interspeech 2008 incorporating SST 08

Bi-Gaussian Score Equalization in an Audio-Visual SVM-based Person Verification System

Pascual Ejarque, Javier Hernando

TALP Research Center
Department of Signal Theory and Communications
Technical University of Catalonia, Barcelona, Spain
pascual@gps.tsc.upc.edu, javier.hernando@upc.edu

Abstract

In multimodal fusion systems a normalization of the features or the scores is needed before the fusion process. In this work, in addition to the conventional methods, histogram equalization, which was recently introduced by the authors in multimodal systems, and Bi-Gaussian equalization, which takes into account the separate statistics of the genuine and impostor scores, and is introduced in this paper, are applied upon the scores in a multimodal SVM-based person verification system composed by prosodic, speech spectrum, and face information. Bi-Gaussian equalization has obtained the best results and outperform in more than a 23.25% the results obtained by Min-Max normalization.

Index Terms: equalization, normalization, Support Vector Machines, multimodal, audio-visual.

1. Introduction

Multimodal score fusion can be performed in two main approaches: the arithmetical or logical combination of the scores and the classification of the score vectors by mean of classificatory techniques [1]. In the combinatorial approach the scores provided by every unimodal system must be normalized before the fusion process due to, without this process, the contribution of a biometric could eliminate the contribution of the rest of the experts [2]. In the classificatory approach, not much importance has been given to score normalization because the same classificatory techniques can adapt themselves to the biometric characteristics.

Concretely, for the SVM based classificatory techniques, the usage of kernels permits the non linear transformation of the input scores in a higher dimensional subspace where the recognition decision can be taken by means of a separator hyperplane [3]. Some efforts have been made for the development of particular kernels for each application, as in the case of spherical normalization developed by Wan et al. [4]. However, most investigators and developers use well-known kernels as radial basis function (RBF) or polynomial kernels for their systems and adapt them by the modification of the kernel parameters. In this case, the number of non linear transformations is limited by the kernel and the chosen parameters.

The aim of this work is to demonstrate the importance of the normalization of the unimodal scores in an SVM fusion system and, more concretely, the application of histogram equalization techniques in the normalization process. Histogram equalization consists in the equalization of the distribution function to a reference signal, and has been used in a wide range of applications including image and speech processing. The authors have recently introduced this technique in multimodal systems [5, 6] with good results.

In other hand, most fusion techniques do not take into account the separate statistics of the genuine and impostor scores. The authors made an effort in 2005 [7] to introduce this information in multimodal fusion techniques with good results. In this work, histogram equalization has been performed upon a reference distribution where the genuine and impostor distributions have been separately generated. The genuine and impostor distributions have been built by means of two Gaussians which variances have been set in order to the reference distribution had the same EER than the original modality. This novel technique has been called Bi-Gaussian equalization and outperforms the results obtained by the rest of tested normalizations.

The multimodal system is composed by three score sources: the first score has obtained by the SVM fusion of 9 voice prosodic features [5, 8], the second one has been obtained by a voice spectrum expert based in the Frequency Filtering front-end and GMM [9], and the last one has been provided by an NMFFaces algorithm face recognition system [10]. The prosodic and spectrum scores have been obtained from voice signals of the Switchboard-I database and the face scores have been obtained from face still images of the XM2VTS database.

The paper is organized as follows: in section 2, the normalization techniques that have been tested in this work are introduced; in section 3 the equalization methods are presented; and finally; in sections 4 and 5, the results and conclusions are presented.

2. Normalization Methods

The normalization process transforms the unimodal scores of all the biometrics in a comparable range of values and is an essential step in multimodal fusion. The most conventional normalization techniques are Min-Max, Z-Score, and Tanh, which have been widely used in previous works [1, 2].

Min-Max normalization maps the scores in the [0, 1] range by means of an affine transformation, i.e.,

$$x_{MM} = \frac{a - \min(a)}{\max(a) - \min(a)} \quad (1)$$

where *min(a)* and *max(a)* are the minimum and maximum values of the unimodal scores *a*.

By means of Z-Score normalization the mean of all the biometric scores is set to 0 and its variance is set to 1 in a non affine transformation. Equation 2 demonstrates the application of this normalization

$$x_{ZS} = \frac{a - mean(a)}{std(a)} \tag{2}$$

where *mean(a)* and *std(a)* are respectively the statistical mean and standard deviation of a unimodal set of scores.

Tanh normalization maps the scores in the [-1, 1] range in a non linear transformation. This normalization is performed by means of the formula in equation 3

$$x_{TANH} = \frac{1}{2}\left\{\tanh\left(k\frac{a-\mu_{GH}}{\sigma_{GH}}\right)+1\right\} \tag{3}$$

where μ_{GH} and σ_{GH} are, respectively, the mean and standard deviation estimates, of the genuine score distribution introduced by Hampel [2], and k is a suitable constant. The main advantage of this normalization is the diminution of the effect of outliers, which is absorbed by the compression of the extreme values.

3. Equalization

The normalization techniques in the previous section permit to control certain statistical characteristics of the normalized scores, as the value of the maximum and the minimum score in the case of the MM and TANH techniques, or the mean and the variance of the scores in the case of the ZS normalization.

The equalization process transforms the source histogram to a known reference distribution and, for this reason, the mean of the scores and the shape of the whole histogram can be set in order to improve the recognition results. In this section, histogram equalization, which has been recently integrated in multimodal person recognition systems by the authors [5, 6], is reviewed. Furthermore, a novel equalization technique that takes into account the separate statistics of the genuine and impostor scores, Bi-Gaussian equalization, is introduced in this paper. These techniques have been used as normalization methods in this work.

3.1. Histogram equalization (HEQ)

By means of histogram equalization, the distribution function of the unimodal biometrics is equalized to a distribution of reference. This non-linear technique has been widely used in image processing [11] and has been applied to speech processing in order to reduce non linear effects introduced by speech systems such as: microphones, amplifiers, etc. [12]. Furthermore, it has also been used in robust speaker verification for the warping of cepstral features streams over a specified interval [13].

The first phase of the equalization process is the division of the cumulative histogram of the scores in M intervals; all those with the same probability of occurrence ($1/M$). Once the source intervals have been defined, the target distribution must also be divided in the same number of intervals, and the point with the "half probability" is selected as the representative point for each interval. In the equalization process, all the scores of the source biometric that are included in an interval are matched to the representative point in the corresponding interval of the reference histogram.

This equalization process can be seen as an increasing transformation from the source to the equalized scores. For this reason, this transformation will not modify the recognition result of the biometric and can be considered a normalization process. Figure 1 illustrates the equalization process. The score x and all the scores in the same interval are matched to the equalized score y in the reference distribution.

The authors have used histogram equalization as a normalization technique in [5, 6] using one of the unimodal histograms as the reference distribution. The good results obtained in these works with combinatorial techniques as simple sum or matcher weighting for the fusion process are confirmed in this paper in an SVM-based fusion scheme.

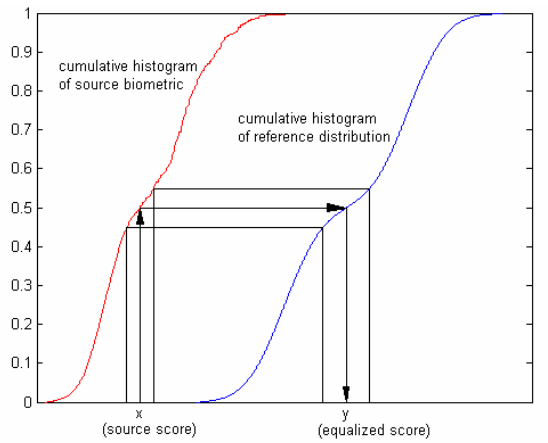

Figure 1: The distribution matching performed by HEQ.

Another technique that can be seen as a particular case of histogram equalization is rank normalization [14], where each feature value is replaced by its rank in the reference distribution to obtain an approximately uniform distribution.

3.2. Bi-Gaussian equalization (BGEQ)

In the normalization process, the separate statistics of the genuine and the impostor scores are not generally taken into account. However, the information relative to these separate statistics can be used to improve the results obtained by multimodal fusion systems [6, 7].

In this section, a new technique is proposed for the equalization of the scores, where the genuine and impostor distributions are separately modelled by means of two independent Gaussian distributions that are overlapped to construct a reference distribution that had the same recognition error than the original biometric. This technique has been called Bi-Gaussian equalization.

The mean of the genuine Gaussian distribution has been set to 1 and the impostor one has been set to -1 and the variance of both Gaussians, which controls the overlapping among them, has been set to the adequate value to accomplish the error condition. The reference distribution is, in this case,

$$f_{ref}(x) = \frac{1}{2\sigma\sqrt{2\pi}}\left[e^{-\frac{(x+1)^2}{2\sigma^2}} + e^{-\frac{(x-1)^2}{2\sigma^2}}\right] \tag{8}$$

where σ is the standard deviation of both Gaussians.

In this work, the Equal Error Rate (EER) of the original biometrics has been used to determine the variance of the Bi-Gaussian Equalization. Due the fact that both genuine and impostor Gaussian distributions have been designed with the same variance, the EER will be obtained by setting the threshold to the average of the means of both Gaussians, in this case, zero. As the EER can be calculated as the probability that the value of an impostor score is greater or a genuine score is less than the threshold, i.e.

$$EER = \int_0^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x+1)^2}{2\sigma^2}} dx = \int_{-\infty}^{0} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-1)^2}{2\sigma^2}} dx \quad (9)$$

and with a change of variable in the previous equation, the variance of the Gaussians for each biometric can be calculated from its EER and the areas of a normal distribution.

Figure 2 shows the Bi-Gaussian equalized biometric histograms for the face, spectrum speech, and prosody scores. As it can be observed in the figure, one of the benefits of the application of Bi-Gaussian equalization is the attenuation of the effect of the outliers.

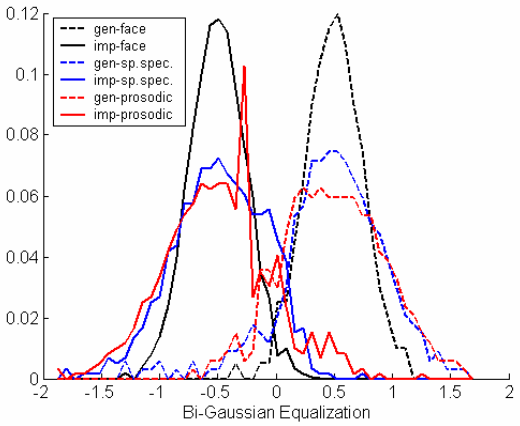

Figure 2: Histogram of the scores for BGEQ.

4. Recognition experiments

In this section, the unimodal recognition systems that provide the scores for the training and the testing of the fusion system are presented. Furthermore, the techniques and databases used in the fusion process are detailed.

The experimental results obtained with the different normalization methods in an SVM fusion scheme will be presented in the second subsection.

4.1. Experimental setup

The unimodal scores used in the experiments have been provided by three experts: an SVM fusion of 9 speech prosodic features, a voice spectrum based speaker recognition system and a facial recognition expert based in the NMFFaces algorithm.

In the prosody based recognition system a 9 prosodic feature vector was extracted for each conversation side [8]. The system was tested with 1 conversation-side, using the k-Nearest Neighbor method. The prosodic vectors have been fused by means of a SVM classificatory system to obtain a single unimodal score.

The spectrum based speaker recognition system was a 32-component GMM system with diagonal covariance matrices; 20 Frequency Filtering parameters were generated [9], and 20 corresponding delta and acceleration coefficients were included. The UBM was trained with 116 conversations.

The face recognition expert is based in the NMFFaces algorithm [10], where non-negative matrix factorization is used to yield sparse representation of localized features to represent the constituent facial parts over the face images.

Equal error rate (EER) and minimum half total error rate (HTER) have been used in this work as the error measures to compare the recognition systems. EER and HTER are respectively 14.65% and 14.27% for the prosodic system, 9.52% and 8.50% for the speech spectrum system, and 2.50% and 2.00% for the facial recognition system.

The fusion process has been performed, after the normalization or equalization of the scores, by means of a Support Vector Machine system using radial basis function (RBF) and polynomial kernels. Radial basis function kernel is based in Gaussian classificatory regions where the parameter σ controls the variance of the Gaussian functions, i.e.

$$k(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \quad (11)$$

Polynomial kernels are based in the dot product of the data vectors and are controlled by the exponent α, i.e.

$$k(x_i, x_j) = \left(\langle x_i, x_j \rangle + 1\right)^{\alpha} \quad (12)$$

Prosodic and spectrum scores have been obtained form speech records of the Switchboard-I database [15] and the face scores have been obtained from still images of the XM2VTS database [16]. The Switchboard-I is a collection of 2,430 two-sided telephone conversations among 543 speakers from the United States. XM2VTS database contains face images of 295 subjects. A chimerical database has been created by the combination of the 1,860 speech experiments (for prosodic and spectrum information) and 33,361 face scores. A total of 5,000 score vectors have been generated for the training of the fusion models and 46,500 score vectors has been used in the test phase.

4.2. Results

In the experiments, several normalization techniques have been applied upon the unimodal scores. Later, these scores have been fused by means of a SVM system. The normalization methods are those presented in sections 2 and 3: Min-Max (MM), Z-Scores (ZS), a tanh based technique (TANH), rank equalization (RANK), histogram equalization to the best unimodal system, the face recognition system, (HEQ), and Bi-Gaussian equalization (BGEQ).

To compare the effect of each normalization method upon the SVM fusion system, RBF and polynomial kernel SVM configurations have been tested. Concretely, for the RBF kernel different values of the Gaussian variance σ have been tested: 1/3, 1, 3, and 9. For the polynomial kernel, values from 1 to 4 have been used for the α parameter. Furthermore, the regularization parameter C has been set to 1, 10, 100, and 200. A cross-validation process has been performed for the selection of the final parameters for each technique and kernel. The EER and the HTER obtained by each normalization technique and kernel are presented in Table 1.

The equalization techniques obtain the best results for both kernels for the two error measurements. Min-Max, one of the most often used normalization techniques in SVM systems, is outperformed by Bi-Gaussian equalization with relative improvements from 23.25% to 32.91% while ZS is improved in more than a 10% and TANH in more than a 5%. Among the equalization techniques, rank normalization is outperformed with relative improvements from 2.39% to 9.74% and HEQ is also improved by BGEQ with the polynomial kernel.

	RBF		Polynomial	
	EER	**HTER**	**EER**	**HTER**
MM	1.005	0.948	0.869	0.827
ZS	0.852	0.686	1.101	1.029
TANH	0.714	0.671	0.703	0.657
RANK	0.739	0.662	0.697	0.627
HEQ	**0.690**	0.652	0.708	0.647
BGEQ	0.701	**0.636**	**0.667**	**0.612**

Table 1: Multimodal results.

The minimum EER and HTER are obtained by the polynomial kernel with the BGEQ normalization, that obtain a relative improvement of a 69% with respect to the best unimodal system, and that improve between a 3% and a 4% the best results obtained with the RBF kernel. MM, TANH and BGEQ obtain the best results with the polynomial kernel while ZS takes a greater advantage of the use of the RBF kernel.

In Figure 2, the DET curve for the comparison of the normalization methods with the polynomial kernel is shown.

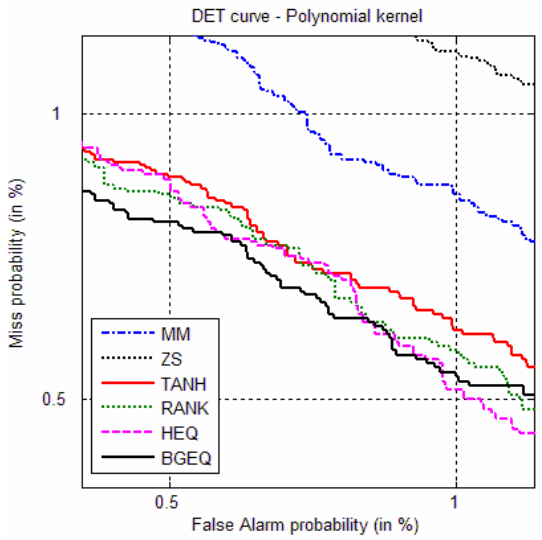

Figure 2: DET curve for polynomial kernel SVM.

For all the ranges of FAR and FRR, the best results are obtained by HEQ and BGEQ. This last technique outperforms the conventional normalizations for all the range of values of FAR and FRR.

5. Conclusions

Support Vector Machines fusion systems need the normalization process for the alignment of the range of values for the features or the scores. In this work, several normalization methods have been applied upon a multimodal score SVM fusion system with RBF and polynomial kernel.

In this work, Bi-Gaussian equalization, a novel equalization technique that takes into account the separate distributions of clients and impostors, obtains the best results in a multimodal SVM-based system for the fusion of the scores of prosody, spectrum speech and face recognition experts. Concretely, Bi-Gaussian equalization obtains relative error improvements from 23.25% to 32.91% with respect to MM normalization and outperforms the conventional normalization techniques for all values of FAR and FRR.

6. Acknowledges

This work has been funded by the Spanish project SAPIRE (TEC2007-65470). We want to thank Dr. A. Tefas who has provided us of face recognition results.

7. References

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

[2] Jain, A. K., Nandakumar, K., and Ross, A., "Score normalization in multimodal bimetric systems", Pattern Recognition, vol. 38, no. 12, pp. 2270-2285, 2005.

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

[4] Wan, V., and Renals, S., "Speaker verification using sequence discriminant support vector machines", IEEE Trans. on Speech and Audio Processing, 13:203-210, 2005.

[5] Farrús, M., Garde, A., Ejarque, P., Luque, J., and Hernando, J., "On the Fusion of Prosody, Voice Spectrum and Face Features for Multimodal Person Verification", Proc. of Interspeech, Pittsburgh.USA, 2006.

[6] Ejarque, P., Garde, A., Anguita, J., and Hernando, J., "On the use of genuine-impostor statistical information for score fusion in multimodal biometrics", Multimodal Biometrics in Annals of Telecommunication 2007.

[7] Ejarque, P., and Hernando, J., "Variance reduction by using separate genuine- impostor statistics in multimodal biometrics", Proc. of Interspeech 2005, pp. 785-788, Lisbon, Portugal, September 2005.

[8] Wolf, J. J., "Efficient acoustic parameters for speaker recognition," Journal of the Acoustical Society of America, vol. 51, pp. 2044-2056, 1972.

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

[10] Tefas, A., Zafeiriou, S., and Pitas, I., "Discriminant NMFfaces for frontal face verification", Proc. of IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2005), Mystic, Connecticut, September 28-30, 2005.

[11] Jain, A., "Fundamentals of Digital Image Processing", Prentice-Hall, 1986, pp 241 - 243.

[12] Balchandran, R., and Mammone, R., "Non parametric estimation and correction of non-linear distortion in speech systems", Proc. IEEE Int. Conf. Acoust. Speech Signal Proc., 1998.

[13] Pelenacos, J., and Sridharan, S., "Feature warping for robust speaker verification", Proc. ISCA Workshop on Speaker Recognition – 2001: A Speaker Oddyssey, June 2001, pp. 213-218.

[14] Stolcke, A., Ferrer, L., Kajarekar, S., Shrigerg, E., and Venkataraman, A., "MLLR Transforms as Features in Speaker Recognition", Proc. Eurospeech, Lisbon, pp. 2425-2428, 2005.

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

[16] Lüttin, J., and Maître, G., "Evaluation Protocol for the Extended M2VTS Database (XM2VTSDB)", IDIAP Communication 98-05 (1998), Martigny, Switzerland.