Speaker Verification Using Adapted User-Dependent Multilevel Fusion

Julian Fierrez-Aguilar, Daniel Garcia-Romero, Javier Ortega-Garcia, and Joaquin Gonzalez-Rodriguez

Biometrics Research Lab./ATVS, Escuela Politecnica Superior, Universidad Autonoma de Madrid, Campus de Cantoblanco, C/ Francisco Tomas y Valiente 11, 28049 Madrid, Spain
{julian.fierrez, javier.ortega, joaquin.gonzalez}@uam.es

Abstract. In this paper we study the application of user-dependent score fusion to multilevel speaker recognition. After reviewing related works in multimodal biometric authentication, a new score fusion technique is described. The method is based on a form of Bayesian adaptation to derive the personalized fusion functions from prior user-independent data. Experimental results are reported using the MIT Lincoln Laboratory's multilevel speaker verification system. It is experimentally shown that the proposed adapted fusion method outperforms both user-independent and non-adapted user-dependent fusion approaches.

1 Introduction

The state of the art in speaker recognition has been widely dominated during the past decade by the Gaussian Mixture Model (GMM) approach working at the short-time spectral level [1]. Recently, new approaches based on Support Vector Machines (SVM) [2] are achieving similar performance, working also at the spectral level. These new techniques provide complementary information for the verification task, which has been exploited by the use of score fusion techniques [3].

On the other hand, higher levels of information conveyed in the speech signal have shown promising discriminative capabilities among speakers, and are a major goal of present speaker recognition research efforts. Some examples in this regard are the SuperSID project [4], and the MIT Lincoln Laboratory's (MIT-LL) speaker recognition system [5] applied to the 2004 NIST Speaker Recognition Evaluation (SRE) [6]. Since the inclusion of the extended data task in the 2002 NIST SRE, major advances have been done in finding, characterizing and modelling new high-level sources of speaker information. However, once the similarity scores from each individual system have been computed, little emphasis has been placed in developing new fusion approaches that take into account the speaker specificities [7].

Related works combining different sources of information for the person verification task are found in the multimodal biometric authentication literature [8].

N.C. Oza et al. (Eds.): MCS 2005, LNCS 3541, pp. 356–365, 2005.

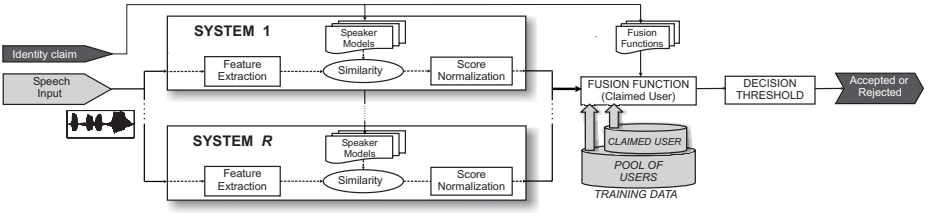

Fig. 1. System model of adapted user-dependent multilevel speaker verification

In this area, it has recently been shown [9, 10, 11, 12, 13] that using personalized fusion functions leads to improved verification performance, when some constraints on the number of training samples are considered. Motivated by the speaker specificities present in the speaker recognition problem [7], the present work is focused on studying user-dependent fusion techniques, and their application to multilevel speaker verification.

This paper is structured as follows. Related works on user-dependent fusion strategies found in the multimodal biometric authentication literature are reviewed in Sect. 2. A new adapted user-dependent score fusion strategy well suited to the common case of small training set size is described in Sect. 3 (see Fig. 1 for the system model). Experiments validating the proposed approach using the established multilevel speaker recognition system from MIT-LL on standard data from NIST SRE evaluations are reported in Sect. 4. Conclusions are finally drawn in Sect. 5.

2 User-Dependent Fusion in Biometric Authentication

The idea of user-dependent fusion in multiple classifier approaches for biometric authentication has probably been introduced in [9], and is receiving increasing attention in the multimodal biometric authentication literature [10, 11, 12, 13, 14, 15, 16].

In the preceding work [9], user-independent weighted linear combination of similarity scores was demonstrated to be improved by using user-dependent weights. A trained user-dependent scheme using support vector machines was subsequently presented in [10], also showing enhanced performance as compared to user-independent fusion. Other attempts to personalized fusion include: using the claimed identity index as a feature for Neural Network learning [14], computing user-dependent combination weights using lambness [7] metrics [15], learning user-dependent polynomial fusion functions [12], and using personalized score normalization techniques based on Fisher ratios [16] prior to user-independent fusion.

The use of general information in user-dependent fusion schemes has recently been introduced [11, 13]. The idea of adapted learning is based on the fact that the amount of available training data in user-dependent learning is usually not

sufficient and representative enough to guarantee good parameter estimation. To cope with this lack of robustness derived from partial knowledge, general user-independent information is considered as prior information from which the user-dependent fusion scheme is built [17].

In the present paper, we describe an efficient adaptation technique based on Bayesian learning [13], and study its application to multilevel speaker verification.

3 Bayesian Adaptation for User-Dependent Fusion

Let the similarity scores $x \in \mathbb{R}$ provided by each one of the R individual systems be combined into a multilevel score $\mathbf{x} = [x_1, \ldots, x_R]'$. Let the fusion training set be $X = (\mathbf{x}_i, y_i)_{i=1}^N$, where N is the number of multilevel scores in the training set, and $y_i \in \{\omega_0, \omega_1\} = \{\text{Impostor}, \text{Client}\}$. Impostor and client score distributions are modelled as the multivariate Gaussians $p(\mathbf{x}|\omega_0) = N(\mathbf{x}|\boldsymbol{\mu}_0, \boldsymbol{\sigma}_0^2)$ and $p(\mathbf{x}|\omega_1) = N(\mathbf{x}|\boldsymbol{\mu}_1, \boldsymbol{\sigma}_1^2)$, respectively[1]. The fused score s_T of a multilevel test $\mathbf{x}_T$ is defined as follows

$$s_T = f(\mathbf{x}_T) = \log p(\mathbf{x}_T|\omega_1) - \log p(\mathbf{x}_T|\omega_0) \tag{1}$$

which is known to be a Quadratic Discriminant function consistent with Bayes estimate for the case of equal impostor and client prior probabilities [18]. The score distributions are estimated using the available training data.

In the user-independent case, the global training set X_{G} includes scores from a pool of users, and the resulting global fusion rule, $f_{\text{G}}(\mathbf{x})$, is obtained by using the standard Maximum Likelihood criterion [18] for estimating $\{\boldsymbol{\mu}_{\text{G},0}, \boldsymbol{\sigma}_{\text{G},0}^2\}$ and $\{\boldsymbol{\mu}_{\text{G},1}, \boldsymbol{\sigma}_{\text{G},1}^2\}$. In the user-dependent case, a different local fusion function, $f_{j,\text{L}}(\mathbf{x})$, is obtained for each client enrolled in the system by using Maximum Likelihood estimates, $\{\boldsymbol{\mu}_{j,\text{L},0}, \boldsymbol{\sigma}_{j,\text{L},0}^2\}$ and $\{\boldsymbol{\mu}_{j,\text{L},1}, \boldsymbol{\sigma}_{j,\text{L},1}^2\}$, computed from a set of development scores X_j of the specific client j.

The proposed adapted fusion function, $f_{j,\text{A}}(\mathbf{x})$, trades off the general knowledge provided by X_{G}, and the user specificities provided by X_j, through Maximum a Posteriori density estimation [19]. This is done by adapting the sufficient statistics as follows [1]:

$$\begin{aligned} \boldsymbol{\mu}_{j,\text{A},i} &= \alpha_i \boldsymbol{\mu}_{j,\text{L},i} + (1-\alpha_i)\boldsymbol{\mu}_{\text{G},i} \\ \boldsymbol{\sigma}_{j,\text{A},i}^2 &= \alpha_i(\boldsymbol{\sigma}_{j,\text{L},i}^2 + \boldsymbol{\mu}_{j,\text{L},i}^2) + (1-\alpha_i)(\boldsymbol{\sigma}_{\text{G},i}^2 + \boldsymbol{\mu}_{\text{G},i}^2) - \boldsymbol{\mu}_{j,\text{A},i}^2 \end{aligned} \tag{2}$$

For each class $i = \{0 = \text{Impostor}, 1 = \text{Client}\}$, a data-dependent adaptation coefficient

$$\alpha_i = \frac{N_i}{N_i + r} \tag{3}$$

[1] We use diagonal covariance matrixes, so $\boldsymbol{\sigma}^2$ is shorthand for $\text{diag}(\Sigma)$. Similarly, $\boldsymbol{\mu}^2$ is shorthand for $\text{diag}(\boldsymbol{\mu}\boldsymbol{\mu}')$.

is used [1], where N_i is the number of local training scores in class i, and r is a fixed relevance factor.

4 Experiments

4.1 Baseline Systems

In the present paper, the scores submitted by the MIT-LL [5] for the 2004 NIST SRE extended data task [6] are used. These scores were computed by using seven systems with speaker information from spectral level, pitch and duration prosodic behavior, and phoneme and word usage. These different types of information were modelled and classified using Gaussian Mixture Models (GMM), Support Vector Machines (SVM) and n-gram language models. In the following, a brief description of the main features of each individual system is presented:

MFCC_GMM. The system is based on a likelihood ratio detector with target and alternative probability distributions modeled by GMMs [1]. A Universal Background Model GMM is used as the alternative hypothesis model, and target models are derived using Bayesian adaptation. The techniques of feature mapping [20] and T-norm [21] are also used.

MFCC_SVM. The spectral SVM system uses a novel sequence kernel [2]. The sequence kernel compares entire utterances using a generalized linear discriminant. It uses the same front-end processing as the MFCC_GMM system.

PHONE_SVM. The SVM phone system uses a kernel for comparing conversation sides based upon methods from information retrieval. Sequences of phones are converted to a vector of probabilities of occurrences of terms, and co-occurrences of terms (bag of unigram, and bag of bigrams, respectively). A weighting based upon a linearization of likelihoods is then used to compare vectors for SVM training.

PHONE_NGM. A phone n-gram system was developed using the output of the MIT-LL phone recognizer. This system used the n-gram approach proposed in [22].

PROSODY_SLOPE. To capture prosodic differences in the realization of intonation, rhythm, and stress, the $F0$ and energy contours are converted into a sequence of tokens reflecting the joint state of the contours (rising or falling). A n-gram system is then used to model and classify distinctive token patterns from token sequences [23].

PROSODY_GMM. The aim of this system is to capture the characteristics of the $F0$ and short-term energy features distribution. This system is based on a likelihood ratio detector that uses adapted GMMs for estimating the likelihoods [24].

WORD_NGM. A word n-gram (idiolect) system was developed using the speech-to-text output from the BBN Byblos real-time system. This system used the idiolect word n-gram approach proposed in [22].

4.2 Database and Experimental Protocol

The experiments presented below were conducted on the 8sides-1side set of the 2004 NIST SRE corpus [6]. This database comprises conversational telephone speech in five different languages (English, Spanish, Russian, Arabic and Mandarin) over three different channels (cellular, cordless and landline), and four types of transducers (speaker-phone, head-mounted, ear-bud, and hand-held). Speaker models were trained with 8 single channel conversation sides of approximately five minutes total duration. Test segments consist of one side of the conversations. All trials were performed between two speakers of the same gender.

In order to provide a development set (DEV) for the experiments, data from Switchboard II phases $1-5$ were used to mimic the conditions in the 8sides-1side set of the 2004 NIST SRE corpus.

The following subsets of the 8sides-1side set were defined for the experiments:

ALL5. All speaker models with at least 5 genuine and 10 impostor attempts. In this way, ALL5 consists of 830 genuine and 4614 impostor attempts of 118 different speaker models.

COMMON5. All speaker models with $> 75\%$ of English enrollment, and at least 5 client and 10 impostor attempts. In this way, COMMON5 consists of 136 genuine and 378 impostor attempts of 19 different speaker models.

Three different types of experiments have been conducted:

User-Independent Fusion. Training on DEV data.

User-Dependent Fusion. For each user and each multilevel test score, 4 different genuine and 9 different impostor multilevel scores of the user at hand are randomly selected (different to the tested one). Local training is performed on the randomly selected multilevel scores. For each multilevel test score, 5 runs of the random sampling are performed.

Adapted User-Dependent Fusion. For each user and each multilevel test score, 4 different genuine and 9 different impostor multilevel scores of the user at hand are randomly selected (different to the tested one). Global training is performed on DEV data whereas local training is carried out on the randomly selected multilevel scores. For each multilevel test score, 5 runs of the random sampling are performed.

4.3 Results

Verification performance of the seven individual systems, along with various user-independent combinations, are given in Tables 1 and 2 for the ALL5 and COMMON5 datasets, respectively. Spectral level systems perform remarkably better than the other systems, and their combination with the high-level system WORD_NGM leads to enhanced performance. Worth noting, not all combinations provide improved performance over the best system, and the relative improvement between the best fused system and the best individual system is not

Table 1. Verification performance on **ALL5** dataset with **user-independent fusion** based on Quadratic Discriminant. EERs in %

information level	system label	individual performance	unilevel fusion	multilevel fusion levels	best/level	all/level
	MFCC_GMM	8.67		*12*	9.28	8.79
1	MFCC_SVM	**7.70**	**7.39**	*13*	7.83	6.98
	PHONE_SVM	16.90		*14*	7.46	**6.91**
2	PHONE_NGM	22.16	18.21	*123*	9.05	8.07
	PROSODY_SLOPE	20.86		*124*	8.98	8.25
3	PROSODY_GMM	22.51	16.76	*134*	7.59	6.98
4	WORD_NGM	22.70		*1234*	9.19	7.96

Table 2. Verification performance on **COMMON5** dataset with **user-independent fusion** based on Quadratic Discriminant. EERs in %

information level	system label	individual performance	unilevel fusion	multilevel fusion levels	best/level	all/level
	MFCC_GMM	5.98		*12*	3.69	3.06
1	MFCC_SVM	**3.06**	**3.56**	*13*	4.32	3.56
	PHONE_SVM	10.31		*14*	3.56	**2.93**
2	PHONE_NGM	18.32	10.94	*123*	3.56	3.56
	PROSODY_SLOPE	22.14		*124*	4.32	**2.93**
3	PROSODY_GMM	19.08	14.63	*134*	3.06	**2.93**
4	WORD_NGM	20.61		*1234*	3.56	3.19

very high (10% and 4% on ALL5 and COMMON5 respectively). Finally, performance on COMMON5 is remarkably better than performance on ALL5, specially for the spectral and phonetic systems (60% and 39% relative improvements in the best system of each level respectively).

Verification performance using non-adapted user-dependent fusion is given in Tables 3 and 4 for the ALL5 and COMMON5 datasets, respectively. The same behavior found in user-independent fusion is also observed here, obtaining similar performance figures. In particular, relative improvements between the best fused system and the best individual system are 9% and 12% for ALL5 and COMMON5 datasets, respectively.

Verification performance using the proposed adapted user-dependent fusion approach ($r = 1$) is given in Tables 5 and 6 for the ALL5 and COMMON5 datasets, respectively. In this case, all combinations are better than the best individual system, which is outperformed significantly by the best combination (i.e., spectral and lexical systems). In particular, relative improvement between the best fused system and the best individual system are 31% and 61% for ALL5 and COMMON5, respectively. Also worth noting, the unilevel combination of the two spectral level systems gives an interesting combination pair (31%

Table 3. Verification performance on **ALL5** dataset with **user-dependent fusion** based on Quadratic Discriminant. EERs in %

information level	system label	individual performance	unilevel fusion	multilevel fusion levels	best/level	all/level
	MFCC_GMM	8.67		*12*	7.86	7.22
1	MFCC_SVM	**7.70**	**6.84**	*13*	8.27	8.15
	PHONE_SVM	16.90		*14*	8.04	**6.98**
2	PHONE_NGM	22.16	15.74	*123*	8.08	7.99
	PROSODY_SLOPE	20.86		*124*	8.46	7.37
3	PROSODY_GMM	22.51	18.46	*134*	8.57	8.04
4	WORD_NGM	22.70		*1234*	8.44	8.11

Table 4. Verification performance on **COMMON5** dataset with **user-dependent fusion** based on Quadratic Discriminant. EERs in %

information level	system label	individual performance	unilevel fusion	multilevel fusion levels	best/level	all/level
	MFCC_GMM	5.98		*12*	4.40	2.98
1	MFCC_SVM	**3.06**	**2.95**	*13*	5.98	4.99
	PHONE_SVM	10.31		*14*	5.42	**2.70**
2	PHONE_NGM	18.32	11.60	*123*	5.60	4.43
	PROSODY_SLOPE	22.14		*124*	5.04	2.77
3	PROSODY_GMM	19.08	18.99	*134*	5.85	3.66
4	WORD_NGM	20.61		*1234*	5.60	3.66

Table 5. Verification performance on **ALL5** dataset with **adapted user-dependent fusion** based on Quadratic Discriminant ($r = 1$). EERs in %

information level	system label	individual performance	unilevel fusion	multilevel fusion levels	best/level	all/level
	MFCC_GMM	8.67		*12*	6.25	5.66
1	MFCC_SVM	**7.70**	**5.35**	*13*	5.85	5.40
	PHONE_SVM	16.90		*14*	6.14	**5.36**
2	PHONE_NGM	22.16	13.61	*123*	5.92	5.39
	PROSODY_SLOPE	20.86		*124*	6.72	5.61
3	PROSODY_GMM	22.51	15.08	*134*	5.95	**5.32**
4	WORD_NGM	22.70		*1234*	6.16	5.37

and 34% relative improvement over the best system for ALL5 and COMMON5, respectively). The effect of varying the relevance factor of the adapted fusion scheme on the verification performance is shown in Fig. 2. A good working point is found at $r = 1$.

Table 6. Verification performance on **COMMON5** dataset with **adapted user-dependent fusion** based on Quadratic Discriminant ($r = 1$). EERs in %

information level	system label	individual performance	unilevel fusion	multilevel fusion levels	best/level	all/level
	MFCC_GMM	5.98		*12*	2.80	2.06
1	MFCC_SVM	**3.06**	**2.03**	*13*	2.37	2.27
	PHONE_SVM	10.31		*14*	2.49	**1.20**
2	PHONE_NGM	18.32	8.70	*123*	2.77	2.11
	PROSODY_SLOPE	22.14		*124*	2.92	1.68
3	PROSODY_GMM	19.08	15.65	*134*	1.91	1.66
4	WORD_NGM	20.61		*1234*	2.42	1.32

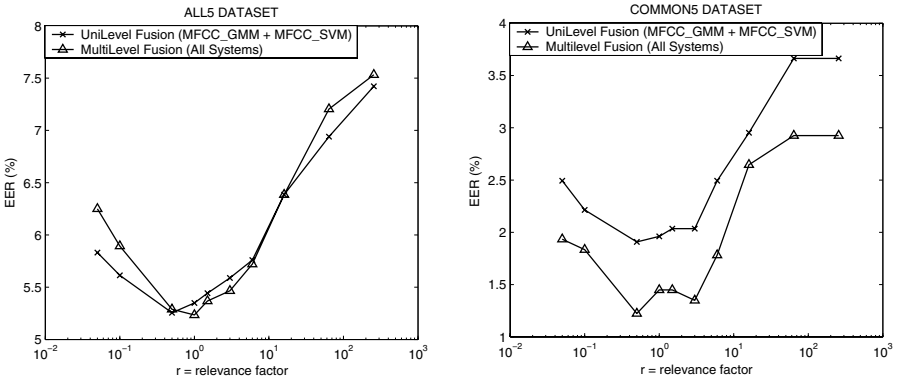

Fig. 2. Verification performance of the adapted fusion scheme on ALL5 (left) and COMMON5 (right) data sets for varying relevance factor

Verification performance results comparing the individual systems to the studied fusion strategies are summarized in Fig. 3 as DET plots [25].

5 Discussion and Conclusions

It can be argued against user-dependent fusion that training data scarcity is a major drawback for its success. In this paper, it has been demonstrated that the performance of multilevel speaker verification is improved in an standard evaluation scenario by considering user-dependent information at the fusion level. This has been achieved by using a novel user-dependent fusion technique based on Bayesian adaptation of the fusion functions and only a few training score samples from each user. Nevertheless, although we have used an un-biased cross-validation experimental procedure, it must be emphasized that we have used post-evaluation results for adapting to the user specificities. The study of the

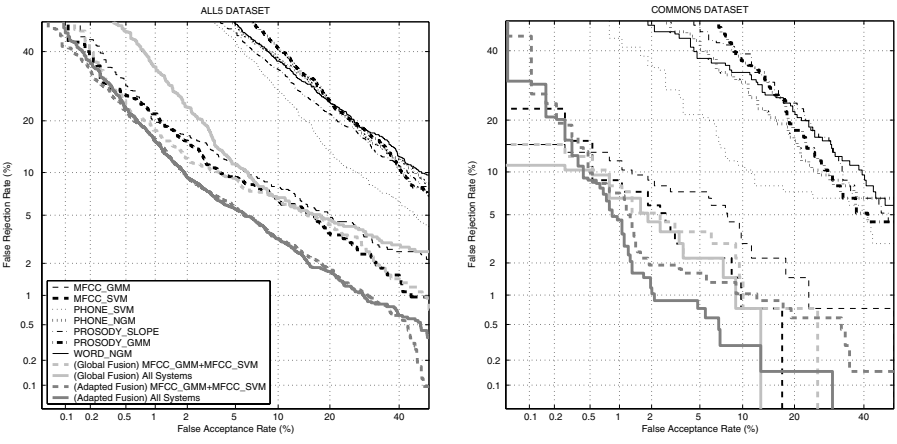

Fig. 3. Verification performance of the individual systems and the adapted fusion scheme on ALL5 (left) and COMMON5 (right) data sets

case of using only the available training data will be addressed in future work. In this regard, it is our belief that for the case of large training set size (such as the 8sides-1side or above scenarios defined by NIST), the use of resampling techniques (e.g., resubstitution, leave-one-out, bootstrap) [26] may result in a significant improvement. As a preliminary justification for this aim, we point out the related work [27], where resampling techniques were applied successfully in the related problem of training user-dependent score normalization techniques applied to signature verification. As a result, the present work is an encouraging starting and reference point for devising personalized fusion schemes with application to multilevel speaker recognition.

Acknowledgements

This work has been supported by the Spanish MCYT projects TIC2003-08382-C05-01 and TIC2003-09068-C01-01. J. F.-A. is also supported by a FPI scholarship from Comunidad de Madrid. The authors would like to thank MIT-LL for providing the speaker recognition scores used in this paper, and also for their substantive comments about the paper.

References

1. Reynolds, D.A., Quatieri, T.F., Dunn, R.B.: Speaker verification using adapted Gaussian mixture models. Digital Signal Processing **10** (2000) 19–41
2. Campbell, W.M.: A SVM/HMM system for speaker recognition. Proc. ICASSP (2003) 209–302

3. Campbell, W.M., Reynolds, D.A., Campbell, J.: Fusing discriminative and generative methods for speaker recognition: Experiments on Switchboard and NFI/TNO field data. Proc. ODYSSEY (2004) 41–44
4. Reynolds, D.A., et al.: The SuperSID project: Exploiting high-level information for high-accuracy speaker recognition. Proc. ICASSP (2003) 784–787
5. Reynolds, D.A., et al.: The 2004 MIT Lincoln Laboratory Speaker Recognition System. Proc. ICASSP (2005) (to appear)
6. NIST SRE Web (http://www.nist.gov/speech/tests/spk/2004/index.htm)
7. Doddington, G., et al.: Sheeps, goats, lambs and wolves: A statistical analysis of speaker performance in the NIST 1998 SRE. Proc. ICSLP (1998)
8. Bigun, E.S., Bigun, J., et al.: Expert conciliation for multi modal person authentication systems by Bayesian statistics. Springer LNCS-1206 (1997) 291–300
9. Jain, A.K., Ross, A.: Learning user-specific parameters in a multibiometric system. Proc. ICIP (2002) 57–60
10. Fierrez-Aguilar, J., et al.: A comparative evaluation of fusion strategies for multimodal biometric verification. Springer LNCS-2688 (2003) 830–837
11. Fierrez-Aguilar, J., et al.: Exploiting general knowledge in user-dependent fusion strategies for multimodal biometric verification. Proc. ICASSP (2004) 617–620
12. Toh, K.A., Jiang, X., Yau, W.Y.: Exploiting local and global decisions for multimodal biometrics verification. IEEE Trans. on SP **52** (2004) 3059–3072
13. Fierrez-Aguilar, J., et al.: Bayesian adaptation for user-dependent multimodal biometric authentication. Pattern Recognition (2005) (to appear)
14. Kumar, A., Zhang, D.: Integrating palmprint with face for user authentication. Proc. MMUA (2003) (available at http://mmua.cs.ucsb.edu/)
15. Snelick, R., et al.: Large scale evaluation of multimodal biometric authentication using state-of-the-art systems. IEEE Trans. PAMI **27** (2005) 450–455
16. Poh, N., Bengio, S.: An Investigation of F-ratio client-dependent normalisation on biometric authentication tasks. Proc. ICASSP (2005) (to appear)
17. Lee, C.H., Huo, Q.: On adaptive decision rules and decision parameter adaptation for automatic speech recognition. Proc. IEEE **88** (2000) 1241–1269
18. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. Wiley (2001)
19. Gauvain, J.L., Lee, C.H.: Maximum a Posteriori estimation for multivariate Gaussian mixture observations of Markov chains. IEEE Trans. on SAP **2** (1994) 291–298
20. Reynolds, D.A.: Channel robust speaker verification via feature mapping. Proc. ICASSP (2003) 53–56
21. Auckenthaler, R., et al.: Score normalization for text-independent speaker verification systems. Digital Signal Processing **10** (2000) 42–54
22. Doddington, G.: Speaker recognition based on idiolectal differences between speakers. Proc. EUROSPEECH (2001) 2521–2524
23. Adami, A., Mihaescu, R., Reynolds, D.A., Godfrey, J.: Modeling prosodic dynamics for speaker recognition. Proc. ICASSP (2003) 788–791
24. Adami, A.G.: Modeling prosodic differences for speaker and language recognition. PhD thesis, OGI (2004)
25. Martin, A., Doddington, G., et al.: The DET curve in assessment of decision task performance. Proc. EUROSPEECH (1997) 1895–1898
26. Jain, A.K., Duin, R.P.W., Mao, J.: Statistical pattern recognition: A review. IEEE Trans. on PAMI **22** (2000) 4–37
27. Fierrez-Aguilar, J., Ortega-Garcia, J., Gonzalez-Rodriguez, J.: Target dependent score normalization techniques and their application to signature verification. IEEE Trans. on SMC-C **35** (2005) (to appear)