On the Fusion of Prosody, Voice Spectrum and Face Features for Multimodal Person Verification

M. Farrús, A. Garde, P. Ejarque, J. Luque, J. Hernando

TALP Research Center
Department of Signal Theory and Communications
Technical University of Catalonia, Barcelona, Spain
{mfarrus, agarde, pascual, aluque, javier}@gps.tsc.upc.edu

Abstract[1]

Multimodal person recognition systems normally use short-term spectral features as voice information. In this paper prosodic information is added to a system based on face and voice spectrum features. By using two fusion techniques, support vector machines and matcher weighting, different fusion strategies based on the fusion of monomodal scores in several steps are proposed. The performance of the system is clearly improved when the prosodic information is added and the best results are achieved when prosodic scores are previously fused and the resulting scores are fused again with spectral and facial scores. Speech and face scores have been obtained upon Switchboard-I and XM2VTS databases respectively.

Index Terms: speaker recognition, multimodality, fusion, prosody, voice spectrum, face

1. Introduction

Multimodal person recognition involves the combination of two or more human traits like voice, face, fingerprints, iris, hand geometry, etc. to achieve better results than using monomodal recognition [1]. In a multimodal biometric system that uses several biometric characteristics fusion is possible at three different levels: feature extraction level, matching score level or decision level. Fusion at the feature extraction level combines different biometric features in the recognition process, while decision level fusion performs logical operations upon the monomodal system decisions to reach a final resolution. Score level fusion matches the individual scores of different recognition systems to obtain a multimodal score. 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 [2]. Since monomodal scores are usually non-homogeneous, the normalization process transforms the different scores of each monomodal system into a comparable range of values.

After normalization, the converted scores are combined in the fusion process in order to obtain a single multimodal score. In some fusion methods each biometric is weighted by a different factor, as in matcher weighting, where each monomodal score is weighted by a factor proportional to the recognition result of the biometric, or in user weighting, where different weighting factors are applied for every user [2].

One of the most currently used fusion techniques in recognition systems 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 [3].

The aim of this work is to add prosodic information to the multimodal biometric recognition systems. Prosodic, vocal tract spectral and facial scores are fused by using two types of fusion: the conventional technique matcher weighting, previously normalized by z-score method, and support vector machines. In order to do it, a new strategy is proposed: score level fusion is carried out in one, two or three steps, considering two different configurations 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, the matcher weighting fusion technique and support vector machines are reviewed in section 3. Finally, experimental results are shown in section 4 for the fusion combinations of prosodic, vocal tract spectrum and face scores obtained upon Switchboard-I and XM2VTS databases. It can be clearly seen that the use of prosodic information improves the performance of voice spectrum and face based systems.

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

[1] 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).

mel-cepstrum coefficients come from a set of mel-scaled log filter-bank energies (LFBE) *S(k), k=1,...,Q*. The sequence of cepstral coefficients is a 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, only its length. In order to overcome them, [4] presents an alternative that consists of a simple linear processing in 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. In most of the experiments that have been done, FF gives comparable or better results than mel-cepstrum coefficients [5].

2.1.2. Prosodic parameters

Lexicon, prosody and phonetics are linguistic levels of information commonly used by humans to recognize others with voice. Prosodic parameters are known as suprasegmental parameters since the segments affected (syllables, words and phrases) are larger than phonetic units. These features are mainly manifested as sound duration, 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 [6]; spectral patterns can be affected by frequency characteristics of the transmission channel, the speech level and the distance between the speaker and the microphone, while fundamental frequency is unaffected by such variations [7].

The prosodic recognition system used in this task was constituted by a total of 9 prosodic features already used in [8]; 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 [9], 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 discriminative information in NMF is to try to discover discriminative 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 [10], in which the final basis images are closer to facial parts.

3. Fusion techniques

In this section the fusion techniques used in this work, matcher weighting and SVM, are reviewed. As it was said in section 1, scores must be normalized before being fused. 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

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

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

In matcher weighting (MW) method [11], 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 weigthing factor and the EER for the mth biometric x^m and M the number of biometrics, the fused score u is expressed as

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

A support vector machine (SVM) is a binary classifier based on a learning fusion technique [12]. 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.

4. Recognition experiments

Prosody, vocal tract spectrum and face based recognition systems used in the fusion experiments are presented in section 4.1. Experimental results obtained by using SVM and matcher weighting fusion methods in different fusion strategies are shown in section 4.2.

4.1 Experimental setup

For the fusion experiments a chimerical database has been created by combining the Switchboard-I speech database [13] and the video and speech XM2VTS database [14] 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 spectral and prosodic based speaker 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 a 32-component GMM-UBM system using short-term feature vectors consisting of 20 Frequency Filtering parameters with a frame size of 30ms and a shift of 10ms. 20 corresponding delta and acceleration coefficients were included. The UBM was trained with 116 conversation sides.

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 1 conversation-side, computing the distance between the test feature vector and the k feature vectors of the claimed speaker, using the k-Nearest Neighbor 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 [14] was followed. The well-known Fisher discriminant criterion was constructed as [15] in order to discover discriminant linear projections and to obtain the facial scores.

In the fusion experiments, the scores obtained from the speech recognition experiments have been combined with the scores obtained from the face recognition experiments. The chimerical database, which contains 30661 users, was created by combining 179 users of the Switchboard-I database and 270 users of the XM2VTS database. Due to the great number of needed experiments for a statistically adequate number of errors, it was necessary to relate one user from one database to more than one user from the other database. Only client experiments were combined to obtain multimodal client experiments and, in the same way, only impostor experiments were combined to obtain multimodal impostor experiments. A total of 46 500 experiments (16 800 client trials and 29 700 impostor trials) have been carried out.

4.2 Verification results

Table 1 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. 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 2.

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

Table 1. EER for each prosodic feature

source	EER (%)				
		SVM		ZS-MW	
prosody		14.65	6.84	15.66	7.44
voice spectrum	10.10	0.99		1.83	
face	2.06				

Table 2. EER for monomodal and bimodal systems

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 matcher weighting fusion is slightly worse than the support vector machines.

4.2.1. One-step fusion

One-step fusion (Figure 1) 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 matcher weighting with z-score normalization) are shown in Table 3.

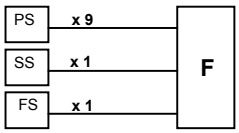

Figure 1. One-step fusion

F	EER (%)
SVM	0.840
ZS-MW	1.320

Table 3. EER for one-step fusion

The results show, once again, that SVM technique outperforms the conventional matcher weighting method with z-score normalization. Furthermore, by using prosodic features the results of the bimodal spectrum and face recognition system are clearly improved.

4.2.2. 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 2). In the first configuration (configuration 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 (configuration B) the scores of the 9 prosodic features are previously fused and the obtained results are then fused with voice spectral and facial scores.

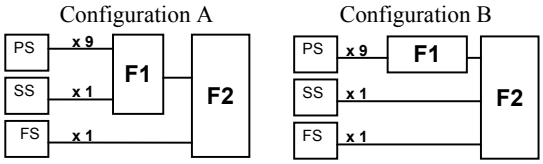

Figure 2. Two configurations of two-step fusion

Table 4 shows the EER for both configurations of the proposed two-step fusion. It can be seen that SVM outperforms, once again, the conventional z-score technique.

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

Table 4. EER (%) for two-step fusion

4.2.3. Three-step fusion

Since the previous results 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 3. EER for three-step SVM fusion are shown in Table 5.

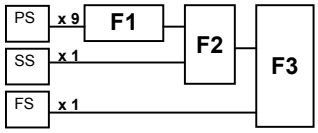

Figure 3. Three-step fusion

F1, F2, F3	EER (%)
SVM	0.868

Table 5. EER for three-step fusion

5. 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 support vector machines outperforms the results obtained by fusing with the matcher weighting technique. The way how the scores are fused is relevant for the performance of the system. The best results are obtained when the three information levels (prosody, voice spectrum and face) are fused at once in the last fusion step. On the other hand, even when the three information levels are not fused in the last step, the results are better if prosodic scores are previously fused. It has been observed that a previous fusion of the voice information (spectral and prosodic scores) does not contribute to the improvement of the system.

6. Acknowledgements

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

7. References

[1] R. M. Bolle, J. H. Connell, S. Pankanti, N. K. Ratha, and A. W. Senior, *Guide to Biometrics*. New York: Springer, 2004.

[2] M. Indovina, U. Uludag, R. Snelik, A. Mink, and A. Jain, "Multimodal Biometric Authentication Methods: A COTS Approach," MMUA, Workshop on Multimodal User Authentication, Santa Barbara, CA, 2003.

[3] M. A. Hearst, "Trends and Controversies: Support Vector Machines," *IEEE Intelligent Systems*, vol. 13, pp. 18-28, 1998.

[4] C. Nadeu, J. B. Mariño, J. Hernando, and A. Nogueiras, "Frequency and time-filtering of filter-bank energies for HMM speech recognition," ICSLP, 1996.

[5] A. Abad, C. Nadeu, J. Hernando, and J. Padrell, "Jacobian Adaptation based on the Frequency-Filtered Spectral Energies," Eurospeech, Geneva, Switzerland, 2003.

[6] M. J. Carey, E. S. Parris, H. Lloyd-Thomas, and S. Bennett, "Robust prosodic features for speaker identification," ICSLP, Philadelphia, 1996.

[7] B. S. Atal, "Automatic speaker recognition based on pitch contours," *Journal of the Acoustical Society of America*, vol. 52, pp. 1687-1697, 1972.

[8] B. Peskin, J. Navratil, J. Abramson, D. Jones, D. Klusacek, D. A. Reynolds, and B. Xiang, "Using prosodic and conversational features for high-performance speaker recognition: Report from JHU WS'02," ICASSP, 2003.

[9] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," Advances in Neural Information Processing Systems: Proceedings of the 2000 Conference, 2001.

[10] S. Zafeiriou, A. Tefas, and I. Pitas, "Discriminant NMF-faces for frontal face verification," IEEE International Workshop on Machine Learning for Signal Processing, Mystic, Connecticut, 2005.

[11] R. Snelick, U. Uludag, A. Mink, M. Indovina, and A. Jain, "Large-Scale Evaluation of Multimodal Biometric Authentication Using State-of-the Art Systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 450-455, 2005.

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

[13] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," ICASSP, 1990.

[14] J. Lüttin, G. Maître, and -. "Evaluation Protocol for the Extended M2VTS Database (XM2VTSDB)," IDIAP, Martigny, Switzerland, IDIAP Communication 05, 1998.

[15] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 711-720, 1997.