Histogram Equalization in SVM Multimodal Person Verification

Mireia Farrús, Pascual Ejarque, Andrey Temko, Javier Hernando

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

Abstract. It has been shown that prosody helps to improve voice spectrum based speaker recognition systems. Therefore, prosodic features can also be used in multimodal person verification in order to achieve better results. In this paper, a multimodal recognition system based on facial and vocal tract spectral features is improved by adding prosodic information. Matcher weighting method and support vector machines have been used as fusion techniques, and histogram equalization has been applied before SVM fusion as a normalization technique. The results show that the performance of a SVM multimodal verification system can be improved by using histogram equalization, especially when the equalization is applied to those scores giving the highest EER values.

Keywords: speaker recognition, multimodality, fusion, support vector machines, histogram equalization, prosody, voice spectrum, face.

1 Introduction

Multimodal biometric systems, which involve the combination of two or more human traits, are used to achieve better results than the ones obtained in a monomodal recognition system [1]. In a multimodal recognition system, fusion is possible at three different levels: feature extraction level, matching score level and decision level. Fusion at the score level matches the monomodal scores of different recognition systems in order to obtain a single multimodal score, and it is the preferred method by most of the systems.

Matching score level fusion is a two-step process which consists of a previous score normalization and the fusion itself [2-5]. The normalization process transforms the non homogeneous monomodal scores into a comparable range of values. Z-score is a conventional affine normalization technique which transforms the scores into a distribution with zero mean and unitary variance [3, 5]. Histogram equalization (HE) is used as a non linear normalization technique which makes equal the statistics of the monomodal scores. HE can be seen as an extension to the whole statistics of the mean and variance equalization performed by the z-score normalization.

The fusion process is a combination of the previously normalized scores. In this paper, two fusion methods are used and compared: matcher weighting and support vector machines. In matcher weighting method, each monomodal score is weighted by a factor proportional to its recognition result. A support vector machine is a binary classifier based on a learning fusion technique, where scores are seen as input patterns to be labeled as accepted or rejected.

The aim of this work is to improve the results obtained in our recent work based on the fusion of prosody, voice spectrum and face features where different step strategies were applied [6]. The improvement is achieved with previous histogram equalization as a normalization of the scores in a SVM based fusion.

In the next section, the monomodal information sources used in this work are described. Z-score and histogram equalization are presented in section 3. Matcher weighting fusion technique and support vector machines are reviewed in section 4 and, finally, experimental results are shown in section 5.

2 Monomodal Sources

2.1 Voice Information

In multimodal person recognition only short-term spectral features are normally used as voice information. However, it has been demonstrated that voice spectrum based systems can be improved by adding prosodic information [7].

Spectral parameters are those which only take into account the acoustical 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 to represent the short-time spectral envelope of a speech frame in current speaker recognition systems. However, Frequency Filtering (FF) parameters, presented in [8] and used in this work, become an alternative to the use of cepstrum in order to overcome some of its disadvantages.

Several linguistic levels like lexicon, prosody or phonetics are used by humans to recognize others with voice. These levels of information are more related to learned habits and style, and they are mainly manifested in the dialect, sociolect or idiolect of the speaker. Prosodic parameters, in particular, are manifested as sound duration, tone and intensity variation. Although these features don’t provide very good results when they are used alone, they give complementary information and improve the results when they are fused with vocal tract spectrum based systems. The prosodic recognition system used in this task consists of a total of 9 prosodic scores already used in [9]:

- number of frames per word averaged over all words
- average length of word-internal voiced segments
- average length of word-internal unvoiced segments
- mean F0 logarithm
- maximum F0 logarithm
- minimum F0 logarithm
- F0 range (maximum F0 – minimum F0) logarithm
- F0 “pseudo slope”: (last F0 – first F0) / (number of frames in word)
- average slope over all segments of a piecewise linear stylization of F0

2.1 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 [10], 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 [11], in which the final basis images are closer to facial parts.

3 Histogram Equalization

Z-score (ZS) is one of the most conventional normalization methods, which transforms the scores into a distribution with zero mean and unitary variance Denoting as *a* the raw matching from the set *A* of all the original monomodal biometric scores, the z-score normalized biometric is computed as:

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

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

Histogram equalization (HE) is a general non parametric method to match the cumulative distribution function (CDF) of some given data to a reference distribution. This technique can be seen as an extension of the statistical normalization made by the z-score to whole biometric statistics.

Histogram equalization 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.

This method has been also developed for the speech recognition adaptation approaches and the correction of non linear effects typically introduced by speech systems such as microphones, amplifiers, clipping and boosting circuits and automatic gain control circuits [12, 13].

The objective of HE is to find a non linear transformation to reduce the mismatch of the statistics of two signals. In [14, 15] this concept was applied to the acoustic features to improve the robustness of a speaker verification system. On the other hand, in this paper HE is applied to the scores. *N* intervals with the same probability are assigned in the distributions of both signals. Each interval in the reference distribution, $x \in [q_i, q_{i+1}[$, is represented by *(x$_i$, F(x$_i$))*. x_i is the average of the scores and $F(x_i)$ is the maximum cumulative distribution value:

$$x_i = \frac{\sum_{j=1}^{k_i} x_{ij}}{k_i} \quad , \qquad\qquad F(x_i) = \frac{K_i}{M} \tag{2}$$

where x_{ij} are the scores in the interval, k_i is the number of scores in the interval, K_i is the number of data in the interval $[q_0, q_{i+1}[$, and M is the total amount of data.

All the scores in each interval of the source distributions are assigned to the corresponding interval in the reference distribution. $F(x_i)$ sets the boundaries $[q'_i, q'_{i+1}[$ of the intervals in the distribution to be equalized. These boundaries limit the interval of values that fulfils the following condition: $F(q_i) \le F(y) < F(q_{i+1})$, and all the values of the source signal lying in the interval $[q'_i, q'_{i+1}[$ will be transformed to their corresponding x_i value.

4 Fusion Techniques and Support Vector Machines

One of the most conventional fusion techniques is the matcher weighting (MW) method, where each monomodal score is weighted by a factor proportional to each biometric 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 m-th biometric x_m and M the number of biometrics, the final fused score u is expressed as [1, 3]:

$$u = \sum_{m=1}^{M} w_m x_m \qquad \text{where} \qquad w_m = \frac{\frac{1}{e_m}}{\sum_{m=1}^{M} \frac{1}{e_m}} \ . \tag{3}$$

In contrast to the MW that is a linear and a data-driven fusion method, non linear and machine learning based methods may lead to a higher performance. 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 labeled as ‘accepted’ or ‘rejected’.

Recent works on statistical machine learning have shown the advantages of discriminative classifiers like SVM [16] in a range of applications. Support vector machine (SVM) is a state-of-the-art binary classifier. Given a linearly separable two-class training data, SVM finds an optimal hyperplane that splits input data in two classes, maximizing the distance of the hyperplane to the nearest data points of each class.

However, data are normally not linearly separable. In this case, non linear decision functions are needed, and an extension to non linear boundaries is achieved by using specific functions called kernel functions [17]. Kernel functions map the data of the input space to a higher dimensional space (feature space) by a non linear transformation. The optimal hyperplane for a non linearly separable data is defined by:

$$f(x) = \sum_{i=1}^{N} \alpha_i t_i \mathrm{K}(x, x_i) + b \qquad \mathbf{(4)}$$

where t_i are labels, K is a chosen kernel function and $\sum_{i=1}^{N} \alpha_i t_i = 0$. The vectors x_i are the support vectors, which determine the optimal separating hyperplane and correspond to the points of each class that are the closest to the separating hyperplane.

5 Recognition Experiments

In the next section, the monomodal recognition systems used in the fusion experiments are described. Experimental results by using different normalization and fusion techniques are shown in section 5.2.

5.1 Experimental Setup

Recognition experiments have been performed with the Switchboard-I speech database [18] and the video and speech XM2VTS database of the University of Surrey [19]. Switchboard-I database, which is a collection of 2430 two-sided telephone conversations among 543 speakers from all areas of the United States, has been used for the speaker recognition experiments. Speaker scores have been obtained by using two different systems: a voice spectrum based recognition system and a prosody based recognition system.

The spectrum based speaker recognition system used is a 32-component GMM system with diagonal covariance matrices; 20 Frequency Filtering parameters were generated with a frame size of 30 ms and a shift of 10 ms, and 20 corresponding delta and acceleration coefficients were included.

In the prosody based recognition system a 9-feature vector was extracted for each conversation side. The mean and standard deviation over all words were computed for each individual feature. The system was tested using the k-Nearest Neighbor classifier (with k=3), comparing the distance of the test feature vector to the k closest vectors of the claimed speakers and the distance of the test vector to the k closest vectors of the cohort speakers, and using the symmetrized Kullback-Leibler divergence as a distance measure.

In both spectral and prosodic systems, each speaker model was trained with 8 conversation sides. Training was performed using splits 1-3 of the Switchboard-I database, and splits 4-6 were provided the cohort speakers and the UBM. Both systems were tested with one conversation-side according to NIST's 2001 Extended Data task.

Face recognition experiments were performed with the XM2VTS database, which is a multimodal database consisting of face images, video sequences and speech recordings of 295 subjects. Only the face images have been used in our experiments. In order to evaluate the verification algorithms on the database, the evaluation protocol described in [19] was followed. The well-known Fisher discriminant criterion was constructed as [20] in order to discover discriminant linear projections and to obtain the facial scores.

Fusion experiments have been done at the matching score level. Since both databases contain biometric characteristics belonging to different users, a chimerical database has been created to perform the experiments. A chimerical database is an artificial database created using two or more monomodal biometric characteristics from different individuals to form artificial (or chimerical) users. In this paper, the chimerical database consists of 30661 users created by combining 179 different voices of the Switchboard-I database with 270 different faces of the XM2VTS database. The scores were then split in two equal sets (development and test) for each recognition system, obtaining a total amount of 46500 scores for each set (16800 clients and 29700 impostors).

The kernel function used in the SVM was a Gaussian radial basis function. Scores are always equalized to the histogram corresponding to the best scores involved in the fusion; i.e. those scores that provided the lowest EER. A 1000-interval histogram was applied before each SVM fusion, and both SVM and HE-SVM techniques are compared to a baseline system which uses MW fusion and z-score normalization.

5.2 Verification Results

Monomodal systems. Table 1 shows the EER obtained for each prosodic feature in the prosody based recognition system. As it can be seen in the table, features based on fundamental frequency measurements achieve the lowest EER.

Table 1. EER for each prosodic feature

Features	EER (%)
Log (#frames/word)	30.3
Average length of word-internal voiced segments	31.5
Average 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
F0 'pseudo-slope'	38.3
Average slope over all segments of PWL stylization of F0	28.7

The EER obtained in each monomodal recognition system are shown in Table 2. Note that fusion is only used in the prosodic system, where there are 9 prosodic scores to be combined. In this case, fusion is carried out in one single step, and the results of the three types of fusion mentioned above are presented: (1) matching score fusion with z-score normalization, (2) support vector machines and (3) support vector machines with a previous histogram equalization.

Table 2. EER (%) for each monomodal recognition system

Source		EER (%)
Prosody	ZS-MW	15.66
	SVM	14.65
	HE-SVM	13.39
Voice spectrum		10.10
Face		2.06

Bimodal systems. Table 3 shows the fusion results for two bimodal systems: a prosody based system fused with the voice spectral recognition system, and a voice spectrum based system fused with the face recognition system, using the same fusion methods as above. As in the monomodal systems (Table 2), matcher weighting fusion is slightly worse than the support vector machines.

Table 3. EER (%) for each bimodal recognition system

Source	ZS-MW	SVM	HE-SVM
Prosody + voice spectrum	7.44	6.84	6.25
Voice spectrum + face	1.83	0.99	1.02

Trimodal system. In [6] several strategies were proposed by fusing the monomodal scores in one, two and three steps. In those experiments the best results were achieved in a two-step configuration, where the 9 prosodic scores were fused in the first step and the obtained scores were then fused in the second step with voice spectral and facial scores (Fig. 1).

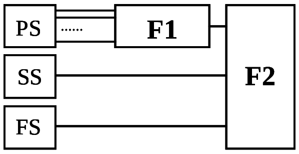

Fig. 1. Two-step fusion

The EER for the selected two-step fusion are presented in Table 4. Once again, matcher weighting fusion method is clearly outperformed by support vector machines.

Table 4. EER (%) for the trimodal system

Fusion technique	EER (%)
ZS-MW	1.493
SVM	0.647

In our trimodal system, equalization has been applied in all the possible combinations:

(1) HE before the first fusion (equalization of the prosodic scores)
(2) HE before the second fusion (equalization of the three modalities)
(3) HE before both fusions F1and F2

The results are shown in Table 5 and they are compared to the non equalized SVM fusion.

Table 5. EER (%) applying HE before SVM fusion

Fusion technique	F1	F2	EER (%)
SVM	-	-	0.647
	HE	-	0.630
	-	HE	0.649
	HE	HE	**0.613**

As it can be seen in the table, the best result is achieved when histogram equalization is used before F1 and F2. By equalizing only the prosodic scores, the performance of the system is also improved. On the other hand, equalization before the second fusion does not improve the performance of the system.

6 Conclusions

In this work, the use of prosody improves the performance of a bimodal system based on vocal tract spectrum and face features. The experiments show that support vector machines based fusion is clearly better than matcher weighting fusion method. In addition, results are improved by applying histogram equalization as a normalization technique before SVM fusion. The best verification results are achieved when the histogram of the scores with the highest values of EER (the prosodic scores in our experiments) are equalized to the distribution of the scores that provide the lowest EER.

References

1. Bolle, R.M., et al., Guide to Biometrics. 2004, New York: Springer. 364.

2. Fox, N.A., et al. Person identification using automatic integration of speech, lip and face experts. ACM SIGMM 2003 Multimedia Biometrics Methods and Applications Workshop. Berkeley, CA (2003).
3. Indovina, M., et al. Multimodal Biometric Authentication Methods: A COTS Approach. MMUA. Workshop on Multimodal User Authentication. Santa Barbara, CA (2003).
4. Lucey, S. and T. Chen. Improved audio-visual speaker recognition via the use of a hybrid combination strategy. The 4th International Conference on Audio- and Video- Based Biometric Person Authentication. Guildford, UK (2003).
5. Wang, Y., Y. Wang, and T. Tan. Combining fingerprint and voiceprint biometrics for identity verification: and experimental comparison. ICBA 2004. Hong Kong, China (2004).
6. Farrús, M., et al. On the Fusion of Prosody, Voice Spectrum and Face Features for Multimodal Person Verification. ICSLP. Pittsburgh (2006).
7. Campbell, J.P., D.A. Reynolds, and R.B. Dunn. Fusing high- and low-level features for speaker recognition. Eurospeech (2003).
8. Nadeu, C., J. Hernando, and M. Gorricho. On the decorrelation of filter bank energies in speech recognition. Eurospeech (1995).
9. Peskin, B., et al. Using prosodic and conversational features for high-performance speaker recognition: Report from JHU WS'02. ICASSP (2003).
10. Lee, D.D. and H.S. Seung. Algorithms for non-negative matrix factorization. in Advances in Neural Information Processing Systems: Proceedings of the 2000 Conference. MIT Press (2001).
11. Zafeiriou, S., A. Tefas, and I. Pitas. Discriminant NMF-faces for frontal face verification. IEEE International Workshop on Machine Learning for Signal Processing. Mystic, Connecticut (2005).
12. Hilger, F. and H. Ney. Quantile based histogram equalization for noise robust speech recognition. Eurospeech. Aalborg, Denmark (2001).
13. Balchandran, R. and R. Mammone. Non parametric estimation and correction of non linear distortion in speech systems. ICASSP (1998).
14. Pelecanos, J. and S. Sridharan. Feature warping for robust speaker verification. in ODYSSEY-2001 (2001).
15. Skosan, M. and D. Mashao, Modified Segmental Histogram Equalization for robust speaker verification. Pattern Recognition Letters **27**:5 (2006) 479-486.
16. Cristianini, N. and J. Shawe-Taylor, An introduction to support vector machines (and other kernel-based learning methods). Cambridge University Press (2000).
17. Burges, C.J.C., A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge discovery **2** (1998) 121-167.
18. Godfrey, J.J., E.C. Holliman, and J. McDaniel. Switchboard: Telephone speech corpus for research and development. ICASSP (1990).
19. Lüttin, J. and G. Maître, Evaluation Protocol for the Extended M2VTS Database (XM2VTSDB). IDIAP: Martigny, Switzerland (1998).
20. Belhumeur, P.N., J.P. Hespanha, and D.J. Kriegman, Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. IEEE Transactions on Pattern Analysis and Machine Intelligence **19**:7 (1997) 711-720.