MAKING BETTER BIOMETRIC DECISIONS WITH QUALITY AND COHORT INFORMATION: A CASE STUDY IN FINGERPRINT VERIFICATION

Norman Poh, Amin Merati and Josef Kittler

CVSSP/Department of Electrical Engineering, University of Surrey
University of Surrey, Guildford, Surrey, UK, GU2 7XH
emails: normanpoh@ieee.org, a.merati@surrey.ac.uk, j.kittler@surrey.ac.uk

ABSTRACT

Automatically recognizing humans using their biometric traits such as face and fingerprint will have very important implications in our daily lives. This problem is challenging because biometric traits can be affected by the acquisition process which is sensitive to the environmental conditions (e.g., lighting) and the user interaction. It has been shown that post-processing the classifier output, so called *score normalization*, is an important mechanism to counteract the above problem. In the literature, two dominant research directions have been explored: cohort normalization and quality-based normalization. The first approach relies on a set of competing cohort models, essentially making use of the resultant cohort scores. A well-established example is the T-norm. In the second approach, the normalization is based on deriving the quality information from the raw biometric signal. We propose to combine both the cohort score- and signal-derived information via logistic regression. Based on 12 independent fingerprint experiments, our proposal is found to be significantly better than the T-norm and two recently proposed cohort-based normalization methods.

1. INTRODUCTION

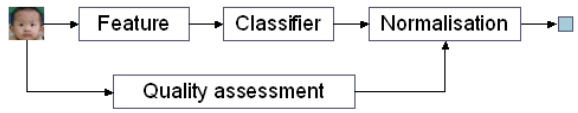
Automatic person identity recognition using biometrics is a challenging problem [1] despite the decades of research. One of the reasons is that raw biometric signals are susceptible to various forms of degradation caused by, the acquisition environment, the manner a subject interacts with a biometric acquisition device, and natural alteration of biometric trait due to sickness or aging. Because a matcher or classifier cannot adequately cope with all of the above corrupting effects, post-processing the resulting match scores, i.e. *score normalization*, has been identified as an important stage. The goal of score normalization is to map the original match scores to a domain where the effect of noise on match score distribution is neutralized.

Biometric match scores can be normalized using one the following approaches:

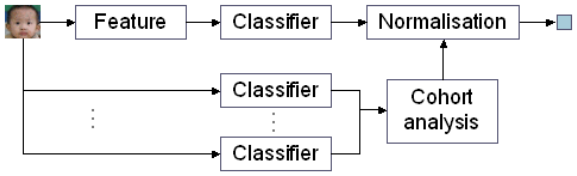
- **biometric trait quality**: Variation in the raw biometric signal can be gauged by directly measuring the signal quality. Historically, this quality information is used in the context of multimodal fusion where the idea is to weigh the biometric modality heavier in the process of computing the final combined score [2, 3, 4, 5, 6]. However, recent studies have shown that the quality information can also be used in unimodal biometrics, e.g., [7, 8]. In fact, it is natural to handle raw signal quality variation for each biometric modality separately because this variation is modality-dependent [9]. For instance, the noise affecting the fingerprint signal is different from that affecting the face signal. Some quality-based fusion can be factorized into two stages wherein the first stage essentially consists of modality-dependent quality-based score normalization and only then the second stage considers multimodal fusion. Examples of such fusion algorithms are [10] using a Bayesian classifier with Gaussian Mixture Model as a density estimator and [4] using logistic regression and support vector machines. In [4], various architectures considering both intramodal and multimodal quality-based fusion are compared.
- **decision score characteristic**: An alternative method that can be used to normalize against score distribution variation between the enrollment and query sessions is based either on the reliability of the decision score or on measuring the degradation effect in relation to a reference cohort of users (competing hypotheses). The motivation for the latter is that all competing client models (including the claimed user model) will be subject to the *same* degradation. It is therefore sensible to normalize the match score using a pool of user models, also known as the *cohort models*. The class of such normalization procedures (with T-norm being a special case[11]), is subsequently called *cohort-based score normalization* in this paper. A disadvantage of this normalization is that for each query, all the cohort models have to be used. When the cohort models used are the models in the gallery (also known as enrollee or client models) other than the claimed model, one effectively performs *identification* in the verification mode. The former is a many-to-one matching whereas the latter is a one-to-one matching.

Two diagrams depicting quality-based and cohort-based score normalization are shown in Figure 1. In (a), the quality assessment module directly provides an estimate of the signal quality known as a quality measure. This measure quantifies the degree of excellence or conformance of biometric samples to some predefined criteria known to influence the system performance. For instance, for the face biometrics, these measures assess image focus, contrast and face detection reliability. Since any quality measure is derived *separately* from the classifier, there is no guarantee that the quality measures will correlate with the performance. The task of selecting/designing an appropriate quality assessment module is left to the system designer.

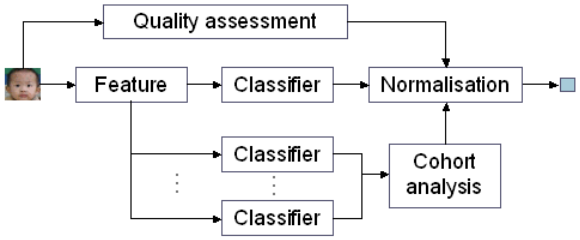
Figure 1(b) shows a cohort-based score normalization procedure. This procedure requires a set of cohort models, shown here as a pool of classifiers, each taking the query biometric sample/features as input and outputting a match score. The resulting set of match scores are submitted to a cohort-analysis module, to calculate the maximum/minimum scores as well as the first and second order moments, for instance. The derived parameters are then used as input to a so-called *cohort-based* score normalization procedure.



(a) quality-based normalization



(b) cohort-based normalization



(c) our proposal

Figure 1: Block diagram of different normalizations method: (a) Quality based normalization, (b) Cohort based normalization, (c) Our proposal of combined Quality and Cohort based normalization.

An important distinguishing characteristic of the two normalization procedures is that in the cohort-based score normalization, only the classifier outputs are used. As a result, there is an inherent *tight coupling* between the score-derived parameters and the system performance. In contrast, in quality-based normalization, this relationship is not guaranteed; additional knowledge from the system designer is needed. We shall therefore describe the cohort-based normalization as *score-derived* quality normalization as opposed to the quality-based normalization which is *signal-derived*. Recognizing that both approaches effectively capture different pieces of information, we propose to combine them (see Figure 1(c)). The role of normalization in this case is to consolidate both pieces of information. Our experiments based on 6 fingers and two acquisition devices show that our proposal is *significantly* superior to the well established T-norm, as well as two recently reported cohort normalization methods in the literature [11, 12].

This paper is organized as follow: Section 2 presents prior works of both score- and signal-derived score normalization procedures. Our proposal is presented in Section 3. This is followed by experimental validation in Section 4 and conclusions in Section 5.

2. RELATED PRIOR WORK

2.1 Signal-derived Normalization

recent advances in multi-biometrics include quality-based fusion, e.g., [2, 3, 4, 5, 6], where the quality associated with the template (or model) as well as the query biometric sample are taken into account in fusion. For this purpose, a plethora of quality measures have recently been proposed in the literature for various biometric modalities, e.g., fingerprint [13, 14], iris [7], face [15], speech [16] and signature [17].

There are two ways quality measures can be incorporated into a fusion classifier, depending on their role, i.e., either as a control parameter or as an evidence. In their primary role, quality measures are used to modify the way a fusion classifier is trained or tested, as suggested in the Bayesian-based classifier called "expert conciliation" [2], reduced polynomial classifier [18], quality-controlled support vector machines [3], and quality-based fixed rule fusion [19]. Alternatively quality measures are often concatenated with the expert outputs to be fed to a fusion classifier. This role lends itself to quality-based score normalization, since the resulting fusion module can be decoupled to process one modality at a time. Examples in this category are [20, 5, 8, 4, 21]. Among these approaches, we shall use the one described in [4] for its simplicity.

Let y be the match score and q be the vector of quality measures derived from the training data (template) as well as the query data. Kittler *et al.* [4] proposed to use discriminative classifiers such as logistic regression and support vector machines to combine y and q. In the case of logistic regression, the resultant classifier score, which can be viewed as a normalization score, can be written as:

$$y_q = P(\mathtt{C}|y,q) \tag{1}$$

Our proposal will be an extension of this normalization method by augmenting the observation space with the cohort information, to be described in Section 3.

2.2 Cohort-based Normalization

Furui [22] described the cohort-based score normalization as a likelihood ratio-based normalization since the expected (mean) score derived from the cohort models can be interpreted as a competing hypothesis. A variation to the above approach includes the claimed model in the pool of cohort models (which the author call posterior probability-based normalization to contrast with the original approach). Both approaches are shown to perform equally well.

While T-norm has been a dominant approach since its proposal, attempts have been made to improve its efficiency, e.g., [23]. Three notable works pursuing the latter direction include [24, 12] and [11].

Let $y^c \in \mathscr{Y}^c$ be a cohort score obtained by comparing a query sample with a cohort model, and $\mathscr{Y}^c$ be a *set* of cohort scores. Note that y is different from y^c because y is a result of comparing a query sample with a *claimed model*, whereas y^c is the result of comparing the sample with a *cohort model*. Furthermore, we define the expected value (mean) of y^c as $\mu^c = \mathtt{E}[y]$ and its variance as $(\sigma^c)^2 = \mathtt{E}\left[(y^c - \mu^c)^2\right]$. In this notation, the T-norm is given by:

$$y_T = \frac{y - \mu^c}{\sigma^c} \tag{2}$$

In [24], a generalized form of T-norm is given by:

$$y_X = y + \frac{(y - \mu^c)^2}{2(\sigma^c)^2}$$

It was reported that the original and generalized form of the T-norm are not significantly different in performance. This can be expected since both formulations use the same information.

Tulyakov *et al.* [11] proposed a different form, which can be described by

$$y_{Tul} = P(\mathtt{C}|y, \max_{y^c \in \mathscr{Y}^c}\{y^c\}), \quad (3)$$

where the posterior probability is obtained via multi-layer Perceptrons using $[y, \max_{y^c \in \mathscr{Y}^c}\{y^c\}]$ as observations. In their work, a generative classifier based on the log-likelihood ratio test was also examined. The term $\max_{y^c \in \mathscr{Y}^c}\{y^c\}$ is the maximum of the impostor scores, and is referred to as the "second best score" by the authors. The "best score" would be y (the score of the claimed identity)[1].

Aggarwal *et al.* [12] proposed the following normalization:

$$y_{Ag} = \frac{y}{\max_{y^c \in \mathscr{Y}^c}\{y^c\}} \quad (4)$$

The advantage of this approach is that no additional training is required, as compared to (3). However, designing a score normalization that is tailored to a specific system output can potentially increase the generalization performance further, hence justifying this approach.

We note that if the scores y and y^c are interpreted as likelihoods, the ratio of (4) is simply a likelihood ratio test, a concept already noted by Furui [22]. However, rather than taking the average, here the maximum operator is used. The motivation here is that only the strongest cohort model (that is the *most* similar to the claimed model) will be used. Following this intuition, Aggarwal *et al.* proposed to use only a subset of cohort models pre-selected for each claimed model to improve the algorithm efficiency without hampering its effectiveness.

3. METHODOLOGY

We propose a novel score normalization method combining the score- and signal-derived quality information. As we shall see in the Experimental Section T-normalization is superior in performance to methods (2) and (3). We shall therefore use it as a basis for combination with the biometric trait quality information to construct a new normalization method:

$$y_p = P(\mathtt{C}|y^T_{norm}, q) \quad (5)$$

In (5) the weights of the two pieces of inherently different information are obtained via training. The logistic regression algorithm optimizes the likelihood of the model given the data, following the usual formulation of the maximum likelihood principle [25]. However, since the model is discriminative, the maximum likelihood solution (unlike the usual generative models) in this case maximizes the class-separability between the genuine and impostor classes. In other words, the feature observations are weighted by a discriminative criterion, hence, justifying our choice of using logistic regression.

A scatter plot of match scores of $[y_{T-norm}, q]$ for a finger print data set is shown in Figure 2. We observe that both features provide *complementary* information for discriminating the genuine accesses from the impostor ones.

[1]It is not clear what the second best scores would mean in the case where the cohort score set contains the genuine score. This is the case when one actually performs verification in the identification mode. We shall limit ourselves to the case where the pool of cohort models and the gallery of client models are *disjoint*, such that $\mathscr{Y}^c$ contains no genuine match score.

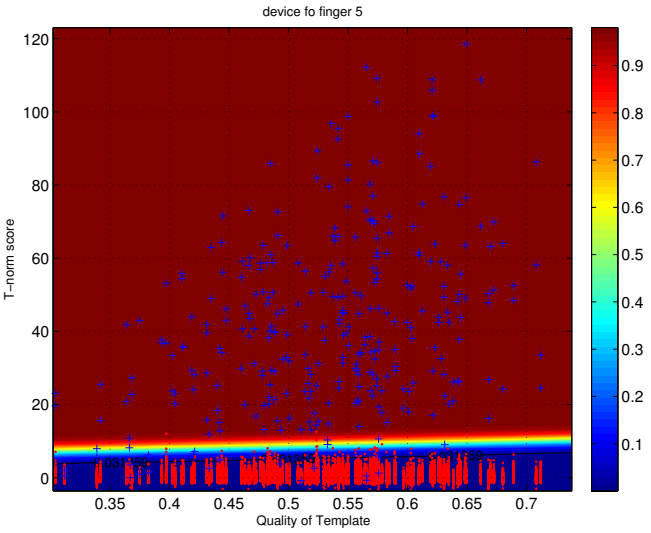


Figure 2: Scatter plot of the T-normalized score (Y-axis) vs the template quality measure of fingerprint. The inferred posterior are shown in the background, with the optimal decision boundary being 0.5.

4. EXPERIMENTAL RESULTS

4.1 Database and matching algorithms

In order to demonstrate the efficiency of our method, we used a subset of the Biosecure data set [26], limiting to the study to the fingerprint modality. We used the NIST fingerprint matcher ("Bozorth3"). This software also has a quality assessment module called "NFIQ" (NIST's Fingerprint Imaging Quality). In this database, six fingers – thumb, middle and index fingers of both hands were used. The fingerprints were scanned with two devices, namely thermal and optical-based sensors. Each subject provides 4 impressions per device and per finger. Therefore, each subject supplied a total of 4 impressions $\times$ 6 fingers $\times$ 2 devices $=$ 48 impressions.

Five *disjoint* groups of users were identified, with the first four groups (respectively referred to as g1–g4) constituting enrollees and the final group forming a separate set of cohort users to provide cohort models. Subjects in g1 and g2 were used as enrollees in the *development* (*dev*) set; and, g3 and g4 as enrollees in the *evaluation* (*eva*) set. The total number of subjects in g1–g4 are $\{84, 83, 83, 81\}$ respectively. The total number of cohort users is 84. For the purpose of obtaining a cohort score, only the first of the four samples of a cohort user was used.

We require that each of *dev* and *eva* sets has its own enrollment and query data set, i.e.,$\mathscr{D}_{d,enrol}$, $\mathscr{D}_{d,query}$ for $d \in \{dev, eva\}$. Recall that there are four impressions per finger and per device. The first impression was used as a template (or model). The second impression was used to generate a genuine score for $\mathscr{D}_{dev,enrol}$ whereas the remaining two query samples were used for $\mathscr{D}_{eva,query}$ (for each client/enrollee).

To generate the impostor match scores, for,$\mathscr{D}_{dev,enrol}$ we used query samples of $g3$; for $\mathscr{D}_{dev,query}$, $g4$; for $\mathscr{D}_{eva,enrol}$, $g1$; and, for $\mathscr{D}_{dev,query}$, $g2$. In this way, the impostor match scores in *all* the four data sets are completely disjoint. This has the advantage that algorithms (operating at the score level) trained will not have seen the impostor match scores presented during testing.

In the empirical evaluation reported, in the next section we use $\mathscr{D}_{dev,query}$ as our training set and $\mathscr{D}_{eva,query}$ as our test set. Note that the enrollees and impostor subjects in these two match scores are completely *disjoint*. This simulates a scenario where the development and operational data have disjoint subjects, a realistic condition.

4.2 Results

We compared the following approaches:

- **Baseline**: the original system output without any post-processing
- **T-norm**: post-processing using T-norm, as in (2)
- **Tulyakov's approach**: post-processing using (3) (labeled as "SecondBestScore")
- **Aggarwal's approach**: post-processing using (4) (labeled as "src/maxOfCohort")
- **Quality-based approach**: Combination of score with biometric trait quality using (1) (labeled as "[baseline SigQlty]")
- **Our proposal**: post-processing using (5) (labeled as "[Tnorm, SigQlty]")

Since there are 12 independent experiments (due to 2 devices $\times$ 6 fingers), we shall summarize the results using relative change of Equal Error Rate (EER) with respect to the performance of the baseline system:

$$\text{rel. change of EER} = \frac{\text{EER}_{algo} - \text{EER}_{baseline}}{\text{EER}_{baseline}}$$

A negative change of EER implies an improvement over the baseline system. This statistic has the advantage that the performance of different systems can be collated, hence establishing confidence intervals when visualized using a boxplot (showing median, the first and third quarter, as well as the fifth and 95-th percentiles of the data) when several *independent* experiments are conducted.

The relative changes of EERs for the above mentioned algorithms are shown in Figure 3. As can be observed, our proposal outperforms all the competing algorithms, including the T-norm. Our results, obtained on a more extensive set of experiments, do not support the claim of Tulyakov *et al.* that their approach is better than the T-norm. Aggarwal *et al.* did not compare their proposal with the T-norm as done here.

Acknowledging that EER is not the only point of interest in many applications, we also plot the more conventional detection error trade-off (DET) curve for one of the 12 experiments selected at random in Figure 4. The result is consistent with Figure 3.

5. CONCLUSION

In this paper, we have proposed a novel score normalization method combining both the signal- and score-derived quality information. By exploiting these two pieces of complementary information, our approach significantly improves over the T-norm and two recently reported methods.

Future works in this direction will investigate (i) the effect of our proposed score normalization on fusion, (ii) including genuine score information, (iii) the choice of cohort users; and (iv) the generality of our proposal on other biometric modalities.

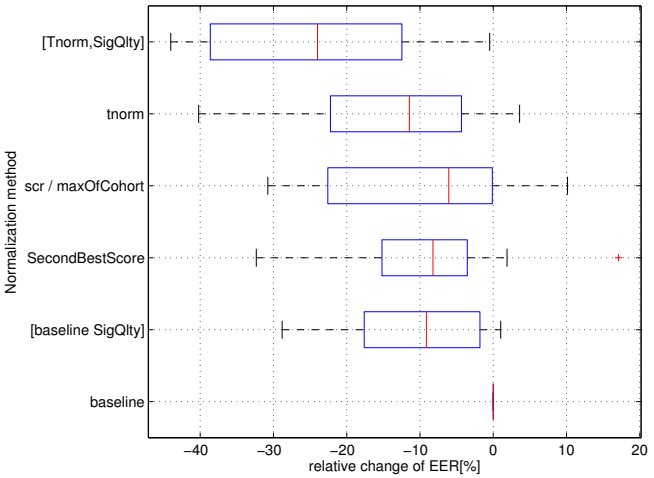


Figure 3: Boxplot of relative change of EER.

Acknowledgment

This work was supported partially by the advanced researcher fellowship PA0022_121477 of the Swiss National Science Foundation and by the EU-funded Mobio project (www.mobioproject.org) grant IST-214324.

REFERENCES

[1] A.K. Jain, S. Pankanti, S. Prabhakar, L. Hong, and A. Ross, "Biometrics: A grand challenge," in *Proc. 17th Int'l Conf. Pattern Recognition (ICPR)*, 2004, pp. II: 935–942.

[2] 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*, Mantova, 2003, pp. 2–13.

[3] 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*, 2004, vol. 5404, pp. 544–554.

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

[5] K. Nandakumar, Y. Chen, S. C. Dass, and A. K. Jain, "Likelihood ratio based biometric score fusion," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 30, pp. 342–347, 2008.

[6] K. Kryszczuk and A. Drygajlo, "Credence estimation and error prediction in biometric identity verification," *Signal Processing*, vol. 88, pp. 916–925, 2008.

[7] Y. Chen, S. Dass, and A. Jain, "Localized iris image quality using 2-d wavelets," in *Proc. Int'l Conf. on Biometrics (ICB)*, Hong Kong, 2006, pp. 373–381.

[8] N. Poh, G. Heusch, and J. Kittler, "On Combination of Face Authentication Experts by a Mixture of Quality

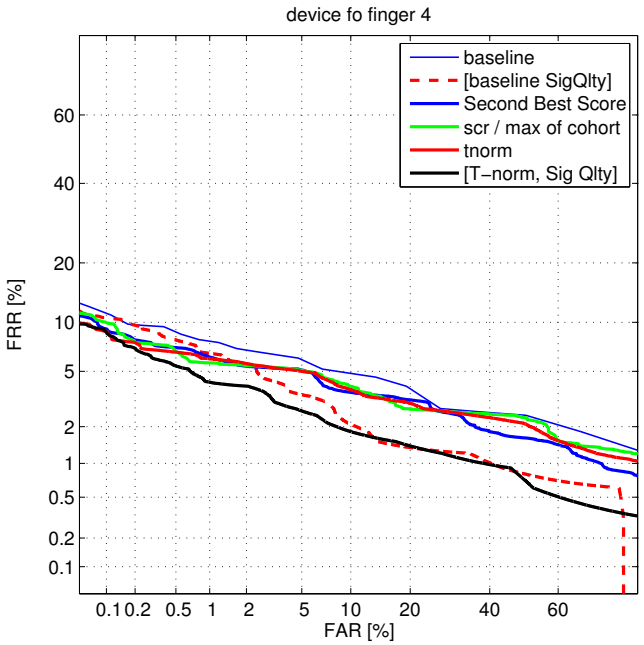


Figure 4: DET curves of different normalization methods for left thumb.

Dependent Fusion Classifiers," in *LNCS 4472, Multiple Classifiers System (MCS)*, Prague, 2007, pp. 344–356.

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

[10] K. Nandakumar, Y. Chen, S.C. Dass, and A.K. Jain, "Quality-based Score Level Fusion in Multibiometric Systems," in *Proc. 18th Int'l Conf. Pattern Recognition (ICPR)*, Hong Kong, 2006, pp. 473–476.

[11] S. Tulyakov, Z. Zhang, and V. Govindaraju, "Comparison of combination methods utilizing t-normalization and second best score model," in *IEEE Conf. on Computer Vision and Pattern Recognition Workshop*, 2008.

[12] G. Aggarwal, N.K. Ratha, R.M Bolle, and R. Chellappa, "Multi-biometric cohort analysis for biometric fusion," in *IEEE Int'l Conf. on Acoustics, Speech and Signal Processing*, 2008.

[13] H. Fronthaler, K. Kollreider, J. Bigun, J. Fierrez, F. Alonso-Fernandez, J. Ortega-Garcia, and J. Gonzalez-Rodriguez, "Fingerprint image-quality estimation and its application to multialgorithm verification," *IEEE Trans. on Information Forensics and Security*, vol. 3, pp. 331–338, 2008.

[14] 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)*, New York, 2005, pp. 160–170.

[15] X. Gao, R. Liu, S. Z. Li, and P. Zhang, "Standardization of face image sample quality," in *LNCS 4642, Proc. Int'l Conf. Biometrics (ICB'07)*, Seoul, 2007, pp. 242–251.

[16] National Institute of Standards and Technology, "Nist speech quality assurance package 2.3 documentation," .

[17] S. Muller and O. Henniger, "Evaluating the biometric sample quality of handwritten signatures," in *LNCS 3832, Proc. Int'l Conf. Biometrics (ICB'07)*, 2007, pp. 407–414.

[18] 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)*, Hong Kong, 2004, pp. 678–685.

[19] O. Fatukasi, J. Kittler, and N. Poh, "Quality Controlled Multimodal Fusion of Biometric Experts," in *12th Iberoamerican Congress on Pattern Recognition CIARP*, Via del Mar-Valparaiso, Chile, 2007, pp. 881–890.

[20] D. E. Maurer and J. P. Baker, "Fusing multimodal biometrics with quality estimates via a bayesian belief network," *Pattern Recognition*, vol. 41, no. 3, pp. 821–832, 2007.

[21] F Alonso-Fernandez, J. Fierrez, D. Ramos, and J. Ortega-Garcia, "Dealing with sensor interoperability in multi-biometrics: The upm experience at the biosecure multimodal evaluation 2007," in *Proc. of SPIE Defense and Security Symposium, Workshop on Biometric Technology for Human Identification*, 2008.

[22] Sadaoki Furui, "Recent advances in speaker recognition," *Pattern Recognition Letters*, vol. 18, no. 9, pp. 859 – 872, 1997, Audio- and Video-Based Person Authentication.

[23] D.E. Sturim and D.A. Reynolds, "Speaker adaptive cohort selection for tnorm in text-independent speaker verification," *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, vol. 1, pp. 741–744, 18-23, 2005.

[24] Johnny Marithoz and Samy Bengio, "A Bayesian Framework for Score Normalization Techniques Applied to Text Independent Speaker Verification," *IEEE Signal Processing Letters*, vol. 12, no. 7, pp. 532–535, 2005.

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

[26] Javier Ortega-Garcia, Julian Fierrez, Fernando Alonso-Fernandez, Javier Galbally, Manuel R. Freire, Joaquin Gonzalez-Rodriguez, Carmen Garcia-Mateo, Jose-Luis Alba-Castro, Elisardo Gonzalez-Agulla, Enrique Otero-Muras, Sonia Garcia-Salicetti, Lorene Allano, Bao Ly-Van, Bernadette Dorizzi, Josef Kittler, Thirimachos Bourlai, Norman Poh, Farzin Deravi, Richard Ng, Michael Fairhust, Jean Hennebert, Andreas Humm, Massimo Tistarelli, Linda Brodo, Jonas Richiardi, Andrzej Drygajlo, Harald Ganster, Federico Sukno, Sri-Kaushik Pavani, Alejandro Frangi, Lale Akarun, and Arman Savran, "The multi-scenario multi-environment biosecure multimodal database (bmdb)," *IEEE Trans. on Pattern Analysis and Machine*, 2009.