User-specific processing at feature or score level.** Chapters 6 and 7 show that user-specific processing at the score-level can improve the system performance. This suggests that the processing at the score-level can be potentially extended to the feature level. While the information is richer at the feature level, the dimensionality is also much higher. Overcoming this possible drawback is thus very challenging but if successful, *significant improvement* could be obtained.

- **Template-updating.** When a biometric system is operational, the user-specific performance changes over time. If the performance degrades, then the algorithm has to update the underlying template/model. There are two important questions to answer: (i) *when* and (ii) *how* the update should proceed. For a completely automatic system, this can be considered a semi-supervised learning. There are certainly many more issues to examine, for instance, what if the wrong template is updated and how the remedial procedure should proceed.

- **Mismatch due to different sensors.** When a system is operational, its sensor may be replaced but not the user's template. In speaker verification, using a different microphone type than the one used during enrollment is a common problem. As a result, the system performance degrades when a different sensor is used. Algorithms developed in speaker verification can certainly be adapted to different biometric modalities. Ultimately, a common noise mismatch framework has to be addressed.

- **User-specific and population assessment.** Current evaluation techniques using standard EPC/DET curve cannot generalize to a different population, size of users and of course the mismatch conditions. This is a drawback because one cannot conclude that if algorithm A is better than B in a database with population X, the result is consistent with another database with population Y. One even has the least idea if algorithm A is better than B for a given user. This issue is particularly important for applications involving personalized biometric devices, e.g., mobile phones and PDAs.

- **User-specific criterion for joint training.** The current fusion systems combine system outputs after the base-systems are trained. A joint training strategy, including the fusion classifier can be potentially useful. It is yet to find out to what extent this training can be beneficial, considering that limited genuine training samples are available per user. We conjecture that joint training is useful in the case where the underlying data streams correlate in time (e.g., audio-visual speech) or in space (e.g., common facial image but different facial features).

8.3 An Open Question

Finally, it should be noted that despite many research works on biometric fusion and its promise of achieving lower verification error rates, it is still an open question why the deployment of multimodal biometric fusion is not widespread after 30 years of research. We conclude this thesis by leaving the reader with the following reflection quoted in [149]:

> "Although multi-modal biometric approaches are theoretically fascinating, the practical path forward in multi-system biometrics is in first fully exploiting the time, cost, and complexity economies of multi-presentation/ instance/sensor/algorithmic data."

Appendix A

Cross-Validation for Score-Level Fusion Algorithms

Algorithm 3 [7] shows how K-fold cross-validation can be used to estimate the correct value of the hyper-parameters of our fusion model, as well as the decision threshold used in the case of authentication. The basic framework of the algorithm is as follows: first perform K-fold cross-validation on the training set by varying the value of the hyper-parameter, and for each hyper-parameter, select the corresponding decision threshold that minimises Half Total Error Rate (HTER); then choose the best hyper-parameter according to this criterion and perform normal training with the best hyper-parameter on the whole training set; finally test the resultant classifier on the test set with HTER evaluated on the previously found decision threshold.

There are several points to note concerning Algorithm 3: $\mathcal{Z}$ is a set of labeled examples of the form $(\mathcal{X}, \mathcal{Y})$, where the first term is a set of patterns and the second term is a set of corresponding labels. The "train" function receives a hyper-parameter θ and a training set, and outputs an optimal classifier $\hat{F}$ by minimising the HTER on the training set. The "test" function receives a classifier $\hat{F}$ and a set of examples, and outputs a set of scores for each associated example. The "thrd$_{HTER}$" function returns a *decision threshold* that minimises HTER by minimising $|\text{FAR}(\Delta) - \text{FRR}(\Delta)|$ with respect to the threshold Δ (FAR(Δ) and FRR(Δ) are false acceptance and false rejection rates, as a function of Δ) while L_{HTER} returns the HTER *value* for a particular decision threshold. What makes this cross-validation different from classical cross-validation is that there is only one single decision threshold and the corresponding HTER value for all the held-out folds and for a given hyper-parameter θ. This is because it is logical to union scores of all held-out folds into one single set of scores to select the decision threshold (and obtain the corresponding HTER).

Algorithm 3 Risk Estimation $(\Theta, K, \mathcal{Z}^{train}, \mathcal{Z}^{test})$

REM: Risk Estimation with K-fold Validation. See [7].
Θ : a set of values for a given hyper-parameter
$\mathcal{Z}^i$: a tuple $(\mathcal{X}^i, \mathcal{Y}^i)$, for $i \in \{train, test\}$ where
$\mathcal{X}$: a set of patterns. Each pattern contains scores/hypothesis from base experts
$\mathcal{Y}$: a set of labels $\in \{client, impostor\}$
Let $\cup_{k=1}^{K} \mathcal{Z}^k = \mathcal{Z}^{train}$ and $\mathcal{Z}^i \cap \mathcal{Z}^j = \emptyset \forall_{i,j}$
for each hyper-parameter $\theta \in \Theta$ **do**
  for each $k = 1, \ldots, K$ **do**
    $\hat{F}_\theta$ = train(θ, $\cup_{j=1, j \neq k}^{K} \mathcal{Z}^j$)
    $\hat{\mathcal{Y}}_\theta^k$ = test($\hat{F}_\theta$, $\mathcal{X}^k$)
  end for
  $\Delta_\theta = \text{thrd}_{HTER}\left(\{\hat{\mathcal{Y}}_\theta^k\}_{k=1}^K, \{\mathcal{Y}^k\}_{k=1}^K\right)$
end for
$\theta^* = \arg\min_\theta \left(L_{HTER}\left(\Delta_\theta, \{\hat{\mathcal{Y}}_\theta^k\}_{k=1}^K, \{\mathcal{Y}^k\}_{k=1}^K\right)\right)$
$\hat{F}_{\theta^*}$ = train(θ^*, $\mathcal{Z}^{train}$)
$\hat{\mathcal{Y}}_{\theta^*}^{test}$ = test($\hat{F}_{\theta^*}$, $\mathcal{X}^{test}$)
return $L_{HTER}(\Delta_{\theta^*}, \hat{\mathcal{Y}}_{\theta^*}^{test}, \mathcal{Y}^{test})$

Appendix B

The WER criterion and Others

The WER criterion of (2.4) (see Section 2.2.2) is similar to the criterion used in the yearly NIST evaluation plans [148, Chap. 8] and also the WER criterion used in the BANCA protocols [5].

The NIST evaluation plans use the C_{DET} point which is defined as:

$$C_{DET}(C_{FR}, C_{FA}) = \underbrace{C_{FR} \times P(C)} \times \text{FRR}(\Delta) + \underbrace{C_{FA} \times P(I)} \times \text{FAR}(\Delta), \tag{B.1}$$

where C_{FA} and C_{FR} are respectively the costs of FA and FR, and $P(k)$ is the prior probability of class $k \in \{C, I\}$.

The BANCA protocols uses a criterion also called "the WER criterion" but is different from (2.4). It is defined as:

$$\text{WER}_{banca}(R, \Delta) = \frac{\text{FRR} + R\,\text{FAR}}{1 + R}, \tag{B.2}$$

where $R \geq 0$ balances between the costs of FAR and FRR.

The two underbraced terms in C_{DET} as well as R of WER_{banca} play the same role as α in (2.5): they adjust for the different costs between FA and FR. Note that this adjustment parameter is not normalized for C_{DET}. Let us explicitly write the grouped underbraced terms in C_{DET} as

$$C_{DET} = \alpha_{FRR}\text{FRR}(\Delta) + \alpha_{FAR}\text{FAR}(\Delta).$$

Since $\min_\Delta C_{DET}$ is equivalent to $\min_\Delta \frac{C_{DET}}{\alpha_{FRR} + \alpha_{FAR}}$, the normalized and non-normalized versions of C_{DET} are equivalent. As a result, (2.5) as well as (2.6) generalizes to both the NIST and BANCA criteria.

In the NIST evaluation, the following constants are used:

$$C_{FR} = 10\,,\, C_{FA} = 1\,,\, P(C) = 0.01 \text{ and } P(I) = 0.99.$$

As a result, $C_{DET} = 0.1 \times \text{FRR} + 0.99 \times \text{FAR}$. By enforcing that the two costs sum to one, it can be observed that $\alpha = 0.91$. For the BANCA protocols, three R values are used, namely 0.1, 1 and 10. They correspond to α values of 0.09, 0.5 and 0.91, respectively.

Appendix C

Experimental Evaluation of the Proposed Parametric Fusion Model

C.1 Validation of F-ratio

This section investigates whether or not the EER derived from the F-ratio is acceptable. This is done by comparing the *theoretical* EER derived using (4.14)) with its *empirical* counterpart, i.e., the minimum HTER as appeared in (2.7). Note that the minimum HTER is found by minimizing WER with respect to the threshold as appeared in (2.6) with $\alpha = 0.5$.

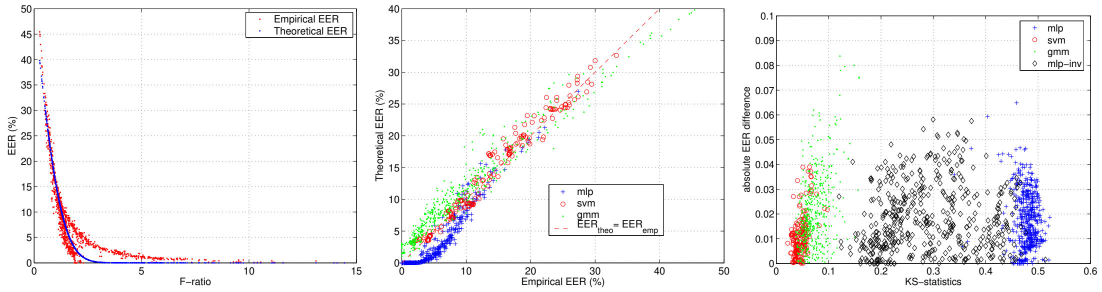
We conducted 1186 experiments on the BANCA database as described in Section 2.1.2 and [80][1]. There are 490 experiments from the output of MLPs; 182 from SVMs; and 514 from GMMs. Two approaches are adopted here. The first approach is to test whether for each of the 1186 experiments, the respective client and impostor scores are normally distributed or not. The second approach is to directly compare the empirical EER against its theoretical counterpart (assuming that client and impostor distributions are normally distributed).

For the first approach, we applied the Lillie-test [24], which evaluates the hypothesis that a set of (client or impostor) scores has a normal distribution with unspecified mean and variance against the alternative that the set of scores does not have a normal distribution. This test is similar to Kolmogorov-Smirnov (KS) test, but it adjusts for the fact that the parameters of the normal distribution are estimated from the set of scores rather than specified in advance. Using this test, we found that 22.85% of impostor scores and 25.89% of client scores (out of 1186 experiments) supported the hypothesis that they are Gaussian distributed. Hence, only approximately a quarter of the distributions are Gaussian according to the Lillie test.

The results of the second approach are shown in Figure C.1. From Figure C.1(a), it can be seen that both the theoretical and empirical EERs are non-linearly and inversely proportional to their F-ratio. Removing the F-ratio, we compared the theoretical EER directly with its empirical counterpart in Figure C.1(b). Here the output of different classifiers are plotted with different symbols. If the theoretical EER matches exactly its empirical EER, the points (each one corresponding to a single experiment) should be on the diagonal line. One measure of agreement is to use correlation. Its value is evaluated to be 0.9573, indicating the the variables are *strongly correlated*. In other words, knowing theoretical EER, one can use the correlation to *approximately* estimate the empirical EER.

One way to understand the effect of deviation from Gaussian assumption on the quality of estimated EER, we plotted the absolute EER difference (between theoretical EER and empirical EER) versus the average KS-statistic of their respective client and impostor distributions in Figure C.1(c). Note that from each experiment, we will have two KS-statistics, one for each distribution. KS-statistic quantifies the degree of divergence from normal distribution. It is an intermediate calculation used in the Lillie test to accept or reject the Gaussian hypothesis. Note that KS-statistic itself is not used to accept or reject the

[1] The NIST2001 and XM2VTS databases have also been used and we obtained similar results and conclusions in [103].



(a) Empirical and theoretical EER vs. F-ratio

(b) Empirical vs. theoretical EER

(c) Absolute EER difference vs. KS-statistic

Figure C.1: Results of experiments carried out using all the available 1186 experiments on the BANCA score database. (a) Theoretical EER and empirical EER (HTER) versus their common F-ratio (b) Theoretical EER versus empirical EER (HTER) using output of different classifiers – 490 MLPs, 182 SVMs and 514 GMMs; the correlation coefficient between the two variables is 0.9573. (c) Absolute EER difference between theoretical EER and empirical EER versus the average KS-statistic between the corresponding client and impostor distributions. KS-statistic measures the degree of deviation from Gaussian assumption. Note that "mlp-inv" denotes the experiments involving MLP outputs that are converted to the logit space where the conditional scores are once again more normally distributed. Their corresponding KS-statistic after such post-processing is much smaller.

Gaussian hypothesis. As can be seen, the output of MLPs (trained using sigmoid output function) gives high KS-statistic whereas the outputs of SVMs and GMMs conform better to the Gaussian assumption.

Prior to this experiment, we thought that deviation from Gaussian would mean large absolute EER difference. If this was the case, absolute EER difference would have been increasing proportionally with respect to the KS-statistic. It turns out that this is not the case. In Figure C.1(c), despite high KS-statistic of MLP outputs, their corresponding absolute EER differences are spread below 0.06; some are even near 0! Hence, deviation from Gaussian does not mean large absolute EER difference. In other words, the theoretical EER is fairly robust to deviation from the Gaussian assumption.

It should be noted that a more interesting issue to investigate is the *relative* values of EER, i.e., if the empirical EER of experiment a is more than the empirical EER of experiment b, does the theoretical EER of these experiments also follow the same trend? Using the data at hand, we calculated all the possible combinations of two EER experiments. This turns out to be ${}^{1186}C_2 = 702,705$ combinations. The number of "disagreements", d, can be calculated as follows:

$$d = \left|(\text{EER}_a^{emp} > \text{EER}_b^{emp}) - (\text{EER}_a^{theo} > \text{EER}_b^{theo})\right| \tag{C.1}$$

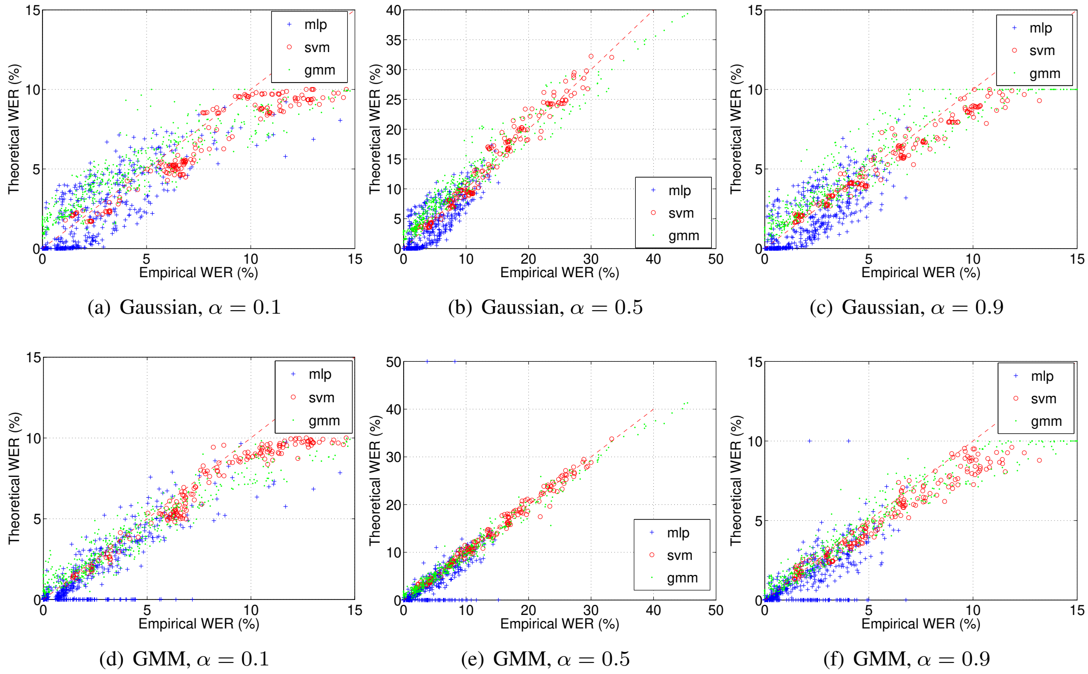
for $(a,b) \in \{(1,2),(1,3),\ldots,(1185,1186)\}$ and

$$(z_1 > z_2) = \begin{cases} 1 & \text{if true} \\ 0 & \text{otherwise.} \end{cases} \tag{C.2}$$

The percentage of disagreement turns out to be 11%. If the 1186 experiments are representative of biometric authentication tasks, we can conclude that to compare any two experiments, the theoretical EER (calculated from the F-ratio) can give a correct answer 89% of the time as compared to using the empirical EER as the ground-truth.

C.2 Beyond EER and Beyond Gaussian Assumption

In the last section, although only the EER point is studied, one can extend the present finding to a more general case, whereby the EER constraint by its definition, i.e, $\text{EER}(\Delta) = \text{FAR}(\Delta) = \text{FRR}(\Delta)$, does not



(a) Gaussian, $\alpha = 0.1$ (b) Gaussian, $\alpha = 0.5$ (c) Gaussian, $\alpha = 0.9$

(d) GMM, $\alpha = 0.1$ (e) GMM, $\alpha = 0.5$ (f) GMM, $\alpha = 0.9$

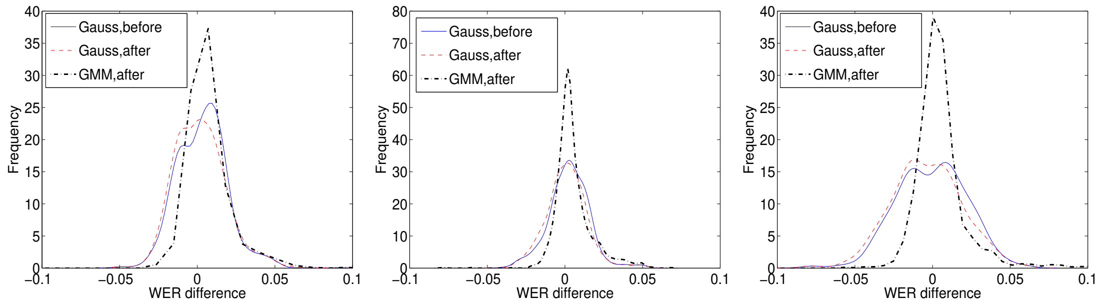
Figure C.2: Empirical WERs vs. approximated WERs. Compare each of a, b and c with d, e and f. The approximated WERs refer to those calculated using the class-conditional Gaussian assumption for a–c and those using the assumption by GMM d–f. For each of a–c or d–f, the following α values are used $\{0.1, 0.5, 0.9\}$. Each point represents one of the 1186 BANCA datasets. For those computed with the Gaussian assumption, we converted the scores into the logit space using $f_{LLR}(y)$. This is the one shown here. We also omitted this pre-processing step but not shown here to avoid cluttering the figures. The distribution of the error deviates of GMM, Gaussian with and without pre-processing are shown in Figure C.3.

hold anymore. In this case, one is interested in WER with varying α values. We choose the following α values: $\{0.1, 0.5, 0.9\}$ which approximate the scenarios in the BANCA protocols.

We also propose here an improvement over the Gaussian assumption by using a mixture of Gaussians (GMM) as appeared in (3.22). Of course, a non-parametric Parzen window with Gaussian kernel could have been used. In either case, any hyper-parameter (number of Gaussian component for GMM; kernel width for Parzen window) are tuned using two-fold cross validation in our case. The results are shown in Figure C.2 and the distribution of their error deviates are shown in Figure C.3. The error deviate is defined as the difference between the empirical WER and the theoretical WER. Recall that the empirical WER is based on empirical FAR and FRR obtained from the data whereas the theoretical WER is based on FAR and FRR with Gaussian assumption, as appeared in (4.12) and (4.11). As can be observed and expected, the GMM solution fits better the distribution (smallest bias) but the Gaussian solution is still robust to different WER values. In both cases, the WER estimates are less accurate towards boundary values (near 0 or 1). In any case, the robustness of Gaussian assumption, as in any practical application, is confirmed.

C.3 The Effectiveness of F-ratio as a Performance Predictor

The goal of this section is to test the effectiveness of F-ratio as a performance predictor compared to the commonly used correlation. We used the BANCA fusion datasets as outlined in Section 2.1.2. For this study, the mean fusion operator is used.



(a) $\alpha = 0.1$ (b) $\alpha = 0.5$ (c) $\alpha = 0.9$

Figure C.3: The distribution of WER error deviates between the empirical and the theoretical counterpart for different α values.

C.3.1 Experimental Results Using Correlation

A naive approach to analyse fusion is to empirically find the relationship between minimum *a posteriori* HTER and the sum of correlation of client and impostor distributions. Let the client and impostor-dependent correlations between two baseline systems (to be fused) be the scalars ρ_C and ρ_I, respectively[2]. The results are shown in Figure. C.4. From this figure, it can be observed that multimodal fusion experiments have less correlated scores while multi-feature fusion experiments have high correlated scores. One would have expected that the minimum *a posteriori* HTER is somewhat proportional to $\rho_C + \rho_I$. This is actually partially true because the variance of participating systems are not taken into account. As a result, there is no clear trend in this graph and one cannot conclude that HTER is proportional to correlation.

C.3.2 Experimental Results Using F-ratio

We distinguish here two concepts: empirical F-ratio and its theoretical counterpart. For each of the parameters to be tested, *empirical* means that the respective parameter is directly estimated on the combined system output y_{COM}; and *theoretical* means that no fusion experiment is performed – only the respective parameters need to be estimated.

Figure C.5(a) shows empirical F-ratio versus its theoretical counterpart (based on (4.18)) calculated uniquely on the development set. As can be seen both empirical and theoretical F-ratios are *exactly* the same. Their equivalence can be shown mathematically (see Section D.5). Figure C.5(b) plots the F-ratio found on the development set versus the F-ratio found on the evaluation set. They are not exactly the same this time because there is a mismatch between these two data sets. Nevertheless, their correlation is 0.90, indicating that knowing F-ratio from the development set, it is possible to predict reasonably F-ratio of the evaluation set. This property will be exploited in Section 4.5.

As a by-product of these set of experiments, Figure C.5(c) plots the following two variables: correlation of client and that of impostor scores. The overall correlation between these two variables is 0.83. This indicates that knowing the covariance (or correlation; since one is proportional to the other as shown in (4.27)) of the impostor scores, one can approximate the covariance of the client scores. Note that all intramodal fusion experiments have high correlation values. Figure C.5(c) thus has two clusters. The cluster in the upper right corner belongs to intramodal fusion experiments whereas the cluster in the lower left corner belongs to multimodal fusion experiments.

Summary

Comparing Figure C.4 with Figure C.5(a) (or Figure C.5(b)), we conclude that F-ratio is an adequate fusion performance predictor.

[2]In general, the correlation of scores of N responses are a matrix of N by N with elements $\rho_{m,n}$. It has the property that $\rho_{m,m} = 1$ and $\rho_{m,n} = \rho_{n,m}$. In the case of two responses, we simply write ρ in place of $\rho_{1,2}$.

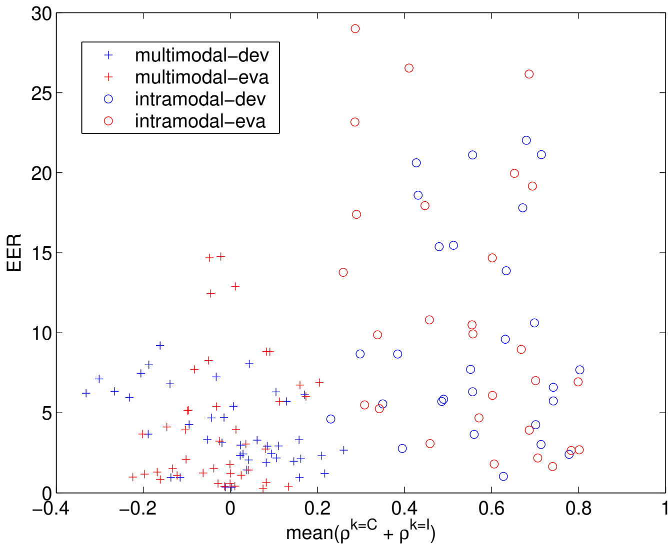
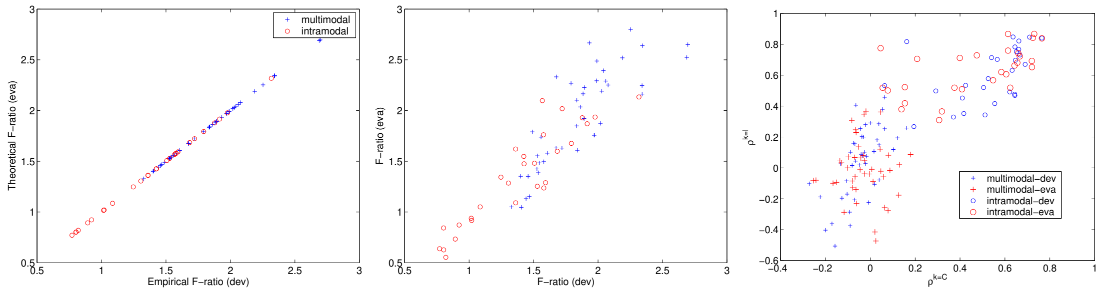


Figure C.4: Empirical EER of combining two baseline systems versus $\rho_C + \rho_I$ using the BANCA database. The crosses represent experiments combining 2 modalities while the circles represent those combining 2 features of the *same* modality. The correlation between the two variables is 0.38.



(a) Empirical F-ratio vs. theoretical F-ratio

(b) F-ratio (dev) vs F-ratio (eva)

(c) Correlation of client vs impostor scores

Figure C.5: Effectiveness of F-ratio as a fusion performance predictor. Experiments carried out on fusion of ${}^5C_2 \times 7 = 70$ experiments, i.e., combining 2 systems each time out of five available systems, for all the 7 BANCA protocols: (a) Empirical F-ratio versus theoretical F-ratio on the development set. (b) F-ratio of development set versus its evaluation set counterpart. The correlation between the two variables is 0.90. (c) Correlation of client scores versus correlation of impostor scores. The correlations between the two variables (class-dependent correlations) on the development and evaluation sets are 0.85 and 0.80, respectively.

Appendix D

Miscellaneous Proves

D.1 On the Redundancy of Linear Score Normalization with Trainable Fusion

Suppose that a linear classifier is used. Then, the fused score can be written as:

$$
\begin{aligned}
y_{COM} &= \sum_i w_i y_i^{lin} - \Delta = \sum_i w_i A_i (y_i - B_i) - \Delta \\
&= \sum_i \underbrace{w_i Ai}\, y_i - \underbrace{\sum_i w_i A_i B_i - \Delta}
\end{aligned}
\tag{D.1}
$$

Comparing (D.1) with the linear combination without normalisation, as in (3.23), we see that the first underbraced term is the new weight whereas the second underbraced term is the new decision threshold. This shows that if $y_i|\forall_i$ are unevenly scaled, their scaling factor A_i may not be necessary as it will be automatically absorbed by the weight. This implies that if scores are not evenly scaled, the weights in the linear combination should be allowed to take on any values, without the constraint $\sum_i w_i = 1$. This shows that linear score normalisation is not *necessary*. □

D.2 Deriving μ^k_{wsum} and $(\sigma^k_{wsum})^2$

The central idea consists of projecting the N dimensional score onto a one dimensional (combined) score. Suppose that the class conditional scores (prior to fusion) are modeled by a multivariate Gaussian with mean $\boldsymbol{\mu}^k = [\mu^k_1, \ldots, \mu^k_N]'$ and covariance $\boldsymbol{\Sigma}^k$ of N-by-N dimensions. Let $\boldsymbol{\Sigma}^k_{i,j}$ be the i-th row and j-th column of covariance matrix $\boldsymbol{\Sigma}^k$ for $k = \{C, I\}$. The linear projection from N dimensions of score to one dimension of score has the same effect on the Gaussian distribution: from N multivariate Gaussian distribution to a single Gaussian distribution with mean μ^k_{wsum} and variance $(\sigma_{wsum})^2$ defined in the fourth row of Table 4.1 for each class k. The mean operator is derived similarly with $w_i = \frac{1}{N}\forall i$. Note that the weight w_i affects both the mean and variance of fused scores.

The expected value of Y^k_{wsum}, for $k = \{C, I\}$, is:

$$
\begin{aligned}
\mu^k_{wsum} &\equiv E[y^k_{wsum}] = \sum_{i=1}^{N} w_i E[y^k_i] \\
&= \sum_{i=1}^{N} \frac{w_i}{A_i}\left(E[y^k_i] - B_i\right) \\
&= \sum_{i=1}^{N} \frac{w_i}{A_i}\left(\mu^k_i - B_i\right)
\end{aligned}
\tag{D.2}
$$

The variance of y^k_{wsum} is:

$$\begin{aligned}
(\sigma^k_{wsum})^2 &= Cov(y^k_{wsum}, y^k_{wsum}) \\
&= E\left[\left(y^k_{wsum} - E[y^k_{wsum}]\right)^2\right] \\
&= E\left[\left(\sum_{i=1}^{N}\frac{w_i(y^k_i - B_i)}{A_i} - \sum_{i=1}^{N}\frac{w_i(\mu^k_i - B_i)}{A_i}\right)^2\right] \\
&= E\left[\left(\sum_{i=1}^{N}\frac{w_i(y^k_i - B_i)}{A_i}\right)^2\right] \\
&= E\left[\left(\sum_{i=1}^{N}\frac{w_i\eta^k_i}{A_i}\right)^2\right] \qquad \text{(D.3)}
\end{aligned}$$

To expand (D.3), one should take care of possible correlation between different η^k_m and η^k_n, as follows:

$$\begin{aligned}
(\sigma^{norm,k}_{wsum})^2 &= E\left[\left(\sum_{m=1}^{N}\sum_{n=1}^{N}\frac{w_m\eta^k_m w_n\eta^k_n}{A_m A_n}\right)\right] \\
&= \sum_{m=1}^{N}\sum_{n=1}^{N}\frac{w_m w_n}{A_m A_n}E\left[\eta^k_m\eta^k_n\right] \qquad \text{(D.4)}
\end{aligned}$$

for any $k \in \{C, I\}$. □

D.3 Proof of $(\sigma^k_{COM})^2 \leq (\sigma^k_{AV})^2$

For simplicity, we will omit the conditioning k. For the case $\rho_{m,n} \neq 0$, the inequality can be written as:

$$\sigma^2_{COM} \leq \sigma^2_{AV}$$

$$\frac{1}{N^2}\sum_{j=1}^{N}\sigma^2_j + \frac{2}{N^2}\sum_{m=1,m<n}^{N}\rho_{m,n}\sigma_m\sigma_n \leq \frac{1}{N}\sum_{j=1}^{N}\sigma^2_j \qquad \text{(D.5)}$$

By multiplying both sides by N^2 and rearranging them, we obtain:

$$0 \leq (N-1)\sum_{j=1}^{N}\sigma^2_j - 2\sum_{m=1,m<n}^{N}\rho_{m,n}\sigma_m\sigma_n.$$

Given that $(N-1)\sum_{i=1}^{N}\sigma^2_i = \sum_{i=1,i<j}^{N}(\sigma^2_i + \sigma^2_j)$ (the proof can be found in the appendix), this inequality can further be simplified to:

$$
\begin{aligned}
0 &\leq \sum_{m=1,m<n}^{N} (\sigma_m^2 + \sigma_n^2) - 2 \sum_{m=1,m<n}^{N} \rho_{m,n}\sigma_m\sigma_n \\
0 &\leq \sum_{m=1,m<n}^{N} \left(\sigma_m^2 - 2\rho_{m,n}\sigma_m\sigma_n + \sigma_n^2\right) \\
0 &\leq \sum_{m=1,m<n}^{N} \left((\sigma_m^2 - 2\rho_{m,n}\sigma_m\sigma_n + \rho_{m,n}^2\sigma_n^2 + (1-\rho_{m,n}^2)\sigma_n^2)\right) \\
0 &\leq \sum_{m=1,m<n}^{N} \left((\sigma_m - \rho_{m,n}\sigma_n)^2 + (1-\rho_{m,n}^2)\sigma_n^2\right). \qquad \text{(D.6)}
\end{aligned}
$$

Hence, regardless of the value of $\rho_{m,n}$, the inequality is always true. □

D.4 Proof of $(N-1)\sum_{i=1}^{N}\sigma_i^2 = \sum_{i=1,i<j}^{N}(\sigma_i^2+\sigma_j^2)$

Let σ_i be a random variable and $i = 1, \ldots, N$. The term $\sum_{i=1,i<j}^{N}(\sigma_i^2+\sigma_j^2)$ can be interpreted as $\sum_{i=1}^{N}\sum_{j=i+1}^{N}(\sigma_i^2+\sigma_j^2)$. The problem now is to count how many σ_k^2 there are in the term, for any $k = 1, \ldots, N$.

There are two cases here. The first case is when $i = k$, the term $\sum_{i=1}^{N}\sum_{j=i+1}^{N}(\sigma_i^2+\sigma_j^2)$ becomes: $\sum_{j=k+1}^{N}(\sigma_k^2+\sigma_j^2)$. There are $(N-k)$ terms of σ_k^2.

In the second case, when $j = k$, the term $\sum_{i=1}^{N}\sum_{j=i+1}^{N}(\sigma_i^2+\sigma_j^2)$ then becomes: $\sum_{i=1}^{k-1}(\sigma_i^2+\sigma_k^2)$. There are $(k-1)$ terms of σ_k^2.

The total number of σ_k^2 is just the sum of these two cases, which is $(N-k)+(k-1) = (N-1)$, for any k drawn from $1, \ldots, N$. The sum of $(N-1)$ σ_k^2 over all possible $k = 1, \ldots, N$ then gives $(N-1)\sum_{k=1}^{N}\sigma_k^2$.

Therefore, $(N-1)\sum_{i=1}^{N}\sigma_i^2 = \sum_{i=1,i<j}^{N}(\sigma_i^2+\sigma_j^2)$. □

D.5 Proof of Equivalence between Empirical F-ratio and Theoretical F-ratio

The estimated theoretical and empirical parameters can be shown to be exactly the same mathematically. Suppose there are M^k accesses, where M^C are the number of client accesses and M^I are the number of impostor accesses. Suppose also that $Y_{i,u}^k$ is the output of the i-system and u-th access given that the class label is $k = \{C, I\}$, and $i = 1, \ldots, N$ and $u = 1, \ldots, M^k$. μ_i^k can be estimated by:

$$\hat{\mu}_i^k \equiv \frac{1}{M}^k \sum_{u=1}^{M^k} Y_{i,u}^k \equiv \bar{Y}_{i,\cdot}^k. \qquad \text{(D.7)}$$

For the u-th access, the combined score is:

$$\frac{1}{N}\sum_{i=1}^{N} Y_{\cdot,u}^k \equiv \bar{Y}_{\cdot,u}^k. \qquad \text{(D.8)}$$

The empirical estimate of μ_{COM}^k, $\hat{\mu}_{COM,emp}^k$ is given by:

$$\frac{1}{M}\sum_{u=1}^{M^k} \bar{Y}_{\cdot,u}^k \equiv \bar{Y}_{\cdot,\cdot}^k. \qquad \text{(D.9)}$$

Note that:

$$\begin{aligned}
\hat{\mu}^k_{COM,emp} &= \frac{1}{M}\sum_{u=1}^{M^k} \bar{Y}_{\cdot,u} \\
&= \frac{1}{N}\sum_{i=1}^{N} \bar{Y}_{i,\cdot} \;\; \text{(interchange the } i \text{ and } u \text{ summations)} \\
&= \frac{1}{N}\sum_{i=1}^{N} \hat{\mu}^k_i \\
&= \hat{\mu}^k_{COM,theo}.
\end{aligned} \tag{D.10}$$

Hence, they are the same. The empirical variance can be calculated as follows:

$$\left(\hat{\sigma}^k_{COM,emp}\right)^2 = \frac{1}{M}\sum_{u=1}^{M}\left(\bar{Y}_{\cdot,u} - \bar{Y}_{\cdot,\cdot}\right) \tag{D.11}$$

The theoretical variance is obtained by estimating the terms $(\sigma^k_i)^2$ and $\rho^k_{i,j}\sigma^k_i\sigma^k_j$ in the expression of $(\sigma^k_{COM})^2$, as shown in (4.26). The estimate of $(\sigma^k_i)^2$ is given by:

$$\frac{1}{M}\sum_{u=1}^{M}\left(Y^k_{i,u} - \bar{Y}^k_{i,\cdot}\right)^2. \tag{D.12}$$

The estimate of $\rho^k_{i,j}\sigma^k_i\sigma^k_j$ is given by:

$$\frac{1}{M}\sum_{u=1}^{M}\left(Y^k_{i,u} - \bar{Y}^k_{i,\cdot}\right)\left(Y^k_{j,u} - \bar{Y}^k_{j,\cdot}\right)^2. \tag{D.13}$$

Plugging in these two estimates into the expression for $(\sigma^k_{COM})^2$, we get the theoretical estimate of the variance of the fused scores as:

$$\begin{aligned}
&\left(\hat{\sigma}^k_{COM,theo}\right)^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{M}\sum_{u=1}^{M}\left(Y^k_{i,u} - \bar{Y}^k_{i,\cdot}\right)\right] \\
&\quad + \frac{2}{N}\sum_{i=1,j>i}^{N}\left[\left(Y^k_{i,u} - \bar{Y}^k_{i,\cdot}\right)\left(Y^k_{j,u} - \bar{Y}^k_{j,\cdot}\right)\right] \\
&= \frac{1}{M}\sum_{u=1}^{M}\left[\frac{1}{N^2}\sum_{i,j=1}^{N}\left(Y^k_{i,u} - \bar{Y}^k_{i,\cdot}\right)\left(Y^k_{j,u} - \bar{Y}^k_{j,\cdot}\right)\right] \\
&= \frac{1}{M}\sum_{u=1}^{M}\left(\bar{Y}_{\cdot,u} - \bar{Y}_{\cdot,\cdot}\right) \\
&= \left(\hat{\sigma}^k_{COM,emp}\right)^2.
\end{aligned} \tag{D.14}$$

Because the empirical and theoretical μ^k_{COM} and σ^k_{COM} are the *same*, the empirical and theoretical F-ratios will be exactly the same. Using the definition of F-ratio in (4.15), the theoretical F-ratio of the combined score can be defined as:

$$\text{F-ratio}_{COM,theo} \equiv \frac{\hat{\mu}^C_{COM,theo} + \hat{\mu}^I_{COM,theo}}{\hat{\sigma}^C_{COM,theo} + \hat{\sigma}^I_{COM,theo}}. \tag{D.15}$$

The empirical F-ratio is:

$$\begin{aligned}
\text{F-ratio}_{COM,emp} &\equiv \frac{\hat{\mu}^C_{COM,emp} + \hat{\mu}^I_{COM,emp}}{\hat{\sigma}^C_{COM,emp} + \hat{\sigma}^I_{COM,emp}} \\
&= \frac{\hat{\mu}^C_{COM,theo} + \hat{\mu}^I_{COM,theo}}{\hat{\sigma}^C_{COM,theo} + \hat{\sigma}^I_{COM,theo}} \\
&= \text{F-ratio}_{COM,theo}
\end{aligned} \tag{D.16}$$

Hence, the theoretical F-ratio is exactly the same as the empirical F-ratio. This applies also for normalised version of Y. $\square$

Bibliography

[1] O. Ushmaev abd S. Novikov. Biometric fusion: Robust approach. In *Workshop on Multimodal User Authentication (MMUA 2006)*, Toulouse, 2006.

[2] A. Adler and M. E. Schuckers. Calculation of a Composite DET Curve. In *LNCS 3546, 5th Int'l. Conf. Audio- and Video-Based Biometric Person Authentication (AVBPA 2005)*, pages 860–868, New York, 2005.

[3] B.C. Arnold, N. Balakrishnan, and H.N. Nagaraja. *A First Course in Order Statistics*. Wiley, New York, 1992.

[4] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas. Score Normalization for Text-Independant Speaker Verification Systems. *Digital Signal Processing (DSP) Journal*, 10:42–54, 2000.

[5] E. Bailly-Baillière, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariéthoz, 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.

[6] M. Ben, R. Blouet, and F. Bimbot. A Monte-Carlo Method For Score Normalization in Automatic Speaker Verification Using Kullback-Leibler Distances. In *Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 689–692, Orlando, 2002.

[7] S. Bengio, C. Marcel, S. Marcel, and J. Mariéthoz. Confidence Measures for Multimodal Identity Verification. *Information Fusion*, 3(4):267–276, 2002.

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

[9] S. Bengio and J. Mariéthoz. A Statistical Significance Test for Person Authentication. In *The Speaker and Language Recognition Workshop (Odyssey)*, pages 237–244, Toledo, 2004.

[10] J. Bigun, J. Fierrez-Aguilar, J. Ortega-Garcia, and J. Gonzalez-Rodriguez. Multimodal Biometric Authentication using Quality Signals in Mobile Communnications. In *12th Int'l Conf. on Image Analysis and Processing*, pages 2–11, Mantova, 2003.

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

[12] R.M. Bolle, N.K. Ratha, and S. Pankanti. Error Analysis of Pattern Recognition Systems: the Subsets Bootstrap. *Computer Vision and Image Understanding*, 93(1):1–33, 2004.

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

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

[15] S. Carcia-Salicetti, C. Beumier, G. Chollet, B. Dorizzi, J. Leroux les Jardins, J. Lunter, Y. Ni, and D. Petrovska-Delacrétaz. BIOMET: A Multimodal Person Authentication Database Including Face, Voice, Fingerprint, Hand and Signature Modalities. In *LNCS 2688, 4th Int'l. Conf. Audio- and Video-Based Biometric Person Authentication (AVBPA 2003)*, pages 845–853, Guildford, 2003.

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

[17] 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.

[18] K. Chen. Towards Better Making a Decision in Speaker Verification. *Pattern Recognition*, 36(2):329–346, 2003.

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

[20] Y. Chen, S. Dass, and A. Jain. Localized iris image quality using 2-d wavelets. In *Proc. Int'l Conf. on Biometrics (ICB)*, pages 373–381, Hong Kong, 2006.

[21] Y. Chen, S.C. Dass, and A.K. Jain. Fingerprint Quality Indices for Predicting Authentication Performance. In *LNCS 3546, 5th Int'l. Conf. Audio- and Video-Based Biometric Person Authentication (AVBPA 2005)*, pages 160–170, New York, 2005.

[22] A. Cohen and Y. Zigel. On Feature Selection for Speaker Verification. In *Proc. COST 275 workshop on The Advent of Biometrics on the Internet*, pages 89–92, Rome, November 2002.

[23] J. Collins, M. Mancilulli, R. Hohlfeld, D. Finch, G. Sandri, and E. Shtatland. A Random Number Generator Based on the Logit Transform of the Logistic Variable. *Computers in Physics*, 6:630–632, 1992.

[24] W. J. Conover. *Practical Nonparametric Statistics*. Wiley, 1980.

[25] Conrad Sanderson. The VidTIMIT Database. Communication 06, IDIAP, 2002.

[26] N. Cristianini and J. Shawe-Taylor. *Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press, 2000.

[27] S. Dass, K. Nandakumar, and A. Jain. A Principled Approach to Score Level Fusion in Multimodal Biometric Systems. In *5th Int'l. Conf. Audio- and Video-Based Biometric Person Authentication (AVBPA 2005)*, pages 1049–1058, New York, 2005.

[28] J. Daugman. High confidence visual recognition of persons by a test of statistical independence. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 15(11):1148–1161, 1993.

[29] J. Daugman. Biometric decision landscapes. Technical Report TR482, University of Cambridge Computer Laboratory, 2000.

[30] T.G. Dietterich. Approximate Statistical Test for Comparing Supervised Classification Learning Algorithms. *Neural Computation*, 10(7):1895–1923, 1998.

[31] T.G. Dietterich. Ensemble Methods in Machine Learning. In *Multiple Classifier Systems*, pages 1–15, 2000.

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

[33] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. Reynolds. Sheep, Goats, Lambs and Woves: A Statistical Analysis of Speaker Performance in the NIST 1998 Speaker Recognition Evaluation. In *Int'l Conf. Spoken Language Processing (ICSLP)*, Sydney, 1998.

[34] P. Domingos and M. Pazzani. On the Optimality of the Simple Bayesian Classifier Under Zero-One Loss. *Machine Learning*, 29(2–3):103–130, 1997.

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

[36] J-L. Dugelay, J-C. Junqua, K. Rose, and M. Turk. *Workshop on Multimodal User Authentication (MMUA 2003)*. no publisher, Santa Barbara, CA, 11–12 December, 2003.

[37] R.P.W. Duin. The Combining Classifier: To Train Or Not To Train? In *Proc. 16th Int'l Conf. Pattern Recognition (ICPR)*, pages 765–770, Quebec, 2002.

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

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

[40] J. Fierrez-Aguilar, D. Garcia-Romero, J. Ortega-Garcia, and J. Gonzalez-Rodriguez. Exploiting General Knowledge in User-Dependent Fusion Strategies For Multimodal Biometric Verification. In *IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, volume 5, pages 617–620, Montreal, 2004.

[41] J. Fierrez-Aguilar, D. Garcia-Romero, J. Ortega-Garcia, and J. Gonzalez-Rodriguez. Bayesian Adaptation for User-Dependent Multimodal Biometric Authentication. *Pattern Recognition*, 38:1317–1319, 2005.

[42] J. Fierrez-Aguilar, J. Ortega-Garcia, D. Garcia-Romero, and J. Gonzalez-Rodriguez. A Comparative Evaluation of Fusion Strategies for Multimodal Biometric Verification. In *LNCS 2688, 4th Int'l. Conf. Audio- and Video-Based Biometric Person Authentication (AVBPA 2003)*, pages 830–837, Guildford, 2003.

[43] J. Fierrez-Aguilar, J. Ortega-Garcia, and J. Gonzalez-Rodriguez. Target Dependent Score Normalisation Techniques and Their Application to Signature Verification. In *LNCS 3072, Int'l Conf. on Biometric Authentication (ICBA)*, pages 498–504, Hong Kong, 2004.

[44] J. Fierrez-Aguilar, J. Ortega-Garcia, J. Gonzalez-Rodriguez, and J. Bigun. Kernel-Based Multimodal Biometric Verification Using Quality Signals. In *Defense and Security Symposium, Workshop on Biometric Technology for Human Identification, Proc. of SPIE*, volume 5404, pages 544–554, 2004.

[45] M.A.T. Figueiredo and A.K. Jain. Unsupervised learning on finite mixture models. *Pattern Analysis and Machine Intelligence*, 24(3), March 2002.

[46] P. J. Flynn, K. W. Bowyer, and P. J. Phillips. Assessment of Time Dependency in Face Recognition: An Initial Study. In *LNCS 2688, 4th Int'l. Conf. Audio- and Video-Based Biometric Person Authentication (AVBPA 2003)*, pages 44–51, Guildford, 2003.

[47] Y. Freund and R. Schapire. A Short Introduction to Boosting. *J. Japan. Soc. for Artificial Intelligence*, 14(5):771–780, 1999.

[48] S. Furui. Cepstral Analysis for Automatic Speaker Verification. *IEEE Trans. Acoustic, Speech and Audio Processing / IEEE Trans. on Signal Processing*, 29(2):254–272, 1981.

[49] D. Garcia-Romero, J. Fierrez-Aguilar, J. Gonzalez-Rodriguez, and J. Ortega-Garcia. On the Use of Quality Measures for Text Independent Speaker Recognition. In *The Speaker and Language Recognition Workshop (Odyssey)*, pages 105–110, Toledo, 2004.

[50] D. Garcia-Romero, J. Gonzalez-Rodriguez, J. Fierrez-Aguilar, and J. Ortega-Garcia. U-Norm Likelihood Normalisation in PIN-Based Speaker Verification Systems. In *LNCS 2688, 4th Int'l. Conf. Audio- and Video-Based Biometric Person Authentication (AVBPA 2003)*, pages 208–213, Guildford, 2003.

[51] J.L. Gauvain and C.-H. Lee. Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Obervation of Markov Chains. *IEEE Tran. Speech Audio Processing*, 2:290–298, 1994.

[52] D. Genoud. *Reconnaissance et Transformation de Locuteur*. PhD thesis, Ecole Polythechnique Fédérale de Lausanne (EPFL), Switzerland, 1998.

[53] G. Gravier, J. Kharroubi, and G. Chollet. On the Use of Prior Knowledge in Normalization Schemes for Speaker Verification. *Digital Signal Processing (DSP) Journal*, 10:213–225, 2000.

[54] P. Griffin. Optimal biometric fusion for identity verification. no. rdnj-03-0064, Identix Corporate Research Center, 2004.

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

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

[57] L. Hong, A.K. Jain, and S. Pankanti. Can Multibiometrics Improve Performance? Technical Report MSU-CSE-99-39, Computer Science and Engineering, Michigan State University, East Lansing, Michigan, 1999.

[58] Y. Huang and C. Suen. A Method of Combining Multiple Experts for the Recognition of Unconstrained Handwritten Numerals. *IEEE Trans. Pattern Recognition and Machine Intelligence*, 17(1):1, 1995.

[59] S. Ikbal, H. Misra, and H. Bourlard. Phase Auto-Correlation (PAC) derived Robust Speech Features. In *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP-03)*, pages 133–136, Hong Kong, 2003.

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

[61] A. Jain and A. Ross. Learning User-Specific Parameters in Multibiometric System. In *Proc. Int'l Conf. of Image Processing (ICIP 2002)*, pages 57–70, New York, 2002.

[62] A.K. Jain, R. Bolle, and S. Pankanti. *Biometrics: Person Identification in a Networked Society*. Kluwer Publications, 1999.

[63] Johnny Mariéthoz and Samy Bengio. A Bayesian Framework for Score Normalization Techniques Applied to Text Independent Speaker Verification. *IEEE Signal Processing Letters*, 12(7):532–535, 2005.

[64] K. Jonsson, J. Kittler, Y. P. Li, and J. Matas. Support vector machines for face authentication. *Image and Vision Computing*, 20:269–275, 2002.

[65] M. S. Kamel and N. M. Wanas. Data Dependence in Combining Classifiers. In *LNCS 2709, Proc. 4th Int'l Workshop on Multiple Classifier Systems (MCS 2003)*, pages 1–14, 2003.

[66] 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.

[67] J. Kittler, G. Matas, K. Jonsson, and M. Sanchez. Combining Evidence in Personal Identity Verification Systems. *Pattern Recognition Letters*, 18(9):845–852, 1997.

[68] J. Kittler, K. Messer, and J. Czyz. Fusion of Intramodal and Multimodal Experts in Personal Identity Authentication Systems. In *Proc. Cost 275 Workshop*, pages 17–24, Rome, 2002.

[69] A. Krogh and J. Vedelsby. Neural Network Ensembles, Cross-Validation and Active-Learning. *Advances in Neural Information Processing Systems*, 7, 1995.

[70] K. Kryszczuk, J. Richiardi, P. Prodanov, and A. Drygajlo. Error Handling in Multimodal Biometric Systems using Reliability Measures. In *Proc. 12th European Conference on Signal Processing*, Antalya, Turkey, September 2005.

[71] A. Kumar and D. Zhang. Integrating Palmprint with Face for User Authentication. In *Workshop on Multimodal User Authentication (MMUA 2003)*, pages 107–112, Santa Barbara, 2003.

[72] L. Kuncheva., J.C. Bezdek, and R.P.W. Duin. Decision Template for Multiple Classifer Fusion: An Experimental Comparison. *Pattern Recognition Letters*, 34:228–237, 2001.

[73] L.I. Kuncheva. A Theoretical Study on Six Classifier Fusion Strategies. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24(2):281–286, February 2002.

[74] 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.

[75] J. Lindberg, J.W. Koolwaaij, H.-P. Hutter, D. Genoud, M. Blomberg, J.-B. Pierrot, and F. Bimbot. Techniques for a priori Decision Threshold Estimation in Speaker Verification. In *Proc. of the Workshop R̈econnaissance du Locuteur et ses Applications Commerciales et Criminalistiques(̈RLA2C)*, pages 89–92, Avignon, 1998.

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

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

[78] J. Lüttin. Evaluation Protocol for the XM2FDB Database (Lausanne Protocol). Communication 98-05, IDIAP, Martigny, Switzerland, 1998.

[79] M.W. Mak. A Two-Level Fusion Approach to Multimodal Biometric Verification. In *IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, volume 5, pages 485–488, Philadelphia, 2005.

[80] Christine Marcel. Multimodal Identity Verification at IDIAP. Communication Report 03-04, IDIAP, Martigny, Switzerland, 2003.

[81] S. Marcel and S. Bengio. Improving Face Verification Using Skin Color Information. In *Proc. 16th Int. Conf. on Pattern Recognition*, page unknown, Quebec, 2002.

[82] 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.

[83] J. Matas, M. Hamouz, K. Jonsson, J. Kittler, Y. Li, C. Kotropoulos, A. Tefas, I. Pitas, T. Tan, H. Yan, F. Smeraldi, J. Begun, N. Capdevielle, W. Gerstner, S. Ben-Yacoub, Y. Abdeljaoued, and E. Mayoraz. Comparison of Face Verification Results on the XM2VTS Database. In *Proc. 15th Int'l Conf. Pattern Recognition*, volume 4, pages 858–863, Barcelona, 2000.

[84] C.E. Metz, B.A. Herman, and J-H. Shen. Maximum Likelihood Estimation of Receiver Operating Characteristic (ROC) Curves From Continuously-Distributed Data. *Statistics in Medicine*, 17(9):1033–1053, December 1998.

[85] 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.

[86] P. R. Morin and J-C. Junqua. A Voice-Centric Multimodal User Authentication System for Fast and Convenient Physical Access Control. In *Workshop on (Multimodal-User Authenticaion (MMUA)*, pages 19–24, 2003.

[87] A. C. Morris, M. P. Cooke, and P. D. Green. Some Solutions to the Missing Features Problem in Data Classification with Application to Noise Robust Automatic Speech Recognition. In *IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, pages 737–740, Seattle, 1998.

[88] K. Nandakumar. *Integration of M]ultiple Cues in Biometric Systems*. PhD thesis, Michigan State University, 2005.

[89] NIST. The 2005 NIST Speaker Recognition Evaluation, 2005, [Available] http://www.itl.nist.gov/iad/894.01/tests/spk/2005/.

[90] J. Ortega-Garcia, J. Fierrez-Aguilar, D. Simon, J. Gonzalez, M. Faundez-Zanuy, V. Espinosa, A. Satue, I. Hernaez, J.-J. Igarza, C. Vivaracho, D. Escudero, and Q.-I. Moro. Biometric on the Internet MCYT Baseline Corpus: a Bimodal Biometric Database. *IEE Proc. Visual Image Signal Processing*, 150(6):395–401, December 2003.

[91] K. K. Paliwal. Spectral Subband Centroids Features for Speech Recognition. In *Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages 617–620, Seattle, 1998.

[92] J-B Pierrot. *Elaboration et Validation d'Approaches en Vérification du Locuteur*. PhD thesis, ENST, Paris, September 1998.

[93] J.B. Pierrot, J. Lindberg, J.W. Koolwaaij, H.P. Hutter, D. Genoud, M. Blomberg, and F.Bimbot. A Comparison of *a priori* Threshold Setting Procedures for Speaker Verification in the CAVE Project. In *Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, pages 125–128, Seattle, 1998.

[94] S. Pigeon, P. Druyts, and P. Verlinde. Applying Logistic Regression to the Fusion of the NIST'99 1-Speaker Submissions. *Digital Signal Processing*, 10(1–3):237–248, 2000.

[95] N. Poh. Improving Biometric Authentication Using Score-Averaging and Error-Correction Output-Coding. DEA thesis (Diplôme d'Étude Approfondie), 2002.

[96] N. Poh. On Estimating Person-Dependent Biometric Authentication Performance Evolution Over Time. Research Report 06-25, IDIAP, Martigny, Switzerland, 2006.

[97] 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.

[98] N. Poh and S. Bengio. Evidences of Equal Error Rate Reduction in Biometric Authentication Fusion. IDIAP Research Report 43, IDIAP, 2004.

[99] N. Poh and S. Bengio. Improving Single Modal and Multimodal Biometric Authentication Using F-ratio Client Dependent Normalisation. Research Report 04-52, IDIAP, Martigny, Switzerland, 2004.

[100] N. Poh and S. Bengio. Noise-Robust Multi-Stream Fusion for Text-Independent Speaker Authentication. In *The Speaker and Language Recognition Workshop (Odyssey)*, pages 199–206, Toledo, 2004.

[101] N. Poh and S. Bengio. A Study of the Effects of Score Normalisation Prior to Fusion in Biometric Authentication Tasks. IDIAP Research Report 69, IDIAP, 2004.

[102] N. Poh and S. Bengio. Towards Predicting Optimal Subsets of Base-Experts in Biometric Authentication Task. In *LNCS 3361, 1st Joint AMI/PASCAL/IM2/M4 Workshop on Multimodal Interaction and Related Machine Learning Algorithms MLMI*, pages 159–172, Martigny, 2004.

[103] N. Poh and S. Bengio. Why Do Multi-Stream, Multi-Band and Multi-Modal Approaches Work on Biometric User Authentication Tasks? In *IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, pages vol. V, 893–896, Montreal, 2004.

[104] N. Poh and S. Bengio. Can Chimeric Persons Be Used in Multimodal Biometric Authentication Experiments? In *LNCS 3869, 2nd Joint AMI/PASCAL/IM2/M4 Workshop on Multimodal Interaction and Related Machine Learning Algorithms MLMI*, pages 87–100, Edinburgh, 2005.

[105] N. Poh and S. Bengio. Compensating User-Specific Information with User-Independent Information in Biometric Authentication Tasks. Research Report 05-44, IDIAP, Martigny, Switzerland, 2005.

[106] 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.

[107] N. Poh and S. Bengio. EER of Fixed and Trainable Classifiers: A Theoretical Study with Application to Biometric Authentication Tasks. In *LNCS 3541, Multiple Classifiers System (MCS)*, pages 74–85, Monterey Bay, 2005.

[108] N. Poh and S. Bengio. F-ratio Client-Dependent Normalisation on Biometric Authentication Tasks. In *IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, pages 721–724, Philadelphia, 2005.

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

[110] N. Poh and S. Bengio. Improving Fusion with Margin-Derived Confidence in Biometric Authentication Tasks. In *LNCS 3546, 5th Int'l. Conf. Audio- and Video-Based Biometric Person Authentication (AVBPA 2005)*, pages 474–483, New York, 2005.

[111] N. Poh and S. Bengio. A Novel Approach to Combining Client-Dependent and Confidence Information in Multimodal Biometric. In *LNCS 3546, 5th Int'l. Conf. Audio- and Video-Based Biometric Person Authentication (AVBPA 2005)*, pages 1120–1129, New York, 2005.

[112] N. Poh and S. Bengio. Revisiting Doddington's Zoo: Employing User-Dependent Performance Criterion For Multibiometric Fusion. Researh Report 06-04, IDIAP, Martigny, Switzerland, 2006.

[113] N. Poh and S. Bengio. Using Chimeric Users to Construct Fusion Classifiers in Biometric Authentication Tasks: An Investigation. In *IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1077–1080, Toulouse, 2006.

[114] N. Poh, S. Bengio, and J. Korczak. A Multi-Sample Multi-source Model for Biometric Authentication. In *IEEE International Workshop on Neural Networks for Signal Processing (NNSP)*, pages 275–284, Martigny, 2002.

[115] N. Poh, S. Bengio, and A. Ross. Revisiting Doddington's Zoo: A Systematic Method to Assess User-Dependent Variabilities. In *Workshop on Multimodal User Authentication (MMUA 2006)*, Toulouse, 2006.

[116] N. Poh, S. Marcel, and S. Bengio. Improving Face Authetication Using Virtual Samples. In *IEEE Int'l Conf. Acoustics, Speech, and Signal Processing*, pages 233–236 (Vol. 3), Hong Kong, 2003.

[117] N. Poh, A. Martin, and S. Bengio. Performance Generalization in Biometric Authentication Using Joint User-Specific and Sample Bootstraps. IDIAP-RR 60, IDIAP, Martigny, 2005. accepted for publication in IEEE Trans. Pattern Analysis and Machine Intelligence.

[118] N. Poh, C. Sanderson, and S. Bengio. An Investigation of Spectral Subband Centroids For Speaker Authentication. In *LNCS 3072, Int'l Conf. on Biometric Authentication (ICBA)*, pages 631–639, Hong Kong, 2004.

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

[120] C.R. Rao. *Linear Statistical Inference and Its Applications (2nd Edition)*. John Wiley & Sons, 1973.

[121] D. A. Reynolds. Automatic Speaker Recognition using Gaussian Mixture Speaker Models. *The Lincoln Laboratory Journal*, 8(2):173–192, 1995.

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

[123] F. Roli and G. Fumera. A Theoretical and Experimental Analysis of Linear Combiners for Multiple Classifier Systems. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 27(6):942–956, 2005.

[124] A. Ross, A. Jain, and J-Z. Qian. Information Fusion in Biometrics. *Pattern Recognition Letter*, 24(13):2115–2125, September 2003.

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

[126] J.R. Saeta and J. Hernando. On the Use of Score Pruning in Speaker Verification for Speaker Dependent Threshold Estimation. In *The Speaker and Language Recognition Workshop (Odyssey)*, pages 215–218, Toledo, 2004.

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

[128] C. Sanderson, S. Bengio, and Y. Gao. On Transforming Statistical Models for Non-Frontal Face Verication. *Pattern Recognition*, 39(2):288–302, 2006.

[129] C. Sanderson and K. K. Paliwal. Information Fusion and Person Verification using Speech and Face Information. IDIAP Research Report 22, IDIAP, 2002.

[130] C. Sanderson and K.K. Paliwal. Likelihood Normalization for Face Authentication in Variable Recording Conditions. In *Int. Conf. on Image Processing*, pages 301–304, New York, 2002.

[131] C. Sanderson and K.K. Paliwal. Fast Features for Face Authentication Under Illumination Direction Changes. *Pattern Recognition Letters*, 24(14):2409–2419, 2003.

[132] C. Sanderson and K.K. Paliwal. Identity Verification using Speech and Face Information. *Digital Signal Processing*, 14(5):449–480, 2004.

[133] S. Sharma, H. Hermansky, and P. Vermuulen. Combining Information from Multiple Classifiers for Speaker Verification. In *Proc. Speaker Recogniton and Its Commercial and Forestic Applications Workshop (RLA2C)*, pages 115–119, Avignon, 1998.

[134] L.L. Shen, A. Kot, and W.M. Koo. Quality Measures of Fingerprint Images. In *3rd Int. Conf. Audio-Visual Biometric Person Authentication (AVBPA 2001)*, pages 266–271, 2001.

[135] C.A. Shipp and L.I. Kuncheva. Realtionships Between Combination Methods and Measures of Diversity in Combining Classifiers. *Information Fusion*, 3:135–148, 2002.

[136] 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.

[137] R. Snelick, U. Uludag, A. Mink, M. Indovina, and A. Jain. Large Scale Evaluation of Multimodal Biometric Authentication Using State-of-the-Art Systems. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 27(3):450–455, 2005.

[138] 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.

[139] K.-A. Toh, X. Jiang, and W.-Y. Yau. Exploiting Global and Local Decision for Multimodal Biometrics Verification. *IEEE Trans. on Signal Processing*, 52(10):3059–3072, October 2004.

[140] K.-A. Toh, W.-Y. Yau, and X. Jiang. A Reduced Multivariate Polynomial Model For Multimodal Biometrics And Classifiers Fusion. *IEEE Trans. Circuits and Systems for Video Technology (Special Issue on Image- and Video-Based Biometrics)*, 14(2):224–233, 2004.

[141] K-A. Toh, W-Y. Yau, E. Lim, L. Chen, and C-H. Ng. Fusion of Auxiliary Information for Multimodal Biometric Authentication. In *LNCS 3072, Int'l Conf. on Biometric Authentication (ICBA)*, pages 678–685, Hong Kong, 2004.

[142] K. Tumer and J. Ghosh. Linear and Order Statistics Combiners for Pattern Classification. In *Combining Artificial Neural Nets, Ensemble and Modular Multi-Net Systems*, pages 127–157, Berlin, 1996. Springer.

[143] K. Tumer and J. Ghosh. Robust Combining of Disparate Classifiers through Order Statistics. *Pattern Analysis and Applications*, 5:189–200, 2002.

[144] N. Ueda and R. Nakano. Generalisation Error of Ensemble Estimators. In *Proc. Int'l conf. on Neural Networks*, pages 90–95, 1990.

[145] I. Ulusoy and C.M.Bishop. Generative Versus Discriminative Models for Object Recognition. In *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR)*, San Diego, 2005.

[146] V. N. Vapnik. *Statistical Learning Theory*. Springer, 1998.

[147] H. Wang and P. J. Flynn. Experimental Evaluation of Eye Location Accuracies and Time-Lapse Effects on Face Recognition Systems. In *LNCS 3546, 5th Int'l. Conf. Audio- and Video-Based Biometric Person Authentication (AVBPA 2005)*, pages 627–636, New York, 2005.

[148] J. Wayman, A. Jain, D. Maltoni, and D. Maio. *Biometric Systems: Technology, Design and Performance Evaluation*. Springer, 2005.

[149] J.L. Wayman. A Path Forward for Multi-biometrics. In *IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1069–10720, Toulouse, 2006.

[150] D. Wolpert. Stacked Generalization. *Neural Networks*, 5(2):241–286, 1992.

[151] Xiaojun Wu, Kittler Josef, Jingyu Yang, Messer Kieron, Shitong Wang, and Jieping Lu. On Dimensionality Reduction for Client Specific Discriminant Analysis with Application to Face Verification. In *LNCS 3338, Advances in Biometric Person Authentication: 5th Chinese Conference on Biometric Recognition, SINOBIOMETRICS*, pages 305–312, Guangzhou, 2004.

[152] J. You, W. k. Kong, D. Zhang, and K. H. Cheung. A New Approach to Perosonal Identification in Large Databases by Hierarchical Palmprint Coding with Multi-Features. In *LNCS 3072, Int'l Conf. on Biometric Authentication (ICBA)*, pages 739–745, Hong Kong, 2004.

Norman Poh

Permanent address:	115 Sin Lian Hin Garden, Jalan Nenas Barat, 93150 Kuching, Sarawak, Malaysia.	Phone:	(+41)079.333.69.58
Current address:	IDIAP, Rue de Simplon 4, Martigny CH-1920, Switzerland	Email:	normanpoh@gmail.com
30 years old	Married (two children) Malaysian	Weblog:	http://www.idiap.ch/∼norman

Fields of Interest

pattern recognition, multiple classifier system, biometric authentication, signal processing

Education

2002–2006	Docteur ès Sciences in the field of *Multimodal Biometrics* École Polytechnique Fédérale de Lausanne
2001–2002	Diplôme d'Etudes Approfondies (DEA) degree in Computer Science Université Louis Pasteur (ULP)
1999–2001	Masters in *Artificial Intelligence*, Computer Science Univeristy Science of Malaysia (USM)
1996-1999	Bachelor Degree in Computer Science with Honours majoring in *Software Engineering* and minoring in *Management* Univeristy Science of Malaysia (USM)

Professional Experience

2002–2006	IDIAP Research Institute, Martigny, Switzerland Research assistant on biometric authentication
1999–2001	LSIIT laboratory, ULP, Strasbourg, France Built a face and speech bimodal biometric person authentication for a PC-based system
1998	ISIT laboratory, USM, Penang, Malaysia Built a Smart School automated scheduling system using graph-coloring algorithm
1997	ISIT laboratory, USM, Penang, Malaysia Built a Marketing Research System that maps consumers' traits with their products Designed an infix mathematical formula-parsing algorithm for an Intelligent Computer Aided Instruction system

International Experience

Oct–Nov 2005	Visiting scholar at the PRIB Lab, **Michigan State University (MSU)**, USA, with Prof. Anil K. Jain
July 2005	Visiting scholar at the Speech group of **Nat'l Institute of Standards and Technology (NIST)**, Gaithersburg, USA, with Dr. Alvin Martin
July 2004	Visiting scholar at the Biometric Lab of **Hong Kong Polytechnical University (HKPolyU)**, with Prof. David Zhang and Dr. Ajay Kumar
Jan-May 1998	Exchange student at **Nanyang Techlogical University (NTU)** (Singapore)
July 1994	Malaysian Delegate to the Asian Session and Councils of Young Christian Students, Manila

Languages

English	fluent
Malay	fluent
French	read and converse well
Chinese	native
Chinese dialects	converse well in Theo Chew, Hai Nam, Hokkien

Awards

- Best student poster award in Int'l. Conf. Audio- and Video-Based Biometric Person Authentication (AVBPA 2005), in recognition of contribution on "biometric fusion"
- Recipient of the RLKA academic fellowship (USM) of the year 1999
- Best student award (Pelajar Mithali) of the academic year 1998/99 in the School of Computer Science (Bachelor degree), USM

Clubs and Voluntary Activities

2006–	Reviewer of IEEE Trans. Pattern Analysis and Machine Intelligence
2005–	Webmaster of Fully Communal Monthly Newsletter association
2005–	Member of Swiss Association for Pattern Recognition
2002–	Member of IEEE Switzerland
1998–1999	Vice president of the Catholic Undergraduates Society
1996–1997	Treasurer of Student Association PERKASA (Perkumpulan Anak-anak Sarawak)
1996–1997	Assistant Secretary General of St. John Ambulance, university division
1994–1995	President of Young Christian Student
1994–1995	Senior Prefect of the School Prefectorial Board
1993-1994	Peer Counselor
1992–1993	Secretary of St. John Ambulance

Computer Experience

Operating Systems	Linux, UNIX, Windows
Languages	C, C++, Java, Matlab, Perl, PHP, Visual Basic, Prolog, Lisp, Power Builder

Projects

IM2.ACP	Multimedia Information Access and Content Protection Funding: (IM)2 Swiss National Center of Competence in Research Duration: July 2002 – June 2004 Task: Research on fusion techniques in multimodal biometrics
PASCAL	Pattern Analysis, Statistical Modelling and Computational Learning Funding: European Network of Excellence, 6th Framework Programme, Information Society Technology, supported by OFES Duration: January 2004 – December 2007 Task: Research on user-specific processing in multimodal biometrics

Publications

Book Chapters and Theses

- N. Poh and J. Korczak Automated Authentication using Hybrid Biometric System, Chapter 16, *Biometric Authentication in the e-World*, Kluwer Academic Publishers, edited by Prof. David Zhang, 2003.
- N. Poh, Improving Biometric Authentication using Score-Averaging and Error-Correction Output-Coding, *DEA Thesis*, Université Louis Pasteur, Strasbourg, France, 2002
- N. Poh, Biometric Authentication System, *Masters Thesis*, University Science of Malaysia, Penang, Malaysia

Journal Publications

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

Conference and Workshop Publications

- N. Poh, S. Bengio, and A. Ross. Revisiting Doddington's Zoo: A Systematic Method to Assess User-Dependent Variabilities. In *Workshop on Multimodal User Authentication (MMUA 2006)*, Toulouse, 2006.
- N. Poh and S. Bengio. Using Chimeric Users to Construct Fusion Classifiers in Biometric Authentication Tasks: An Investigation. In *IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Toulouse, 2006.
- N. Poh and S. Bengio. Can Chimeric Persons Be Used in Multimodal Biometric Authentication Experiments? In *LNCS 3869, 2nd Joint AMI/PASCAL/IM2/M4 Workshop on Multimodal Interaction and Related Machine Learning Algorithms MLMI*, pages 87–100, Edinburgh, 2005.
- N. Poh and S. Bengio. EER of Fixed and Trainable Classifiers: A Theoretical Study with Application to Biometric Authentication Tasks. In *LNCS 3541, Multiple Classifiers System (MCS)*, pages 74–85, Monterey Bay, 2005.
- N. Poh and S. Bengio. F-ratio Client-Dependent Normalization on Biometric Authentication Tasks. In *IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, pages 721–724, Philadelphia, 2005.
- N. Poh and S. Bengio. Improving Fusion with Margin-Derived Confidence in Biometric Authentication Tasks. In *LNCS 3546, 5th Int'l. Conf. Audio- and Video-Based Biometric Person Authentication (AVBPA)*, pages 474–483, New York, 2005.
- N. Poh and S. Bengio. Towards Predicting Optimal Subsets of Base-Experts in Biometric Authentication Task. In *LNCS 3361, 1st Joint AMI/PASCAL/IM2/M4 Workshop on Multimodal Interaction and Related Machine Learning Algorithms MLMI*, pages 159–172, Martigny, 2004.
- N. Poh and S. Bengio. Why Do Multi-Stream, Multi-Band and Multi-Modal Approaches Work on Biometric User Authentication Tasks? In *IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, pages vol. V, 893–896, Montreal, 2004.

- N. Poh and S. Bengio. Noise-Robust Multi-Stream Fusion for Text-Independent Speaker Authentication. In *The Speaker and Language Recognition Workshop (Odyssey)*, pages 199–206, Toledo, 2004.
- N. Poh, C. Sanderson, and S. Bengio. An Investigation of Spectral Subband Centroids For Speaker Authentication. In *LNCS 3072, Int'l Conf. on Biometric Authentication (ICBA)*, pages 631–639, Hong Kong, 2004.
- 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 Int'l Conf. Biometric Authentication (ICBA), pages 6-15, Hong Kong, 2004.
- 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), vol. 4, pp. 523-532. IEEE Press, Cambridge, 2004.
- N. Poh and S. Bengio. A Novel Approach to Combining Client-Dependent and Confidence Information in Multimodal Biometric. In *LNCS 3546, 5th Int'l. Conf. Audio- and Video-Based Biometric Person Authentication (AVBPA 2003)*, pages 1120–1129, New York, 2005 ((winner of the Best Student Poster award)).
- 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.
- N. Poh, S. Bengio, and J. Korczak. A Multi-Sample Multi-source Model for Biometric Authentication. In *IEEE International Workshop on Neural Networks for Signal Processing (NNSP)*, pages 275–284, Martigny, 2002.

Technical Reports

- N. Poh and S. Bengio. A Study of the Effects of Score Normalisation Prior to Fusion in Biometric Authentication Tasks. IDIAP Research Report 69, IDIAP, 2004.
- N. Poh and S. Bengio. Improving Single Modal and Multimodal Biometric Authentication Using F-ratio Client Dependent Normalisation. Research Report 04-52, IDIAP, Martigny, Switzerland, 2004.
- N. Poh and S. Bengio. Compensating User-Specific Information with User-Independent Information in Biometric Authentication Tasks. Research Report 05-44, IDIAP, Martigny, Switzerland, 2005.
- N. Poh, A. Martin, and S. Bengio. Performance Generalization in Biometric Authentication Using Joint User-Specific and Sample Bootstraps. IDIAP-RR 60, IDIAP, Martigny, 2005.
- N. Poh. On Estimating Person-Dependent Biometric Authentication Performance Evolution Over Time. Research Report 06-25, IDIAP, Martigny, Switzerland, 2006.