Quality Controlled Multimodal Fusion of Biometric Experts

Omolara Fatukasi, Josef Kittler, and Norman Poh

Centre for Vision, Speech and Signal Processing,
University of Surrey Guildford, GU2 7XH Surrey, UK
{O.Fatukasi,J.Kittler,N.Poh}@surrey.ac.uk

Abstract. The quality of biometric samples used by multimodal biometric experts to produce matching scores has a significant impact on their fusion. We address the problem of quality controlled fusion of multiple biometric experts and focus on the fusion problem in a scenario where biometric trait quality expressed in terms of quality measures can be coarsely quantised. We develop a fusion methodology based on fixed rules that exploit the respective advantages of the sum and product rules and can be easily trained. We show in experimental studies on the XM2VTS database that the proposed method is very promising.

Keywords: Biometric authentication, fixed rules, multiple classifiers system, multimodal fusion, quality dependent fusion.

1 Introduction

Biometric authentication is the verification of a user's identity by means of his/her physical and behavioural characteristics. Studies, e.g. [1] have shown that the fusion of experts improves the system performance when compared with individual experts. However poor quality biometric data may have the opposite effect [2,3]. This finding motivated the investigation of quality based fusion. It has been shown in [4,5,6,7,8,9,10] that quality based fusion improves significantly the performance, as compared to conventional fusion methods (fusion without the use of quality information).

The recent research into quality based score fusion shows that it is beneficial to include quality information as input to the fusion process. In confidence based decision fusion, quality information is also used as a control parameter to select which modality's decision to follow. Most of the quality based multimodal fusion techniques deploy training for the fusion stage design [4,5,6,10]. The exception is [7], where the product rule is used, after adapting the scores by computing the likelihood ratio of estimated densities.

In this paper we address the problem of quality controlled fusion of multiple biometric experts. We focus on the fusion problem in a scenario where biometric trait quality expressed in terms of quality measures can be coarsely quantised. We develop a fusion methodology based on fixed rules that can be easily trained. The methodology involves a two stage process whereby in the first stage expert

L. Rueda, D. Mery, and J. Kittler (Eds.): CIARP 2007, LNCS 4756, pp. 881–890, 2007.

scores are grouped according to the quality of the underlying biometric sample. In each quality group the scores are combined by averaging. The resulting group scores are finally combined by product. We argue that the proposed scheme exploits the properties of fixed fusion rules in the best possible way and provide experimental evidence to support this argument. The proposed scheme is experimentally evaluated on the XM2VTS database. The results show that significant performance gains can be achieved. The performance is comparable to the state of the art method reported in [10] but the proposed fusion system is much easier to design and requires less data for training.

The rest of the paper is organised as follows. In Section 2 we introduce the proposed methodology. The database used in the study is described in Section 3. An overview of the biometric experts used for experiments is presented in Section 4. Section 5 discusses the quality measures used to characterise biometric sample quality. We also report in this section the coefficients of correlation between expert scores in different quality categories. The fusion experiments carried out are described in Section 6 where the results of experiments are also discussed. Section 7 draws the paper to conclusion.

2 Proposed Methodology

The study of fixed fusion rules in [1] demonstrates that the sum rule outperforms all other fixed rules. Alkoot *et al.* showed in [11] that the product rule may outperform even the sum rule, provided the veto effect of conflicting low valued scores is suppressed. The product rule and the sum rule have been compared by Kittler *et al.* in [1] and Tax *et al.* in [12]. These studies demonstrate that the sum rule is robust to noise. The sensitivity of the product rule to noise is due to the veto effect. Tax *et al.* also show that the product rule outperforms the sum rule when the correlation between data is low and noise is low. However if the noise is high, the product rule becomes unreliable even when correlation is low. These studies lead to the following conclusion:

- if a high level of noise is present, the sum rule is preferable.
- for low noise and low correlation, the product rule should be favoured as it outperforms the sum rule in these conditions.
- when experts are highly correlated, even when the noise level is low the sum rule should be chosen, as it outperforms the product rule under these conditions.

[1] shows theoretically that the product rule is more sensitive to noise than the sum rule, hence why it deteriorates on noisy data.

In this paper we consider the problem of fusing multiple experts providing scores on biometric data of varying quality. The scores are assumed to be normalised, so that any fixed rule, including the product rule can be used for fusion. Thus the score values are confined to the interval $[0, 1]$. Without loss of generality, we assume that a score is high (close to 1) for a good match, i.e. when comparing a probe of a genuine claimant against a template of the true client

identity. Impostor score values, of course, would be lower. Clearly for biometric samples of low quality, both the client scores and impostor scores would shift towards the lower end of the score range. It is evident, that when the quality of the biometrics trait varies, a single threshold would not be adequate, as the scores generated by high quality traits of imposters are likely to exceed the scores of true clients derived from low quality biometric data. This problem can be solved by considering the threshold to be a function of the set of quality measures characterising the biometric data. However learning the regression function requires a large amount of data which is not always available.

A similar problem, but greatly amplified, arises in multiple expert fusion. The additional complexity derives from the fact the threshold for the fused score becomes a function of the quality measures of all biometric modality traits jointly. The reason for this is that the fusion potentially involves expert scores associated with different qualities and this will impact on the optimal threshold to be applied to the fused score. The regression function defining the optimal threshold is much more difficult to learn, as the number of variables involved in regression increases without the commensurate increase in the number of training samples. This problem was investigated in detail in [10] where it was demonstrated that significant gains in performance can be obtained by quality dependent fusion where the fusion was realised as a Support Vector machine using both component expert scores and biometric trait qualities as features.

It would appear, therefore, that the key advantage of fixed fusion rules, namely their simplicity and ease of training, is seriously compromised when the experts to be fused use data of different quality. However, in many situations the biometric data quality will not necessarily be uniformly distributed with respect to the various quality measures. Instead, it is likely to be clustered. For instance, if the biometric data is collected in a small set of distinct environments, or using a small set of devices supplied by different manufacturers or involving sensor technology for a particular biometric trait designed on different principles, the data acquired will tend to cluster into a number of quality states corresponding to the distinct conditions of data acquisition. In such situation it would be feasible to group the experts according to the quality state of the biometric data used for computing their score. In each group, it should then be possible to use a fixed fusion rule and subsequently, combine the group scores to produce the final fused result.

We shall develop the above ideas into a practical fusion methodology applicable under the assumption that the biometric data can sensibly be divided into two quality states. We shall see in Section 5 that this assumption is valid for the biometric database, XM2VTS, used for our experiments. In order to be more specific, we shall introduce the necessary mathematical notation.

Let $i = 1 : n$ samples, $j = 1 : R$ experts, and $m = 1 : M$ modalities. The decision whether to assign the quality of a biometric sample $x_{j,i}$ to high or low quality, is dependent on the quality measure, $q_{i,m}$, of the sample, its mean $\overline{q_m}$ and the standard deviation σ_{q_m} and biometric modality in the evaluation data set. A sample $x_{j,i}$ is marked as high quality if $q_{i,m} \geq \overline{q_m} - \sigma_{q_m}$, else it is of

low quality. Let $r_{z,i}$ be the number of experts working with samples of quality $z\epsilon\{high, low\}$. Based on this decision rule we can identify three situations: i) all-high $r_{low,i} = 0$, ii) all-low $r_{high,i} = 0$, and iii) mixed where both $r_{high,i}$ and $r_{low,i}$ are nonzero.

We shall see in Section 5 that experts tend to be correlated. Thus for every sample, within each group, the preferable fixed fusion rule is the sum rule. The fused score for the ith sample in group with quality z is thus given as

$$S_z(i) = \left\{ \begin{array}{ll} \frac{1}{r_{z,i}} \sum_{p=1}^{r_{z,i}} x_{p,i} & if \quad r_{z,i} \geq 1 \\ 1 & if \quad r_{z,i} = 0 \end{array} \right\} \tag{1}$$

Setting the sum to 1 when a group contains no expert is for a later convenience.

Now, in each group we will end up with two averaged scores $S_{high}(i)$ and $S_{low}(i)$. Especially in the mixed group these two scores can further be combined by a fixed rule. We shall see later that the score averaging process in each group results in fused scores $S_z(i), z\epsilon\{high, low\}$ which are much less noisy, and surprisingly, also less correlated. This suggests that the optimal fixed fusion rule for this second fusion stage should be the product rule. Accordingly, the final fused score $S(i)$ for sample i will be given as

$$S(i) = S_{high}(i) \times S_{low}(i) \tag{2}$$

The resulting score $S(i)$ is then compared against the threshold D_θ where $\theta\epsilon\{high, low, mixed\}$. These thresholds are estimated from the training data but it is a relatively simple task.

3 Database

In the current study, we used the original XM2VTS database[13] and its degraded version [14] in both the training and the test phase of the fusion methods. The original database contains mugshot images with well controlled illumination. The low quality section contains images taken under strong side illumination, which has been shown to degrade significantly face verification performance [14]. This database contains 295 individuals, divided into 200 clients, 25 impostors for the algorithm development (training), and 70 impostors for algorithm evaluation (testing). For each subject, face and speech biometric modalities are acquired in four sessions; the first three are used for training the classifiers and the last one for testing. For the face modality we consider the dark data set with left illumination as the "fifth session" and the one with right illumination as the "sixth" session. There is unfortunately no equivalent of degraded speech data that can be paired with the degraded face images. We created degraded biometric data by first introducing additive white noise with a uniform random distribution between 0 and 20dB signal-to-noise ratio on the clean speech database, hence resulting in a degraded speech database with exactly the same size as the clean database. We then paired the degraded face images with the degraded speech data according to Table 1. For instance, the first row shows that the first shot

Table 1. Matching of degraded face and speech data

Degraded face		Degraded speech	
session	shot	session	shot
5	1	1	2
5	2	2	2
6	1	3	2
6	2	4	2

of degraded face image in the fifth session is matched with the second shot of the degraded speech recorded in session one, and so on.

Experimentation with good and degraded data set is important as it reflects a more realistic scenario than the use of only good data. During the data capture of the development data set the environment can be controlled, however in operation the quality is likely to be more varied. Having a good biometric data for the development set and mixed quality biometric data for the operational phase can lead to bad system performance as degraded data is not taken into account in the development stage. It is therefore essential to have representative examples of degradation also for the development.

Unfortunately, the way the experimental data set has been constructed does not allow us to test systematically the merit of fusion when one modality is of good quality and the other one is degraded. Although this is more realistic, there is no obvious solution to introducing this scenario.

The original experimental protocols known as the Lausanne Protocols, did not envisage that for the XM2VTS database the degraded data sets would be used for algorithm development. However, in order to make degraded data avaliable for training, we used the 25-impostor data set in which good and degraded quality data is available. For clients, we divided these 200 subjects into 20- and 180-client data sets such that the 20-client data set is set aside uniquely for algorithm development and the 180-client for both algorithm development and evaluation. The resulting protocol for mixed quality scenario is summarised in Table 2.

Table 2. The XM2VTS clean and degraded protocol

Sessions	Shots	180 Clients	20 Clients	25 Imposter	70 Imposter
S1	1	Training	Training	Evaluation	Test
	2	Evaluation	Evaluation		
S2	1	Training	Training		
	2	Evaluation	Evaluation		
S3	1	Training	Training		
	2	Evaluation	Evalaution		
S4	1	Test	Test		
	2				
Degraded	L1,R1	Test	Evalaution	Evaluation	Test
	L2,R2	degraded	degarded	degraded	degraded

4 Experts

The classifiers used for the face experts in this paper can be found in [15]. There are two classifiers with three types of pre-processing, hence resulting in a matrix of six classifiers. The two classifiers used are Linear Discriminant Analysis (LDA) with correlation as a measure of similarity [16] and Gaussian Mixture Model (GMM) with maximum a posteriori adaptation, described in [17]. The use of the GMM in face authentication was proposed in [18]. The face pre-processing algorithms used include the photometric normalisation as proposed by Gross and Brajovic [19], histogram equalisation and local binary pattern (LBP) as reported in [15]. The feature extraction and classification algorithms are implemented in the open-source Torch Vision Library[1].

The speech system used is implemented with the ALIZE toolkit [20].

5 Quality Measures

In this paper, we used a set of proprietary quality measures developed by Omniperception Ltd for the face image quality assessment. These measures are: "frontal quality", measuring the deviation from the frontal face; and "illumination quality", quantifying the uniformity of illumination of the face.

Two quality measures are used for the speech system: signal-to-noise ratio (SNR) and "entropy quality". Both measures are used for voice activity detection, i.e., to separate speech from non-speech.

These measures can be found in [21]. Thus each modality has two quality measures; "frontal quality" and "illumination quality" for face, signal-to-noise ratio (SNR) and "entropy quality" for speech. These are averaged for each modality.

Table 3. Coefficient of correlation between the six face and one speech experts computed on the development set for the **client** (in bold) and *imposter* (in italic). f1 to f6 are the six face experts and v1 is the speech expert. (a), (b) and (c) shows the correlation coefficient for claims where the quality measure for the biometric data is mixed, low, or high for all experts repectively.

	f1	f2	f3	f4	f5	f6	v1
(a) Mixed quality dataset							
f1	**1.00**/*1.00*	**0.82**/*0.51*	**0.76**/*0.39*	**0.74**/*0.20*	**0.71**/*0.06*	**0.70**/*0.08*	**0.32**/*0.00*
f2	**0.82**/*0.51*	**1.00**/*1.00*	**0.85**/*0.47*	**0.84**/*0.17*	**0.84**/*0.07*	**0.79**/*0.03*	**0.42**/*-0.02*
f3	**0.76**/*0.39*	**0.85**/*0.47*	**1.00**/*1.00*	**0.78**/*0.16*	**0.73**/*0.08*	**0.80**/*0.11*	**0.20**/*-0.01*
f4	**0.74**/*0.20*	**0.84**/*0.17*	**0.78**/*0.16*	**1.00**/*1.00*	**0.93**/*0.39*	**0.02**/*0.31*	**0.38**/*0.05*
f5	**0.71**/*0.06*	**0.84**/*0.07*	**0.73**/*0.08*	**0.93**/*0.39*	**1.00**/*1.00*	**0.91**/*0.37*	**0.49**/*0.09*
f6	**0.70**/*0.08*	**0.79**/*0.03*	**0.80**/*0.11*	**0.92**/*0.31*	**0.91**/*0.37*	**1.00**/*1.00*	**0.29**/*0.07*
v1	**0.32**/*0.00*	**0.42**/*-0.02*	**0.20**/*-0.01*	**0.38**/*0.05*	**0.49**/*0.09*	**0.29**/*0.07*	**1.00**/*1.00*

[1] Available at "http://torch3vision.idiap.ch". See also a tutorial at "http://www.idiap.ch/marcel/labs/faceverif.php".

It is interesting to note the correlation between the experts using low and high quality data. Table 3 shows the correlation coefficient between all the experts for clients (in bold) and imposters (in italic) in the development data set. It can be noted that all the face experts are correlated, but the speech expert is not correlated to any of the face experts.

Most importantly, for the mixed quality scenario the resulting two fused scores have low correlation. In fact the correlation coefficient of the combined scores obtained by averaging in each group is **0.3684**/*-0.2888* client/imposter for the development set and **0.2946**/*-0.3235* client/imposter for the evaluation set. This confirms that these group scores are better suited for fusion by the product rule, as proposed in Section 2.

6 Experiments and Results

We have designed experiments to compare the following:

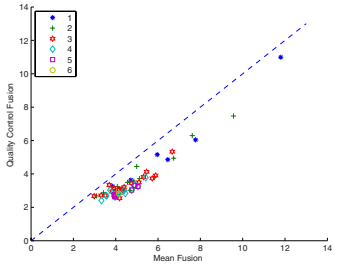
- fixed rule fusion with trained fusion.
- fixed rule fusion with quality and conventional fixed rule fusion
- using quality as a feature in the fusion process and using quality controlled fusion.

The performance of the six face and one speech experts is shown in Table 4. The overall performance is not high due to the influence of low quality biometric data. We consider the set of all $2^6 - 1$ possible combinations of the face experts to be fused with the speech expert for multimodal authentication.

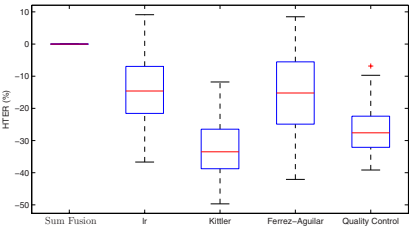
Table 4. Baseline systems, *a priori* half total error rate (HTER) (%) of good + degraded test data, with the *a priori* HTER (%) of the good and degraded data sets recorded separately. The separate good and degraded data results were obtain by using the threshold (Δ) set on the good + degraded training data.

		good + degraded	**good**	**degraded**
modality	no.	HTER (%)	HTER (%)	HTER (%)
face	1	11.06	6.66	13.50
face	2	7.67	3.48	9.78
face	3	8.29	5.86	9.57
face	4	10.39	2.13	17.17
face	5	24.56	2.97	39.28
face	6	16.96	5.51	23.42
speech	1	11.40	1.15	17.48

Figure 1(a) shows the result of the sum rule vs the proposed quality controlled fusion. It can be seen that the quality controlled fusion outperforms the sum rule in all fusion tasks. Another interesting point to note is that the best performance was not obtained when all the face experts were used jointly, but when two, three, or at most four experts are fused together.



(a) Sum Fusion vs Quality Controlled Fusion



(b) Relative *A priori* HTER(%)

Fig. 1. (a) *A priori* HTER (%) of good + degraded test data, with Mean Fusion vs Proposed Quality Controlled Fusion. Each point in the figure represents one of the possible 63 multimodal fusion tasks. The numeric labels in the legend indicate the number of face experts fused with the only speech expert. (b) Relative *a priori* HTER(%) of the sum fusion, logistic regression without quality measure, logistic regression with quality (Kittler *et al.* [10]), SVM with quality measure (Fierrez-Aguilar *et al.* [4]), and the proposed quality controlled fusion.

Figure 1(b) shows the relative *a priori* HTER (%) of conventional sum fusion, a logistic regression with just the expert score as the input to the fusion process, logistic regression with quality measure added as an input feature, proposed in [10], SVM with quality measure used to normalised the score proposed by Fierrez-Aguilar *et al.* in [4], and the proposed quality controlled fusion method. It is interesting to note the following:

1. For logistic regression the average observed relative improvement is 14% with the best improvement realising 39%. This is expected as a trained rule is likely to outperform a fixed rule when the performance of expert varies, as shown in Table 4. However for certain sets of experts, the logistic regression can degrade the performance by as much as 9%.
2. For the method proposed in [10], there is an improvement in all fusion tasks with an average of 33% but as much as 49% can be achieved.
3. For the method proposed by Fierrez-Aguilar *et al.* in [4] an average improvement of 15% with a peak gain of 42% and the worst loss of 8%.
4. For our proposed quality controlled fusion method, there is an improvement in all the fusion tasks with an average improvement of 27%, but up to 39% can be achieved.

These observations highlight the following:

1. Fusion using quality information outperforms conventional fusion.
2. In score level fusion, quality measures can be used in two ways; as input to the fusion process, or as a control parameter.
3. When using quality measures as part of the input to the score level fusion, the method proposed in [10] provides the best average performance and clear improvement in all the fusion experiments. This is evident from Figure 1(b).

4. The score level fusion with the proposed quality control offers very good average performance, and it also provides improvement in performance in all fusion tasks in the experimental comparison with the sum rule. In fact the proposed quality control with a fixed rule performs better than the logistic regression, as shown in Figure 1(b).

7 Discussion and Conclusion

We addressed the problem of quality controlled fusion of multiple biometric experts. We focused on the fusion problem in a scenario where biometric trait quality expressed in terms of quality measures can be coarsely quantised. We developed a fusion methodology based on fixed rules that can be easily trained. The methodology involves a two stage process whereby in the first stage expert scores are grouped according to the quality of the underlying biometric sample. In each quality group the scores are combined by averaging. The resulting group scores are finally combined by product. We argued that the proposed scheme exploits the properties of fixed fusion rules in the best possible way and provided experimental evidence in support of this argument. The proposed scheme was experimentally evaluated on the XM2VTS database. The results showed significant performance gains over conventional fusion. The performance is comparable to the state of the art method reported in [10] but the proposed fusion system is much easier to design and requires less data for training. The proposed method can be used not only for multimodal fusion, but also for intramodal fusion, provided the quality measures of the biometric sample is different for each expert [9].

References

1. Kittler, J., Hatef, M., Duin, R.P.W., Matas, J.: On combining classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence 20, 226–239 (1998)
2. Poh, N., Bengio, S.: A score-level fusion benchmark database for biometric authentication. In: Kanade, T., Jain, A., Ratha, N.K. (eds.) AVBPA 2005. LNCS, vol. 3546, pp. 1059–1070. Springer, Heidelberg (2005)
3. Tabassi, E., Wilson, C., Watson, C.: Fingerprint image quality: Nistir 7151. Technical report, NIST (2004)
4. Fierrez-Aguilar, J., Ortega-Garcia, J., Gonzalez-Rodriguez, J., Bigun, J.: Kernel-Based Multimodal Biometric Verification Using Quality Signals. In: Defense and Security Symposium, Workshop on Biometric Technology for Human Identification, Proc. of SPIE. vol. 5404, pp. 544–554 (2004)
5. Bigun, J., Fierrez-Aguilar, J., Ortega-Garcia, J., Gonzalez-Rodriguez, J.: Multimodal Biometric Authentication using Quality Signals in Mobile Communnications. In: 12th Int'l Conf. on Image Analysis and Processing, Mantova, pp. 2–11 (2003)
6. Kryszczuk, K., Richiardi, J., Prodanov, P., Drygajlo, A.: Error Handling in Multimodal Biometric Systems using Reliability Measures. In: Proc. 12th European Conference on Signal Processing, Antalya, Turkey (September 2005)
7. Nandakumar, K., Chen, Y., Dass, S., Jain, A.: Quality-based score level fusion in multibiometric systems. In: ICPR, Hong Kong, pp. 473–476 (2006)

8. Kryszczuk, K., Drygajlo, A.: On combining evidence for reliability estimation in face verification. In: Proc. 13th European Conference on Signal Processing, Florence, Italy (2006)
9. Fierrez-Aguilar, J., Chen, Y., Ortega-Garcia, J., Jain, A.K.: Incorporating image quality in multi-algorithm fingerprint verification. In: ICB (2006)
10. Kittler, J., Poh, N., Fatukasi, O., Messer, K., Kryszczuk, K., Richiardi, J., Drygajlo, A.: Quality dependent fusion of intramodal and multimodal biometric experts. In: Proceedings of SPIE. vol. 6539 Orlando (2007)
11. Alkoot, F.M., Kittler, J.: Improving the performance of the product fusion strategy. In: ICPR, vol. 02, pp. 164–167. IEEE Computer Society, Los Alamitos, CA, USA (2000)
12. Tax, D., van Breukelen, M., Duin, R.: Combining multiple classifiers by averaging or by multiplying? Pattern Recognition 33, 1475–1485 (2000)
13. Matas, J., Hamouz, M.: K.Jonsson, Kittler, J., Li, Y., Kotropoulos, C.,Tefas, A., Pitas, I., Tan, T., Yan, H., Smeraldi, F., Begun, J., Capdevielle, N., Gerstner, W., Ben-Yacoub, S., Abdeljaoued, Y., Mayoraz, E.: Comparison of face verification results on xm2vts database. In: Proceedings of SPIE. Pattern Recognition. vol. 6539 Orlando (2007)
14. Messer, K., Kittler, J., Short, J., Heusch, G., Cardinaux, F., Marcel, S., Rodriguez, Y., Shan, S., Su, Y., Gao, W.: Performance characterisation of face recognition algorithms and their sensitivity to severe illumination changes. In: Zhang, D., Jain, A.K. (eds.) Advances in Biometrics. LNCS, vol. 3832, pp. 1–11. Springer, Heidelberg (2005)
15. Heusch, G., Rodriguez, Y., Marcel, S.: Local binary pattern as an image preprocessing face authentication. In: Proc. FGR 2006, Washington, DC, 9–14 (2006)
16. Kittler, J., Li, Y., Matas, J.: On matching score for lda-based face verification. In: BMVC (2000)
17. Reynolds, D.A., Quatieri, T., Dunn, T.: Speaker verification using adapted guassian mixture models. In: Digital Signal Processing, pp. 19–41 (2000)
18. Cardinaux, F., Sanderson, C., Bengio, S.: User authentication via adapted statistical models of face images. In: IEEE Trans. on Signal Processing, pp. 361–373 (January 2006)
19. Gross, R., Brajovic, V.: An image preprocessing algorithm for illumination invariant face recognition. In: AVBPA 2003, 10–18 (2003)
20. Bonastre, J.F., Wils, F., Meignier, S.: ALIZE, a free toolkit for speaker recognition. In: Proc. IEEE International Conference on Speech, Acoustics and Signal Processing, Philadelphia pp. 73–740 (2005)
21. Richiardi, J., Prodanov, P., Drygajlo, A.: Speaker verification with confidence and reliability measures. In: Proc. 2006 IEEE International Conference on Speech, Acoustics and Signal Processing, Toulouse, France (May 2006)