PERSON VERIFICATION BY FUSION OF PROSODIC, VOICE SPECTRAL AND FACIAL PARAMETERS

Javier Hernando, Mireia Farrús, Pascual Ejarque, Ainara Garde, Jordi Luque
TALP Research Center, Department of Signal Theory and Communications
Technical University of Catalonia, Barcelona, Spain

Keywords: Biometrics, Multimodality, Fusion, Face, Prosody, Support Vector Machines, Matcher Weighting.

Abstract: Prosodic information can be used successfully for automatic speaker recognition, although most of the speaker recognition systems use only short-term spectral features as voice information. In this work, prosody information is added to a multimodal system based on face and voice characteristics in order to improve the performance of the system. Fusion is carried out by using various fusion strategies and two different fusion techniques: support vector machines and matcher weighting. Results are clearly improved when a previous normalization based on histogram equalization is done before the fusion of the monomodal scores.

1 INTRODUCTION

A multimodal biometric system involves the combination of two or more human characteristics in order to achieve better results than using monomodal recognition (Bolle, Connell et al. 2004). When several biometric traits are used in a multimodal recognition system, fusion is possible at three different levels: feature extraction level, matching score level or decision level. Fusion at the matching score level is usually preferred by most of the systems, which is, in fact, a two-step process: normalization and fusion itself (Indovina, Uludag et al. 2003). Since monomodal scores are usually non-homogeneous, the normalization process transforms the different scores of each monomodal system into a comparable range of values. One conventional affine normalization technique is z-score, which transforms the scores into a distribution with zero mean and unitary variance. Histogram equalization is another method whose purpose is to equalize the statistics of two monomodal biometrics.

After normalization, the converted scores are combined in the fusion process in order to obtain a single multimodal score. In matcher weighting fusion method, each monomodal score is weighted by a factor proportional to the recognition result of the biometric. One of the most currently used fusion techniques is support vector machines (SVM). The SVM algorithm constructs models that contain a large class of neural nets, radial basis function nets and polynomial classifiers as special cases. The algorithm is simple enough to be analyzed mathematically, since it can be shown to correspond to a linear method in a high-dimensional feature space non-linearly related to input space (Hearst 1998).

Prosody is mostly used to refer to speech elements such as tone, rhythm and intensity. The aim of this work is to add prosodic information to the multimodal biometric recognition systems in order to improve the performance of the system. Prosodic, vocal tract spectral and facial scores are fused by using two types of fusion, and different fusion strategies are proposed: score level fusion is carried in one, two or three steps, considering two different combinations in the two-step fusion.

This paper is organized as follows. In the next section the monomodal information sources used in this work are described. The conventional normalization method z-score and histogram equalization are presented in section 3. Matcher weighting fusion technique and support vector machines are reviewed in section 4. Finally, experimental results are shown in section 5 for the fusion combinations of prosodic, vocal tract spectrum and face scores.

Hernando J., Farrús M., Ejarque P., Garde A. and Luque J. (2006).
PERSON VERIFICATION BY FUSION OF PROSODIC, VOICE SPECTRAL AND FACIAL PARAMETERS.
In *Proceedings of the International Conference on Security and Cryptography*, pages 17-23
DOI: 10.5220/0002105200170023

2 MONOMODAL SOURCES

2.1 Voice Information

2.1.1 Spectral Parameters

Spectral parameters are those which only take into account the acoustic level of the signal, like spectral magnitudes, formant frequencies, etc., and they are more related to the physical traits of the speaker. Cepstral coefficients are the usual way of representing the short-time spectral envelope of a speech frame in current speaker recognition systems. These parameters are the most prevalent representations of the speech signal and contain a high degree of speaker specificity. The conventional mel-cepstrum coefficients come from a set of mel-scaled log filter bank energies (LFBE) *S(k)*. The sequence of cepstral coefficients is quasi-uncorrelated and compact representation of speech spectra. However, cepstral coefficients have some disadvantages: they do not possess a clear and useful physical meaning as LFBE have, they require a linear transformation from either LFBE or the LPC coefficients and in continuous observation Gaussian density HMM with diagonal covariance matrices the shape of the cepstral window has no effect so that only its length . In order to overcome them, (Nadeu, Mariño et al. 1996) presents an alternative that consists of a simple linear processing on the LFBE domain. The transformation of the sequence *S(k)* to cepstral coefficients is avoided by filtering that sequence. This operation is called frequency filtering (FF) to denote that the convolution is performed in the frequency domain.

2.1.2 Prosodic Parameters

Humans tend to use several linguistic levels of information like lexical, prosodic or phonetic features to recognize others with voice. Prosodic parameters are called suprasegmental features since the segments affected (syllables, words and phrases) are larger than phonetic units. These features are basically manifested as durations, tone and intensity variation.

Although these features don't provide very good results when used alone, they give complementary information and improve the results when they are fused with vocal tract spectrum based systems. Moreover, some of these features have the advantage of being more robust to noise than the low-level ones (Carey, Parris et al. 1996). Spectral patterns can be affected by frequency features of the transmission channel, and spectral information depends also on the speech level and the distance between the speaker and the array, while fundamental frequency is unaffected by such variations (Atal 1972). The prosodic recognition system used in this task was constituted by a total of 9 prosodic features already used in (Peskin, Navratil et al. 2003); i.e. three features related to word and segmental durations: number of frames per word and length of word-internal voiced and unvoiced segments, and six more features related to pitch: mean pitch, maximum pitch, minimum pitch, pitch range, pitch "pseudo-slope" defined as (last F0 - first F0)/(number of frames in word) and average slope over all segments of piecewise linear stylization of F0, all of them averaged over all words with voiced frames.

2.2 Face Information

Facial recognition systems are based on the conceptualization that a face can be represented as a collection of sparsely distributed parts: eyes, nose, cheeks, mouth, etc. Non-negative matrix factorization (NMF), introduced in (Lee and Seung 2001), is an appearance-based face recognition technique based on the conventional component analysis techniques which does not use the information about how the various facial images are separated into different facial classes. The most straightforward way in order to exploit discriminant information in NMF is to try to discover discriminant projections for the facial image vectors after the projection. The face recognition scores used in this work have been calculated in this way with the NMF-faces method (Zafeiriou, Tefas et al. 2005), in which the final basis images are closer to facial parts.

3 NORMALIZATION AND HISTOGRAM EQUALIZATION

One of the most conventional normalization methods is z-score (ZS), which normalizes the global mean and variance of the scores of a monomodal biometric. Denoting a raw matching score as a from the set A of all the original monomodal biometric scores, the z-score normalized biometric x_{ZS} is calculated according to

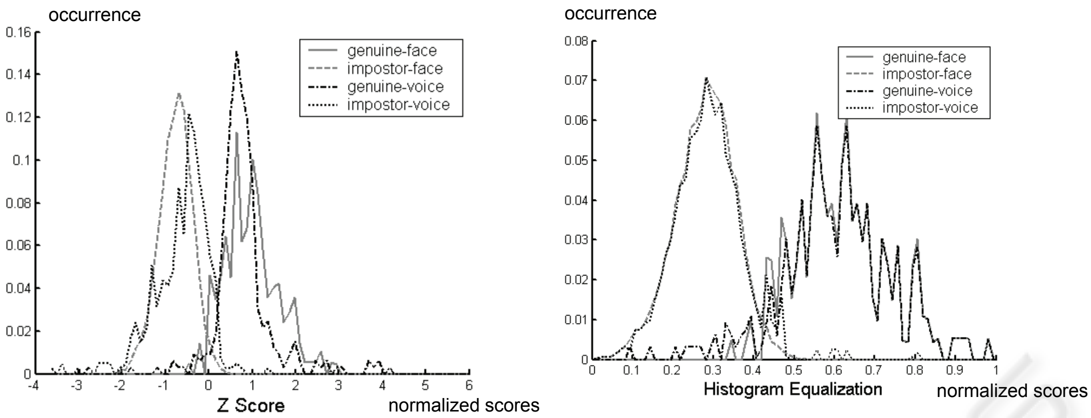

Figure 1: Scores distribution of face and speech biometrics after the presented normalizations.

$$x_{ZS} = \frac{a - mean(A)}{std(A)} \qquad (1)$$

where *mean(A)* is the statistical mean of *A* and *std(A)* is the standard deviation.

Multimodal variances are reduced when the variances of the monomodal scores are similar. Unfortunately, these variances are not usually similar. In order to solve this problem and equalization of the histograms of the monomodal scores is proposed in this paper as a non affine normalization process. Thus, the genuine and impostor statistics and, consequently, the variances, will probably be equalized.

Histogram equalization (HE) or cumulative distribution function (CDF) equalization is a general non parametric method to make the CDF of some given data match to a reference distribution. HE is a widely used non linear method designed for the enhancement of images. HE employs a monotonic, non linear mapping which re-assigns the intensity values of pixels in the input image in order to control the shape of the output image intensity histogram to achieve a uniform distribution of intensities or to highlight certain intensity levels (Torre, Peinado et al. 2005).

CDF equalization method was mainly developed for the speech recognition adaptation approaches or for the correction of non linear effects typically introduced by speech systems such as: microphones, amplifiers, clipping and boosting circuits and automatic gain control circuits. The principle of this method is to find a non linear transformation to reduce the mismatch of the statistics of two signals. By means of the CDF a transformation that maps the distribution of a signal back to the distribution of a reference signal is defined.

In this work, the statistical matching technique matches de CDF obtained from the speaker verification scores and the CDF obtained from the face verification scores, both evaluated over the training data. The designed equalization takes as a reference the histogram of the biometric with a better accuracy, which is expected to have lower separate variances, in order to obtain a bigger variance reduction.

In figure 1 two histograms of face and voice scores are plotted after the application of the presented normalizations in order to compare the transformations produced by each of them.

4 FUSION TECHNIQUES AND SUPPORT VECTOR MACHINES

In matcher weighting (MW) fusion, one of the most conventional fusion techniques, each monomodal score is weighted by a factor proportional to the recognition rate, so that the weights for more accurate matchers are higher than those of less accurate matchers. When using the Equal Error Rates (EER) the weighting factor for every biometric is proportional to the inverse of its EER. Denoting w^m and e^m the weighting factor and the EER for the mth biometric x^m and M the number of biometrics, the fused score u is expressed as:

$$u = \sum_{m=1}^{M} w^m x^m \text{ , where } w^m = \frac{\frac{1}{e^m}}{\sum_{m=1}^{M} \frac{1}{e^m}} \qquad (2)\ (3)$$

A support vector machine (SVM) is a binary classifier based on a learning fusion technique (Cristianini and Shawe-Taylor 2000). Learning based fusion can be treated as a pattern classification problem in which the scores obtained with individual classifiers are seen as input patterns to be labelled as 'accepted' or 'rejected'. Given a linearly separable two-class training data, the aim is to find an optimal hyperplane that splits input data in two classes: 1 and -1 (the target values that correspond to the 'accepted' and 'rejected' labels respectively) maximizing the distance of the hyperplane to the nearest data of each class. The optimal hyperplane is then constructed in the feature space, creating a non linear boundary in the input space.

5 RECOGNITION EXPERIMENTS

In section 5.1 some preliminary experiments involving face and speech multimodal identification by using matcher weighting fusion are presented. The prosody, vocal tract spectrum and face based recognition systems used in our fusion experiments are presented in section 5.2. Experimental results obtained by using SVM and matcher weighting fusion methods and combined according to three different fusion strategies are shown in section 5.3.

5.1 Preliminary Experiments

In this section, the audio, video and multimodal person identification experiments in the CLEAR'06 Evaluation Campaign (http://www.clear-evaluation.org) are presented. A set of audiovisual recordings of seminars have been used, consisting of short video sequences and matching far-field audio recordings.

For the acoustic speaker identification, 20 Frequency Filtering parameters were generated with a frame size of 30ms and a shift of 10ms, and 20 corresponding delta and acceleration coefficients were included. Gaussian Mixture Models (GMM) with diagonal covariance matrices were used.

For the visual identification, a projection-based technique was developed, which combines the information of several images to perform the recognition (Luque, Morros et al. 2006). Models for all the users were created using segments of 15 seconds. The XM2VTS database was used as training data for estimating the projection matrix. For each test segment, the face images of the same user were gathered into a group; then, for each group, the system compared the images with the person model.

Segments of different durations (1 and 5 seconds) corresponding to 26 personal identities have been used for testing. Table 1 shows the correct identification rates obtained for both audio and video monomodalities and the fusion identification rate. The identification results are clearly improved when the multimodal fusion technique is used.

Table 1: Correct identification for both audio and video systems and multimodal fusion.

	Correct identification (%)		
duration (s)	Speech	Video	Fusion
1	75.0	20.2	76.8
5	89.3	21.4	92.0

5.2 Experimental Setup

A chimerical database has been created by relating the speakers of the Switchboard-I speech database (Godfrey, Holliman et al. 1990) to the faces of the video and speech XM2VTS database (Lüttin, Maître et al. 1998) of the University of Surrey. The Switchboard-I database has been used for the speaker recognition experiments. It is a collection of 2430 two-sided telephone conversations among 543 speakers (302 male, 241 female) from all areas of the United States. Each conversation of the Switchboard-I database contains two conversation sides. For both recognition systems each speaker model was trained with 8 conversation sides and tested according to NIST's 2001 Extended Data task.

Speech scores have been obtained by using two different systems: a voice spectrum based speaker recognition system and a prosody based recognition system. The spectrum based recognition system was the same GMM system used in the preliminary experiments and the UBM was a 32-component Gaussian mixture model trained with 116 conversations.

In the prosody based recognition system a 9 prosodic feature vector was extracted for each conversation side. Mean and standard deviation were computed for each individual feature. The system was tested with one conversation side, computing the distance between the test feature vector and the k feature vectors of the claimed speaker, using the k-Nearest Neighbour method with k=3 and the symmetrized Kullback-Leibler divergence.

XM2VTS database was used for the face recognition experiments. It is a multimodal database consisting of face images, video sequences and speech recordings of 295 subjects. Only the face images (four frontal face images per subject) were used in our experiments. In order to evaluate verification algorithms on the database, the evaluation protocol described in (Lüttin, Maître et al. 1998) was followed. The well-known Fisher discriminant criterion was constructed as (Belhumeur, Hespanha et al. 1997) in order to discover discriminant linear projections and to obtain the facial scores.

The fusion experiments combine the scores for the users of all the recognition systems. A chimerical database with 30661 users has been created by combining 170 users of the Switchboard-I database and 270 users of the XM2VTS database. A total of 46500 experiments were carried out.

5.3 Verification Results

Table 2 shows the EER obtained for each prosodic feature used in the prosody based recognition system. As it can be seen, features based on pitch measurements achieve the best results.

Table 2: EER for each prosodic feature.

Features	EER (%)
log(#frames/word)	30.3
length of word-internal voiced segments	31.5
length of word-internal unvoiced segments	31.5
log(mean F0)	19.2
log(max F0)	21.3
log(min F0)	21.5
log(range F0)	26.6
pitch “pseudo slope”	38.3
slope over PWL stylization of F0	28.7

The EER obtained in each monomodal recognition system, in the fusion of prosodic and voice spectral scores and in the fusion of spectral and facial scores when using SVM and MW methods are shown in table 3.

Table 3: EER for monomodal and bimodal systems.

Source	EER (%)				
		SVM		ZS-MW	
Prosody		14.65	6.84	15.66	7.44
Voice spect.	10.10	0.99		1.83	
Face	2.06				

Note that fusion was only used in the monomodal prosodic system, where 9 different prosodic scores where fused, and in both bimodal systems. No fusion was involved in the monomodal voice spectral and facial recognition systems. It can be seen that the performance of MW fusion is slightly worse that the SVM.

5.3.1 One-step Fusion

One-step fusion (figure 2) consists in fusing at once all the scores obtained from the 11 extracted features: prosodic scores (PS) obtained from 9 prosodic parameters, voice spectral scores (SS) obtained from spectral parameters and face scores (FS) obtained from image face parameters. The EER obtained for both types of fusion (SVM and MW with ZS normalization) are shown in table 4.

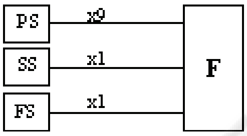

Figure 2: One-step fusion.

Table 4: EER for one-step fusion.

F	EER (%)
SVM	0.840
ZS-MW	1.320

The results show, once again, that SVM technique outperforms the conventional MW method wit ZS normalization. Furthermore, by using prosodic features the results of the bimodal voice spectrum and face recognition system are clearly improved.

5.3.2 Two-step Fusion

Two-step fusion consists in fusing all the scores obtained from the 11 parameters in two consecutive steps. In this kind of fusion two different configurations have been considered (figure 3). In the first configuration (config. A) the scores of all the speech features (9 prosodic features and 1 spectral feature) are previously fused and the obtained results are then fused again with the facial scores. In the second configuration (config. B) the scores of the 9 prosodic features are previously fused and the obtained results are then fused again with voice spectral and facial scores.

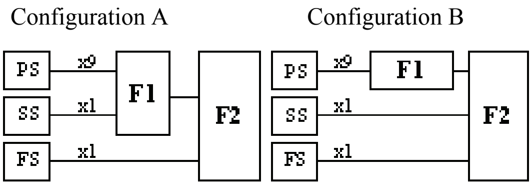

Figure 3: Two configurations of two-step fusion.

Table 5 shows the EER for both configurations of the proposed two-step fusion. It can be seen that

SVM outperforms, once again, the conventional ZS technique.

Table 5: EER (%) for two-step fusion.

F1	F2	Config. A	Config. B
SVM	SVM	0.987	**0.647**
ZS-MW	ZS-MW	2.054	1.493
SVM	ZS-MW	1.583	1.303
ZS-MW	SVM	1.880	0.785

5.3.3 Three-step Fusion

Since the previous tables show that the best results are achieved by SVM fusion, another possibility is now considered: a three-step fusion with SVM. First of all, scores related to the 9 prosodic features are fused by SVM. The obtained results are then fused with voice spectral scores, and the new results are, once again, fused with the facial scores, as it can be seen in figure 4. EER for three-step SVM fusion are shown in table 6.

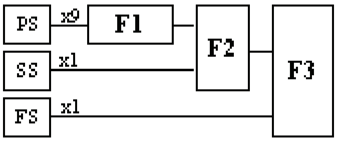

Figure 4: Three-step fusion.

Table 6: EER for three-step fusion.

F1, F2, F3	EER (%)
SVM	0.868

5.3.4 Histogram Equalization

In order to analyze how the fusion process is influenced by a previous histogram equalization of the scores, this normalization method is applied to the fusion strategy where the best results were achieved, i.e. configuration B in the two-step fusion. In table 7 the results obtained with equalized and non equalized scores are compared. It can be clearly seen that the results are always improved when histogram equalization is previously applied.

Table 7: EER (%) with equalized and non equalized scores in the best fusion strategy.

F1	F2	non equalized	equalized
SVM	SVM	0.647	0.630
ZS-MW	ZS-MW	1.493	0.987
SVM	ZS-MW	1.303	0.886
ZS-MW	SVM	0.785	0.774

6 CONCLUSIONS

The performance of a bimodal system based on facial and spectral information is clearly improved in this work when prosodic information is added to the system. In our experiments the use of SVM outperforms the results obtained by fusion with the matcher weighting technique. The way how the scores are fused is relevant for the performance of the system. The best results have been obtained when prosodic scores are previously fused and the resulting scores are fused at once with spectral and face scores. Furthermore, a previous histogram equalization as a normalization technique improves the results obtained with non equalized scores. It has also been observed that a previous fusion of the voice information (spectral and prosodic scores) does not contribute to the improvement of the system.

ACKNOWLEDGEMENTS

The authors would like to thank A. Temko for his helpful discussions on the fusion techniques used in this work, and Dr. A. Tefas, who has provided the face recognition scores.

This work has been partially supported by the European Union (under CHIL IST-2002-506909 and BIOSEC IST-2002-001766) and by the Spanish Government (under ACESCA project TIN2005-08852 and grant AP2003-3598).

REFERENCES

http://www.clear-evaluation.org/

Atal, B. S. (1972). "Automatic speaker recognition based on pitch contours." Journal of the Acoustical Society of America **52**: 1687-1697.

Belhumeur, P. N., J. P. Hespanha, et al. (1997). "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection." IEEE Transactions on Pattern Analysis and Machine Intelligence **19**(7): 711-720.

Bolle, R. M., J. H. Connell, et al. (2004). Guide to Biometrics. New York, Springer.

Carey, M. J., E. S. Parris, et al. (1996). Robust prosodic features for speaker identification. ICSLP, Philadelphia.

Cristianini, N. and J. Shawe-Taylor (2000). An introduction to support vector machines (and other kernel-based learning methods), Cambridge University Press.

Godfrey, J. J., E. C. Holliman, et al. (1990). Switchboard: Telephone speech corpus for research and development. ICASSP.

Hearst, M. A. (1998). "Trends and Controversies: Support Vector Machines." IEEE Intelligent Systems **13**: 18-28.

Indovina, M., U. Uludag, et al. (2003). Multimodal Biometric Authentication Methods: A COTS Approach. MMUA, Workshop on Multimodal User Authentication, Santa Barbara, CA.

Lee, D. D. and H. S. Seung (2001). Algorithms for non-negative matrix factorization. Advances in Neural Information Processing Systems: Proceedings of the 2000 Conference, MIT Press.

Luque, J., R. Morros, et al. (2006). Audio, video and multimodal person identification in a smart room. CLEAR 06 Workshop, Southampton.

Lüttin, J., G. Maître, et al. (1998). Evaluation Protocol for the Extended M2VTS Database (XM2VTSDB). Martigny, Switzerland, IDIAP.

Nadeu, C., J. B. Mariño, et al. (1996). Frequency and time-filtering of filter-bank energies for HMM speech recognition. ICSLP.

Peskin, B., J. Navratil, et al. (2003). Using prosodic and conversational features for high-performance speaker recognition: Report from JHU WS'02. ICASSP.

Torre, Á. d. l., A. M. Peinado, et al. (2005). "Histogram Equalization of Speech Representation for Robust Speech Recognition." IEEE Transactions on Speech and Audio Processing **13**(3): 355-366.

Zafeiriou, S., A. Tefas, et al. (2005). Discriminant NMF-faces for frontal face verification. IEEE International Workshop on Machine Learning for Signal Processing, Mystic, Connecticut.