QUALITY MEASURES IN UNIMODAL AND MULTIMODAL BIOMETRIC VERIFICATION

Jonas Richiardi, Krzysztof Kryszczuk, Andrzej Drygajlo

Signal Processing Institute, Swiss Federal Institute of Technology Lausanne (EPFL)
1015 Lausanne, Switzerland
phone: + (41) 21 693 46 91, email: {jonas.richiardi,krzysztof.kryszczuk,andrzej.drygajlo}@epfl.ch
web: http://scgwww.epfl.ch

ABSTRACT

Real-world deployment of unimodal and multimodal biometric systems often have to contend with degraded signal quality and erratic behaviour of the biometric data being modelled. We review approaches that have been used to extract additional information about the biometric data that can then be used to improve performance in degraded conditions, with a special emphasis on speech (where we present new approaches for signal quality estimation in biometric verification), face, fingerprint, and signature modalities. We also present approaches that do not depend on specific modalities, including new user-model based quality measure. We show how this information can be used in a unimodal and multimodal context, and we perform objective evaluation of quality measures on multimodal benchmarking databases.

1. INTRODUCTION

The scale of deployment of biometric identity verification systems has recently increased dramatically, with the progressive introduction of biometric passports and the burgeoning of the biometric industry. Biometric technology has been moving out of the laboratories into the real world, where a new set of constraints raises difficult technical issues. One of the main problems facing biometric recognition systems in large-scale deployments is error rates, since even low error rates will incommode an objectively large fraction of the population. One key route of addressing errors is the use of quality measures, which we define as *information that helps assess the probability that a biometric verification decision is correct*.

The importance of quality measures in biometric verification is now being increasingly recognized, with specialized workshops organized (e.g. NIST Biometric Quality Workshop) and standardization under way (for instance ANSI/INCITS 379 and ISO/IEC 19794-6 for the iris modality).

To be useful in automated biometric authentication systems, quality measures should be statistically correlated with the classifier output scores and classifier decisions [19, 26]. They constitute additional information about the classification process that can be modeled appropriately. From a machine learning perspective, these quality measures are features, that can be for instance be fed to a second-level classifier or be concatenated to the base feature vector.

We provide a classification of the different types of errors in biometric verification (Section 2) and propose a taxonomy of quality measures (Section 3). In Section 4 we review existing modality-independent quality measures and propose two new user model-based quality measures that can be used with statistical models independently of the modality. Section 5 reviews modality-dependent quality measures and proposes three quality measures that can be used in speaker verification. Section 7 provides a systematic overview of the issues associated with the use of quality measures in biometric verification. We perform experiments on the presented quality measures in Section 8, and conclude in Section 9.

2. WHY DO BIOMETRIC VERIFICATION CLASSIFIERS MAKE MISTAKES?

We distinguish three types of classification errors in biometric identity verification: *systematic*, *presentation-dependent*, and *user-dependent*.

Systematic errors are those caused by design problems inherent to the pattern recognition system engineering task. These include wrong assumptions about the form or family of the distributions of features under consideration, poor choice of features leading to excessive overlap between classes, insufficient amount of training data, poor estimation of model parameters (for example insufficient number of iterations, or aggressive variance flooring), or inadequate decision threshold setting.

Presentation-dependent errors are those caused by unforeseen variability in the signal source. These can be caused by degraded environmental conditions (e.g. ligthing variation for face, specular reflection for iris, additive noise or channel noise for speech, residual fingerprints traces), or by extra variability in a signal (e.g. elastic skin distortion for fingerprints, expression of the face, badly executed signature)

User-dependent errors happen only with certain users that do not fit the otherwise correct assumptions about the user population. This is a well-known problem in biometrics, and one of its incarnations in speaker recognition tasks is called the "Doddington Zoo effect" [8].

The goal of developing quality measures is to find quantities that are indicative of these three types of errors.

3. A SHORT TAXONOMY OF QUALITY MEASURES

Quality measures can be *modality-dependent* and *modality-independent*. *Modality-dependent* measures (such as "frontalness" in face recognition) are not applicable to other modalities, as they exploit specific domain knowledge that can not be transferred to other signals. *Modality-independent* quality measures (such as distance to decision threshold) are more generic and can be exploited across different modalities.

Quality measures can be *absolute* or *relative*. *Relative* quality measures need reference biometric data, and output a comparison to this reference data taken as a "gold standard" of quality. For instance, correlation with average face is a relative measure of quality. *Absolute* measures do not need reference data, except for initial development of the algorithm. A hybrid approach can also be used, whereby an absolute quality measure is extracted and further normalized by some function of the quality of enrollment data [10]).

Lastly, quality measures can be extracted *automatically*, or *hand-labeled* (as in [10]). In this paper we consider only automatically extracted quality measures.

4. MODALITY-INDEPENDENT QUALITY MEASURES

Keeping in mind that the goal of quality measures is to help predict verification errors, we can use some information that does not directly depend on the underlying signal properties. Here we review three approaches that are generic enough to be used with many

modalities and classifiers, though each approach may need to be adapted to fit different classifier families.

4.1 Score-based measures

Many classifiers provide a continuous-valued output (measurement-level) indicating how close a particular sample is to a particular class, a quantity called *score* in biometrics. The probability of classification error increases as the distance gets closer to the decision boundary between classes. This "soft" classifier output, and its distribution (which can be modeled in several ways, see Section 7.1), constitute valuable data for error prediction, and are applicable to any biometric modality whose classifier produces a non-discrete yield output. The use of the score as a quality measure forms the basis of many confidence models [11, 23, 3, 25].

Quantities derived from the score are also used, for instance variance of the score (provided by human expert knowledge of the problem domain) and distance from normalized score to "hard" (decision-level) classifier output (assuming the classifier decisions are the integer extremal points in the score interval, which is typically $[0,1]$) [4]. Indeed, the distance from the score to the decision threshold constitutes a quality measure: it is more probable that the classifier will make a mistake if a score is close to the decision boundary, as noise alone could have moved that score over the threshold. This is the idea behind the method of margins [25].

The distance from user-specific to user-independent decision threshold can be used as a quality measure. In a verification system with a user-independent threshold[1], some users will be more systematically subjected to false rejects, respectively false accepts, than others. Combining this quality measure with the score quality measure simplifies the subsequent classification or regression task [26].

4.2 User model-based quality measures

Information about the user models can be used to detect systematic errors.

First, the closer (in feature space) the user models are to the impostor models, the more likely it is that the classifier will make an error. Thus, an estimate of the "amount of overlap" between the user models and the impostor models in feature space can be used as a quality measure. A method of estimating the amount of overlap for Vector Quantization and Gaussian Mixture Models (GMM) is used for speaker recognition in [14]. In [16] a sum of log-likelihoods for client model and the world model is used as a quality measure of face images. This measure encodes the divergence of the test image quality from the reference quality of the images from the training gallery.

Second, parameter estimation errors can be taken into account. In the case of statistical models such as GMMs, the distance (likelihood) computation rests upon the Mahalanobis distance between the user's model (mean vectors, covariance matrices, and mixing coefficients) and the biometric pattern. The Mahalanobis distance is expressed as follows:

$$d_{Mahal} = (\mathbf{o}-\mu)'\mathbf{\Sigma}^{-1}(\mathbf{o}-\mu) \tag{1}$$

As can be seen from Eq. (1), this distance requires an inversion of the covariance matrix $\mathbf{\Sigma}$. Because this covariance matrix is typically estimated from a limited amount of data using a maximum likelihood procedure, it may be ill-conditioned, meaning that the quality of inversion will be low, which in turn entails errors in the Mahalanobis distance computation. We can therefore use the logarithm of the determinant of the covariance matrix of a Gaussian mixture component as a quality measure QM_{ldet}. If the determinant for a covariance matrix is close to zero, the matrix may be badly conditioned. Another quality measure to use is the logarithm of the *condition number*, which is the ratio of the largest singular value in a matrix to the smallest singular value. A large condition number indicates an ill-conditioned matrix. We denote this quality measure as QM_{lcond}. For both quality measures, we take a weighted sum of the quality measures over all Gaussian mixture components, where the weights are provided by the mixing coefficients of each mixture component.

[1] For instance because it has recently been deployed and there is not enough data for each user to reliably set a personalised threshold.

5. MODALITY-DEPENDENT (SIGNAL-DOMAIN) QUALITY MEASURES

5.1 Speech quality measures

5.1.1 Measures based on voice activity detection

Voice activity detection (VAD), also called speech/pause segmentation, can be used to obtain an estimate of the signal-to-noise ratio. This is done by assuming the average energy in pauses represents the noise energy, and the energy in speech represents the signal energy. The formulation for this family of speech-based quality measures QM_{VAD} is:

$$QM_{VAD} = 10\log_{10}\frac{\sum_{i=1}^{N} Is(i)s^2(i)}{\sum_{i=1}^{N} In(i)s^2(i)}, \tag{2}$$

where $\{s(i)\}, i = 1,\ldots,N$ is the acquired speech signal containing N samples, $Is(i)$ and $In(i)$ are the indicator functions of the current sample $s(i)$ being speech or noise during pauses (e.g. $Is(i)$=1 if $s(i)$ is a speech sample, $Is(i)$=0 otherwise) as reported by the voice activity detector.

In [29] an energy-based VAD and a spectral entropy-based VAD are used, but any robust VAD algorithm can be used for that purpose (e.g. [22]).

5.1.2 Measures using higher-order statistics

Since the amplitude of clean speech has a very distinctive distribution (sharp peak at sample value 0 - a large amount of speech is actually silence if no VAD preprocessing is applied), we can exploit this knowledge to infer when the signal is noisy. The energy of the additive noise we are concerned about contributes to modifying the time-domain distribution of amplitudes.

Higher order statistics can be used to summarize the shape of unimodal distributions in a meaningful way. The skewness (or Fisher skewness) measures the asymmetry of a distribution with respect to its mode. Any symmetrical distribution (such as Laplace, Gaussian, or uniform) has a skewness of 0. Negative skewness indicates that the distribution has a longer tail on the left of the mode, while positive skewness indicates the opposite.

$$QM_{skew} = \frac{E[s-\mu_s]^3}{E[s^2]^{3/2}} = \frac{E[s-\mu_s]^3}{\sigma_s^3} \tag{3}$$

Kurtosis (or Fisher kurtosis), defined in Eq. (4), corresponds to the "peakiness" of the distribution. By definition, a Gaussian distribution has a kurtosis of 3[2]. A leptokurtic (or supergaussian) distribution has a kurtosis higher than 3 and is "peakier", while a platykurtic (or subgaussian) distribution has a kurtosis lower than 3 and is "flatter", that is its probability density is spread over a larger dynamic input range.

$$QM_{kurt} = \frac{E[s-\mu_s]^4}{E[s^2]^2} = \frac{E[s-\mu_s]^4}{\sigma_s^4} \tag{4}$$

Unfortunately, kurtosis estimation is very sensitive to outliers. We therefore introduce a third related measure, called the center bin measure, to approximate kurtosis and estimate the peakiness of the distribution. First, the signal sample amplitudes are binned in 100 equally-spaced bins, then the measure is defined as the ratio of the number of samples in the bin containing the most samples to the total number of samples in the other bins.

$$QM_{bin} = \frac{N_{max}(s)}{\left(\sum_B N_b(s)\right) - N_{max}(s)}, \tag{5}$$

[2] Or 0, as some definitions of kurtosis subtract 3 to have kurtosis of 0 for the normal distribution

where $N_b(s)$ represents the number of samples in bin b, and $N_{max}(s)$ represents the number of samples in the bin that contains the most samples.

5.1.3 *Measures based on an explicit noise model*

A statistical model of noise can be built during enrollment and then compared to the deployment conditions [31], thus forming a relative quality measure.

5.2 Face quality measures

In comparison with other modalities, relatively few works on automatic face quality measures are present. In [17] adversely illuminated face images are segmented using statistical methods, and the face image area left after segmentation is used as a quality measure that helps find an optimal decision threshold. A more systematic approach towards the use of quality measures for face verification has been reported in [18], where two face quality measures are used as evidence in the process of reliability estimation. Those quality measures are image contrast (QM_{f2}), and normalized 2D correlation with an average face template (QM_{f1}).

5.3 Fingerprint quality measures

The fingerprint modality is the biometric modality for which most signal quality estimation algorithms have been developed. A recent review of the state-of-the art fingerprint quality measures is given in [1]. The authors divide the automatic fingerprint quality measures into local, global, and 'based on classifiers' groups. Actually, it must be noted that the quality measures baptized as 'based on classifiers' are measuring the separation between the match and non-match fingerprint feature distributions and as such are not strictly modality-specific, falling into the category described in Section 4. This method has been used in the publicly available quality measure estimation module NFIQ of the NIST/NFIS2 fingerprint verification package [32].

5.4 Signature quality measures

For signature, no signal degradation is present and the modality-independent quality measures described in Section 4 can be used.

6. EVALUATING QUALITY MEASURES

Since one aim of using quality measures is to predict verification errors, one important way of looking at quality measures is to plot their distributions with respect to two classes: the class of correct classification decisions, and the class of incorrect classifications, which we denote $DR = 1$ (Decision Reliable) and $DR = 0$ respectively.

In [15] Koval et al. have proven that dependent features allow for better class separation than independent features. Therefore, quality measures that are statistically dependent on the features or scores are expected to allow for better class separation than features or scores alone [19]. The intuitive graphical interpretation of this fact can be seen in Figure 1. Consequently, quality measures can be evaluated by measuring their statistical dependence on the scores. Under the assumption of linearity this dependence can be estimated by computing the correlation coefficient between the quality measures and scores. Additionally, the linear correlation coefficient between the *DR* variable and the value of the quality measure gives an indication of the ability of the quality measure to predict errors.

It is also possible to use the mean squared Mahalanobis distance between the distributions of each quality measure for the correct classifier decision and erroneous classifier decision cases. Higher Mahalanobis distance between the distributions for correct and erroneous decisions distributions indicates the quality measure is a good predictor of classifier errors, but sports an implicit Gaussian assumption about the distributions.

Another objective measure of goodness for quality measures is normalized cross-entropy [24] (normalized mutual information). It

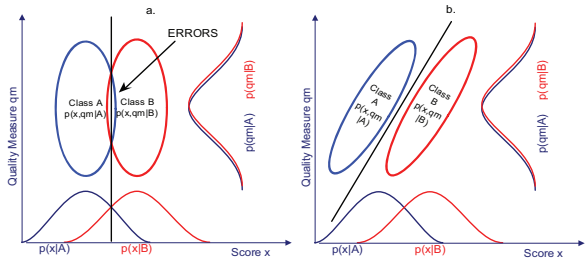

Figure 1: Relationship between scores and quality measures and the impact of their statistical dependence on class separation. Ellipses symbolize two dimensional class-conditional distributions in a space defined by quality measures and scores: a. for independent quality measures and scores and b. for linearly correlated quality measures and scores.

can be phrased as the "relative decrease in uncertainty about the classifier's decision provided by the quality measure".

An important point is that the ultimate evaluation for a quality measure is to apply it to a biometric verification task dataset and see if it leads to improvements in terms of final error rate or rejection rate. While a quality measure may seem to poorly separate the error-conditional distributions, as pointed out by a low Mahalanobis distance, there may still exist a classifier which can make good use of the quality data.

7. USING QUALITY MEASURES IN BIOMETRIC VERIFICATION

7.1 Modeling quality measures

Quality measures can be modeled using generative or discriminative training paradigms, with parametric or non-parametric models. The aim in this case is to build a second-level classifier that can provide additional information on the reliability of the biometric verification result, or to improve classification accuracy directly. We give here a short overview of model families that have been used.

A second-order regression model is used for speech in [14]. A single Gaussian distribution has been used in [11, 23] for speech and [3] for speech and face. A Bayesian network with Gaussian distributions has been used in [28, 20] respectively for speech, and speech and face and in [29] with mixtures of Gaussian distributions for speech data. A multi-layer perceptron is used in [6] on a speaker recognition task. A kernel-based modeling approach is taken for the margins confidence estimation method [25]. Non-parametric modeling of scores has been used in [3], where a histogram-based method for speech and face is presented.

Ensemble classifiers are also used to model quality measures, for instance random forests have been used in [26] to improve classification accuracies of speaker verification and signature classifiers, and to perform multiple classifier fusion on signature data.

7.2 Single classifier systems with quality measures

In the unimodal context, the output of a model including quality information can be used for either automatic processing (such a matching algorithm choice [12]) or human consideration (such as forensic expertise [6]). Quality measures have been shown to provide evidence for computing the reliability of classification decisions which can in turn be used to discard unreliable decisions [18] and request a repeated acquisition [28]. Quality measures can also play an integral role in the classification process and can be used directly as a classification feature for a stacked classifier ensemble, like in the *Q-stack* approach [19]. The latter approach is taken for the evaluation of quality measures presented in Table 1.

7.3 Multiple classifier systems with quality measures

In recent years, the essential contribution of quality information to the fusion of multiple classifiers has been increasingly acknowledged. For multiple classifier systems the use of quality measures can be divided into *heuristic* and *statistical* methods.

The *heuristic* methods embody an intuitive notion that if two classifiers arrive at a decision at unequal confidence levels, the more confident classifier should be trusted more. For multimodal biometrics this rule translates into trusting a modality for which a higher-quality signal is available. Examples of heuristic fusion schemes with quality measures include quality-based decision and score weighting approaches [20] The performance of the heuristic methods depends on how accurate the heuristics is in each particular case, but they can be applied to unseen data.

The *statistical* methods learn the impact of the quality measures on classification errors from available training data with associated quality labels. Examples of classifier fusion schemes in biometrics include [10, 25], as well as [30] for classifier selection in face recognition. For multiple classifier systems, the quality information can also be employed in the *Q-stack* scenario [19], and in the *rigged majority voting* approach [26]. The applicability of the statistical methods hinges on the availability of the relevant data. Namely, sufficient training data of quality compatible with that encountered during testing must be available in order to accurately model the dependencies between quality measures and scores. In general, given sufficient and relevant training data the statistical methods outperform the heuristic methods [19].

Fusion of speech and fingerprint using (hand-labeled) signal quality measures is shown in [4], resulting in classification improvement if the fingerprint signal quality is taken into account. A speech quality measure based on an explicit noise model is used to weight the contribution of a speech expert to a speech and face multimodal system, achieving good results in degraded acoustic conditions [31]. Fusion of fingerprint and speech making use of fingerprint quality measures with polynomial regression models achieved about 2% reduction in error rates compared to the baseline fusion method without quality measure [33].

7.4 Using several quality measures

To obtain better modeling of error conditions, quality measures can be combined. For example, the score quality measure by itself may not lead to very high accuracy in recognizing errors, but combining it with the distance to a decision threshold yields much better results [26]. Likewise, adding an entropy-based quality measure for speech helps compensate the deficiencies of energy-based quality measures in high noise situations [29]. In face verification, combining several signal quality measures also improves the estimation of reliability [18].

Lastly, some quality measures are themselves an arithmetic aggregate of other quality measures, that each take into account a different aspect of the signal [13, 21].

8. EXPERIMENTS

We compute quality measures for all modalities on various reference databases and report on their intrinsic performance. We then train second-level stacking classifiers (Gaussian mixture models, instance-based classifiers, or decision-tree based classifiers) on two-dimensional feature vectors comprising of the score and the quality measures, and report on the decrease in error rate compared to the baseline. When no fusion protocol is specified with the database, we perform 10-fold cross validation.

8.1 Databases and systems

The database used for speech experiments is BANCA [2]. The classifier used for speaker verification is a GMM based on the ALIZE toolkit [5], trained following BANCA protocol P.

Face results are given for the BioSec database [9], a PCA/LDA-based classifier, and quality measures described in detail in [18]. QM_{f1} is the correlation coefficient with an average face template, and QM_{f2} is an image contrast measure.

Fingerprint results are computed for the BioSec database [9], optical sensor. Scores computed using the NFIS2 system [32], quality measures: QM_{fp1} and QM_{fp2} as described in [7], QM_{NFIQ} computed by the `NFIQ` quality measure routine, native to the NFIS2 package.

Signature results are given using a 2-components Gaussian mixture model classifier with diagonal covariance matrices. 12 global features are used [27] (for space reasons, results for a classifier based on local features are not shown here). The database used for signature experiments is the MCYT-100 database.

8.2 Results

Quality measure	ρ_D	ρ_{Sc}	d_{Mahal}	$\Delta_{HTER}[\%]$
Speech (baseline HTER: 8.4 %)				
QM_{VAD_E}	0.141	0.149	0.98	27.7
QM_{kurt}	0.081	0.151	6.80	11.0
QM_{skew}	-0.074	0.026	2.25	34.2
QM_{bin}	0.08	0.132	1.22	26.0
QM_{VAD_H}	0.1273	0.125	0.83	34.2
Face (baseline HTER: 25.6 %)				
QM_{f1}	0.303	0.363	899.58	18.7
QM_{f2}	-0.203	-0.110	325.42	3.1
Fingerprint (baseline HTER: 0.6 %)				
QM_{fp1}	0.017	0.132	7.07	22.2
QM_{fp2}	0.117	0.188	14.73	21.2
QM_{NFIQ}	-0.031	-0.082	7.912	23.6
Signature (baseline HTER: 19.0 %)				
QM_{ldet}	0.053	-0.033	1.14	21.0
QM_{lcond}	-0.041	0.076	1.082	27.4

Table 1: Linear correlation coefficient between the decision correctness indicator (DR) and the quality measure (ρ_D) and between the quality measure and the score (ρ_{Sc}), mean squared Mahalanobis distance (d_{Mahal}) between the DR-conditional distributions of quality measures, and relative reduction in HTER Δ_{HTER}, in percentage. Here, the error rates of baseline systems are compared to results obtained using the $Q-stack$ method. The modalities are speech (BANCA G2 data), face (BioSec data), fingerprints (BioSec data), and signature (MCYT100 data)

9. CONCLUSION

We have presented a systematic classification of the types of quality measures currently used in biometric identity verification, and evaluated modality-independent quality measures. The results highlight the importance of performing experimental evaluation of quality measures in the context of their use: while indicators of performance such as those presented in Section 6 provide an useful overview into the inherent usefulness of the quality measures, the decision boundaries of a feature space including quality measures are often too complex to be accounted for by simple measures such as correlation coefficients. We have introduced new modality-dependent quality measures that can be used in speaker verification, as well as new modality-independent quality measures accounting for some deficiencies of the parameter estimation process in statistical models. These constitute valuable information for second-level classifiers to improve upon the baseline results, as was demonstrated by an application to signature models. Using both modality-dependent and modality-independent quality measures is likely to improve classification accuracy even further, as the information contained in these two types of measures is weakly interdependent.

10. ACKNOWLEDGEMENTS

We thank J. Ortega-Garcia and J. Fierrez-Aguilar for the provision of the MCYT-100 and BioSec corpora.

REFERENCES

[1] F. Alonso-Fernandez, J. Fierrez-Aguilar, and J. Ortega-Garcia. A review of schemes for fingerprint image quality computation. In *Proceedngs, 3rd COST-275 Workshop on Biometrics on the Internet*, pages 3–6, Hatfield, UK, 2005.

[2] 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 J. Kittler and M.S. Nixon, editors, *Proceedings of 4th Int. Conf. on Audio- and Video-Based Biometric Person Authentication (AVBPA)*, volume LNCS 2688, pages 625–638, 2003.

[3] S. Bengio, C. Marcel, S. Marcel, and J. Mariethoz. Confidence measures for multimodal identity verification. *Information Fusion*, 3(4):267–276, December 2002.

[4] J. Bigun, J. Fierrez-Aguilar, J. Ortega-Garcia, and J. Gonzalez-Rodriguez. Multimodal biometric authentication using quality signals in mobile communications. In *Proc. 12th Int. Conf. on Image Analysis and Processing*, pages 2–11, 2003.

[5] J.-F. Bonastre, F. Wils, and S. Meignier. ALIZE, a free toolkit for speaker recognition. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005)*, pages 737–740, Philadelphia, USA, March 2005.

[6] W.M. Campbell, D.A. Reynolds, J.P. Campbell, and K.J. Brady. Estimating and evaluating confidence for forensic speaker recognition. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 717–720, 2005.

[7] Y. Chen, Sarat C. Dass, and A.K. Jain. Fingerprint quality indices for predicting authentication performance. In *Proceedings of AVBPA*, Rye Brook, NY, 2005.

[8] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D.A. Reynolds. SHEEP, GOATS, LAMBSand WOLVES: a statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation. In *Proc. 5th Int. Conf. on Spoken Language Processing (ICSLP)*, Sydney, Australia, November-December 1998.

[9] J. Fierrez, J. Ortega-Garcia, D. Torre-Toledano, and J. Gonzalez-Rodriguez. Biosec baseline corpus: A multimodal biometric database. *Pattern Recognition*, 40(4):1389–1392, April 2007.

[10] J. Fierrez-Aguilar, J. Ortega-Garcia, J. Gonzalez-Rodriguez, and J. Bigun. Discriminative multimodal biometric authentication based on quality measures. *Pattern Recognition*, 38(5):777–779, May 2005.

[11] H. Gish and M. Schmidt. Text-independent speaker identification. *IEEE Signal Processing Magazine*, 11(4):18–32, October 1994.

[12] Patrick Grother and Elham Tabassi. Performance of biometric quality measures. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29(4):531–543, 2007.

[13] L. Hong, Y. Wan, and A. Jain. Fingerprint image enhancement: algorithm and performance evaluation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(8):777–789, 1998.

[14] M.C. Huggins and J.J. Grieco. Confidence metrics for speaker identification. In *Proc. 7th Int'l Conf. on Spoken Language Processing (ICSLP)*, 2002.

[15] O. Koval, S. Voloshynovskiy, and T. Pun. Error exponent analysis of person identification based on fusion of dependent/independent modalities. In *In Proceedings of SPIE Photonics West, Electronic Imaging 2006, Multimedia Content Analysis, Management, and Retrieval 2006 (EI122)*, 2006.

[16] K. Kryszczuk and A Drygajlo. Addressing the vulnerabilities of likelihood-ratio-based face verification. In *Proc. 5th AVBPA*, Rye Brook NY, USA, 2005.

[17] K. Kryszczuk and A. Drygajlo. Gradient-based image segmentation for face recognition robust to directional illumination. In *Visual communications and image processing 2005 : 12-15 July 2005, Beijing, Chine*, 2005.

[18] K. Kryszczuk and A. Drygajlo. On combining evidence for reliability estimation in face verification. In *Proc. of the EUSIPCO 2006*, Florence, September 2006.

[19] K. Kryszczuk and A. Drygajlo. Q-stack: uni- and multimodal classifier stacking with quality measures. In *Proc. 7th International Workshop on Multiple Classifier Systems*, Prague, Czech republic, 2007.

[20] 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 (EUSIPCO)*, Antalya, Turkey, September 2005.

[21] E. Lim, X. Jiang, and W. Yau. Fingerprint quality and validity analysis. In *Proc. Int. Conf. on Image Processing (ICIP)*, volume 1, pages 469–472, 2002.

[22] M. Marzinzik and B. Kollmeier. Speech pause detection for noise spectrum estimation by tracking power envelope dynamics. *Speech and Audio Processing, IEEE Transactions on*, 10(2):109–118, 2002.

[23] H. Nakasone and S.D. Beck. Forensic automatic speaker recognition. In *Proc. 2001: A Speaker Odyssey*, 2001.

[24] National Institute of Standards and Technology. The 2001 NIST evaluation plan for recognition of conversational speech over the telephone, Oct. 2000.

[25] N. Poh and S. Bengio. Improving fusion with margin-derived confidence in biometric authentication tasks. In *Fifth Int. Conf. Audio- and Video-Based Biometric Person Authentication (AVBPA)*, 2005.

[26] J. Richiardi and A. Drygajlo. Reliability-based voting schemes using modality-independent features in multi-classifier biometric authentication. In *Proc. 7th International Workshop on Multiple Classifier Systems*, Prague, Czech republic, 2007.

[27] J. Richiardi, H. Ketabdar, and A. Drygajlo. Local and global feature selection for on-line signature verification. In *Proc. IAPR 8th International Conference on Document Analysis and Recognition (ICDAR 2005)*, volume 2, pages 625–629, Seoul, Korea, August-September 2005.

[28] J. Richiardi, P. Prodanov, and A. Drygajlo. A probabilistic measure of modality reliability in speaker verification. In *Proc. IEEE International Conf. on Acoustics, Speech and Signal Processing 2005*, pages 709–712, Philadelphia, USA, March 2005.

[29] J. Richiardi, P. Prodanov, and A. Drygajlo. Speaker verification with confidence and reliability measures. In *Proc. 2006 IEEE International Conference on Speech, Acoustics and Signal Processing*, Toulouse, France, May 2006.

[30] M.T. Sadeghi and J. Kittler. Confidence based gating of multiple face authentication experts. In *Proc. Joint IAPR Int. Workshops, Structural, Syntactic, and Statistical Pattern Recognition 2006*, volume 4109/2006, pages 667–676, August 2006.

[31] C. Sanderson and K.K. Paliwal. Noise compensation in a person verification system using face and multiple speech features. *Pattern Recognition*, 36(2):293–302, February 2003.

[32] E. Tabassi, C.L. Wilson, and C. Watson. Nist fingerprint image quality. Technical Report NISTIR 7151, NIST, August 2004.

[33] K.-A. Toh, W.-Y. Yau, Eyung L., L. Chen, and C.-H. Ng. *Fusion of Auxiliary Information for Multi-modal Biometrics Authentication*, volume 3072 of *LNCS*. Springer, 2004.