DOI: 10.21917/ijivp.2011.0043

FUSING SPEECH SIGNAL AND PALMPRINT FEATURES FOR AN SECURED AUTHENTICATION SYSTEM

P.K. Mahesh[1] and M.N. Shanmukha Swamy[2]

Department of Electronics and Communication Engineering, J.S.S. Research Foundation, Sri Jayachamarajendra College of Engineering, Karnataka, India

E-mail: [1]mahesh24pk@gmail.com and [2]mnsjce@gmail.com

Abstract

In the application of Biometric authentication, personal identification is regarded as an effective method for automatic recognition, with a high confidence, a person's identity. Using multimodal biometric systems we typically get better performance compare to single biometric modality. This paper proposes the multimodal biometrics system for identity verification using two traits, i.e., speech signal and palmprint. Integrating the palmprint and speech information increases robustness of person authentication. The proposed system is designed for applications where the training data contains a speech signal and palmprint. It is well known that the performance of person authentication using only speech signal or palmprint is deteriorated by feature changes with time. The final decision is made by fusion at matching score level architecture in which feature vectors are created independently for query measures and are then compared to the enrolment templates, which are stored during database preparation.

Keywords:

Multimodal Biometrics, Speech Signal, Palmprint, Fusion, Matching Score

1. INTRODUCTION

A multimodal biometric authentication, which identifies an individual person using physiological and/or behavioural characteristics, such as face, fingerprints, hand geometry, iris, retina, vein and speech is one of the most attractive and effective methods. These methods are more reliable and capable than knowledge-based (e.g. Password) or token-based (e.g. Key) techniques. Since biometric features are hardly stolen or forgotten.

However, a single biometric feature sometimes fails to be exact enough for verifying the identity of a person. By combining multiple modalities enhanced performance reliability could be achieved. Due to its promising applications as well as the theoretical challenges, multimodal biometric has drawn more and more attention in recent years [1]. Speech Signal and palmprint multimodal biometrics are advantageous due to the use of non-invasive and low-cost speech and image acquisition. In this method we can easily acquire palmprint images using touchless sensors and speech signal using microphone. Existing studies in this approach [2, 3] employ holistic features for palmprint and speech signal representation and results are shown with different techniques of fusion and algorithms.

Multimodal system also provides anti-spooling measures by making it difficult for an intruder to spool multiple biometric traits simultaneously. However, an integration scheme is required to fuse the information presented by the individual modalities.

This paper presents a novel fusion strategy for personal identification using speech signal and palmprint features at the features level fusion Scheme. The proposed paper shows that integration of speech signal and palmprint biometrics can achieve higher performance that may not be possible using a single biometric indicator alone. This paper presents MFCC with different window techniques for speech signal and Haar wavelet for palmprint, which gives better performance and better accuracy for both traits (speech signal & palmprint).

The rest of this paper is organized as fallows. Section 2 presents the system structure, which is used to increase the performance of individual biometric trait; multiple classifiers are combined using matching scores. Section 3 presents feature extraction method used for palmprint and section 4 for speech signal. Section 5, the individual traits are fused at matching score level using weighted sum of score techniques. The experimental results are given in section 6. Finally, Conclusions are given in the last section.

2. SYSTEM STRUCTURE

The multimodal biometric system is developed using two traits i.e. speech signal and palmprint as shown in Fig. 1. For the speech signal and palmprint Recognition, the input image is recognized using Mel Frequency Cepstral Coefficients (MFCC) with different window techniques and Haar wavelet method respectively. When we are using a Haar wavelet, the matching score is calculated using weighted ecludiean distance also when we are using MFCC, Gaussian Mixture Model (GMM) is used. The modules based on the individual traits returns an integer vector after matching the database and query feature vectors. The integer vectors are normalized before fusion. The final score is generated by using sum of score technique using False Acceptance Rate (FAR) and False Rejection Rate (FRR) at matching score level, which is passed to the decision module. In decision module person is detected as an imposter or genuine depending on the threshold.

3. FEATURE EXTRACTION USING MFCC

3.1 SPEECH FEATURE EXTRACTION

Firstly speech feature extraction is done by converting the speech waveform to parametric representation (at a considerably lower information rate). The speech signal is a slowly time varying signal (it is called *quasi-stationar*y). When examined the characteristics are fairly stationary over short period of time (between 5 and 100 ms). However, the signal characteristics change to reflect the different speech sounds being spoken over long periods of time (on the order of 0.2s or more). Therefore, short-time spectral analysis *is* the most common way to characterize the speech signal. We have chosen of about 30ms

frame length with overlap. There are more than one techniques exist for parametrically representing the speech signal for the speaker recognition task, such as Mel-Frequency Cepstrum Coefficients (MFCC), Linear Prediction Coding (LPC), and others. The MFCC are motivated by studies of the human peripheral auditory system. MFCC is perhaps the most popular and best known. This method has been used in this paper for feature. MFCC's are based on the known variation of the human ear's critical bandwidths with frequency. The MFCC mainly makes use of two types of filter, namely, linearly spaced filters and logarithmically spaced filters. Speech signal is expressed in the Mel frequency scale, to capture the phonetically important characteristics of speech. This scale has a linear frequency spacing below 1000Hz and a logarithmic spacing above 1000 Hz. MFCC's are less susceptible for variations with respect to change in physical condition of speakers' vocal cord.

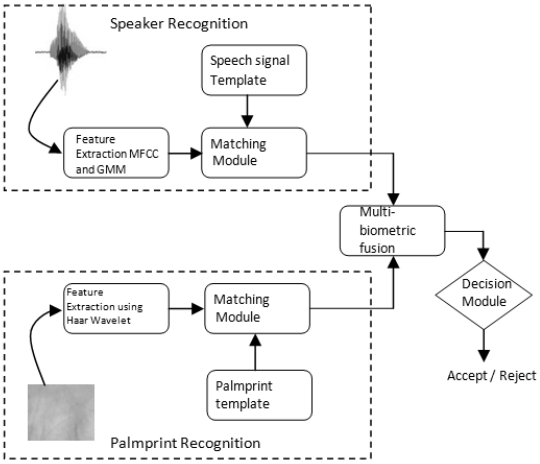

Fig.1. Block diagram of speech signal and palmprint multimodal biometric system

3.2 THE MFCC PROCESSOR

A block diagram of the structure of an MFCC processor is given in Fig. 2. To minimize the aliasing effect in analog to digital converter, we have chosen the sampling rate of 22050Hz.

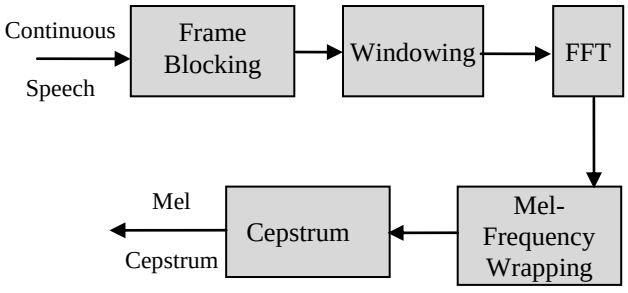

Fig.2. Block diagram of the MFCC processor

3.3 MEL-FREQUENCY WRAPPING

The speech signal consists of tones with different frequencies. For each tone with an actual Frequency, a subjective pitch is measured on the 'Mel' scale. The mel-*frequency* scale is linear frequency spacing below 1000Hz and a logarithmic spacing above 1000Hz. As a reference point, the pitch of a 1kHz tone, 40dB above the perceptual hearing threshold, is defined as 1000 mels. Therefore we can use the following formula to compute the Mels for a given frequency *f* in Hz [4]:

$$\text{mel(f)} = 2595*\log10(1+f/700). \quad (1)$$

One approach to simulating the subjective spectrum is to use a filter bank, one filter for each desired mel-frequency component. The filter bank has a triangular bandpass frequency response, and the spacing as well as the bandwidth is determined by a constant mel-frequency interval.

3.4 CEPSTRUM

In the final step, the log mel spectrum has to be converted back to time. The result is called the mel frequency cepstrum coefficients (MFCCs). Because the mel spectrum coefficients are real numbers (and so are their logarithms), they may be converted to the time domain using the Discrete Cosine Transform (DCT). The MFCCs may be calculated using this equation,

$$\tilde{C}n = \sum_{k=1}^{K} \left(\log \tilde{S}_k\right)[n(k-\tfrac{1}{2})\tfrac{\pi}{K}] \quad \text{where } n = 1, 2.....K \quad (2)$$

K, the coefficient length is typically chosen as 20. The C_0 component, is excluded since it carries little speaker specific information. $\tilde{S}_k$ is the cepstrum. By applying for each speech frame a set of mel-frequency cepstrum coefficients is computed. This set of coefficients is called an *acoustic vecto*r. These acoustic vectors can be used to represent and recognize the voice characteristic of the speaker [5]. Therefore each input utterance is transformed into a sequence of acoustic vectors.

3.5 GAUSSIAN MIXTURE MODEL

In this study, a Gaussian Mixture Model approach proposed in [6] is used where speakers are modeled as a mixture of Gaussian densities. The use of this model is motivated by the interpretation that the Gaussian components represent some general speaker-dependent spectral shapes and the capability of Gaussian mixtures to model arbitrary densities.

The Gausssian Mixture Model is a linear combination of M Gaussian mixture densities, and given by the equation,

$$p(\vec{x} \mid \lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x}) \quad (3)$$

where, $\vec{x}$ is a D-dimensional random vector, $b_i(\vec{x}), i = 1,...M$ are the component densities and p_i, i=1,…M are the mixture weights. Each component density is a D-dimensional Gaussian function of the form

$$bi(\vec{x}) = \frac{1}{(2\Pi)^{D/2} |\sum_i|^{1/2}} \exp\left\{-\frac{1}{2}(\vec{x}-\overrightarrow{\mu_i})^T \sum_i^{-1} (\vec{x}-\overrightarrow{\mu_i})\right\} \quad (4)$$

where $\overrightarrow{\mu_i}$ denotes the mean vector and Σ_i denotes the covariance matrix. The mixture weights satisfy the law of total

probability, $\sum_{i=1}^{M} p_i = 1$. The major advantage of this representation of speaker models is the mathematical tractibility where the complete Gaussian mixture density is represented by only the mean vectors, covariance matrices and mixture weights from all component densities.

4. FEATURE EXTRACTION USING HAAR WAVELET

Features are the attributes or values extracted to get the unique characteristics from the image and speech signal.

4.1 PALMPRINT FEATURE EXTRACTION METHODOLOGY

Details of the algorithm are as follows:

4.1.1 Identify Hand Image From Background:

Our designed system is such that palmprint images are captured using contact-less without pegs, keeping the image background relatively uniform and relatively low intensity when compared to the hand image. Using the statistical information of the background, the algorithm estimates an adaptive threshold to segment the image of the hand from the background. Pixels with intensity above the threshold are considered to be part of the hand image.

4.1.2 Locate Region-Of-Interest:

The palm area is extracted from the binary image of the hand. After translating the original image into binary image, we find two key positioning points in the palmprint image using automatic detecting method. The first valley in the graph is the gaps between little finger and ring finger, Key Point 1. The third valley in the graph is the gaps between middle finger and index finger, Key Point 2. The key point is circled in Fig.3. The hand image is rotated by θ degrees. The hand images are rotated to align the hand images into a predefined direction. θ is calculated using the key points as shown in the Fig.3. Since the size of the original image is large, a smaller hand image is cropped out from the original hand image after image alignment using key points. Fig.4 shows the proposed image alignment and ROI selection method.

$$I(x_i, y_i) = \begin{cases} 0, & if \quad |I(x_i, y_i)| \le std(I(x, y)) \\ \ln(\ |I(x_i, y_i)| - std(I(x, y)) + 1\ ), & o.w. \end{cases} \quad (5)$$

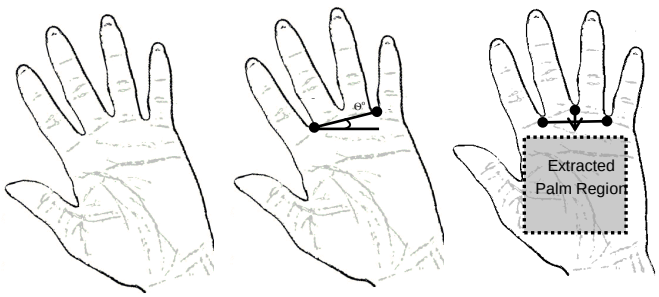

Fig.3. Schematic diagram of image alignment

Fig.4. Segmentation of ROI

4.2 FEATURE EXTRACTION

Firstly, a 2-D lowpass filter is applied to the image. The result is subtracted from the image to minimize the non-uniform illumination effect. Secondly, a Gaussian window is used to smooth out the image since Haar wavelet, due to its rectangular wave nature, is sensitive to noise and also it can be manually tuned.

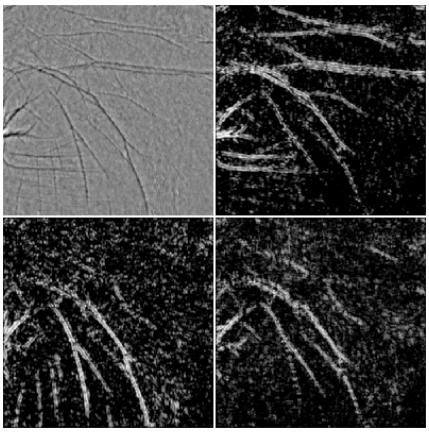

Fig.5. Haar wavelet transform of Palmprint

A 1-level decomposition of the image by the Haar wavelet is carried out. For each of the three detail images obtained, i.e. image consisting of the horizontal, vertical and diagonal details, a smoothing mask is applied to remove noise. It was found that most of the low frequency components are attributable to the redness underneath the skin and should preferably be excluded from features for identification. Thus, pixels with frequency values within one standard deviation are set to zero. Values of the rest of the pixels are projected onto a logarithm scale so as to minimize the absolute differences in the magnitude of the frequency components between two images. That is, where $I(x_i,y_i)$ is the frequency value in a detail image. The processed image is shown in Fig.5.

4.3 MATCHING SCORE CALCULATION

Since the palm images under process are divided into square cells of same widths regardless of the size of the original image, different palm sizes will result in feature vectors of different lengths. Due to the possibility of having variations in the extent the hand is stretched, the resultant maximum palm area may vary within the same subject. Therefore, the distance measure used

must be able to fairly compare two feature vectors with unequal dimension.

The score is calculated as the mean of the absolute difference between two feature vectors. If featureVi represents a feature vector of Ni elements, the score between two images is given as:

$$Score(i, j) = \frac{\sum_{n=1}^{\min(N_i, N_j)} | \; featureV_i(n) - featureV_j(n) \; |}{\min(N_i, N_j)} \quad (6)$$

5. FUSION

The biometrics systems is integrated at multi-modality level to improve the performance of the verification system. At multi-modality level, matching score are combined to give a final score. The following steps are performed for fusion:

1. Given a query image and speech signal as input, features are extracted by the individual recognition and then the matching score of each individual trait is calculated.
2. The weights a and b are calculated using FAR and FRR.
3. Finally, the final score after combining the matching score of each trait is calculated by weighted sum of score technique,

$$MS_{fusion} = \frac{a * MS_{Palm} + b * MS_{Speech}}{2} \quad (7)$$

where, *a* and *b* are the weights assigned to both the traits. The final matching score (MS_{fusion}) is compared against a certain threshold value to recognize the person as genuine or an imposter.

6. EXPERIMENTAL RESULTS

We evaluate the proposed multimodal system on a data set including 720 pairs of images from 120 subjects. The training database contains a speech signals and palmprint images for each individual for each subject. Each subject has 6 palm images taken at different time intervals and 6 different words, which is stored in the database. Before extracting features of palmprint, we locate palmprint images to 128 x 128.

Fig.6 shows identification rate when triangular, or rectangular or hamming window is used for framing in a linear frequency scale. The table clearly shows that as codebook size increases, the identification rate for each of the three cases increases, and when codebook size is 16, identification rate is 100% for the hamming window. However, in case of Fig.7 the same windows are used along with a Mel scale instead of a linear scale. Here, too, identification rate increaseswith increase in the size of thecodebook. In this case, 100% identification rate is obtained with a codebook size of 8 when hamming window is used.

The accuracy of Unimodal vs Multimodal is as shown in Fig.8. The multimodal system has been designed at matching score level. At first experimental the individual systems were developed and tested for FAR, FRR & accuracy. In the last experiment both the traits are combined at matching score level using sum of score technique. The results are found to be very encouraging and promoting for the research in this field. The overall accuracy of the system is more than 98%, FAR & FRR of 1.8% & 0.8% respectively. Table.1 shows FAR, FRR & Accuracy of the systems.

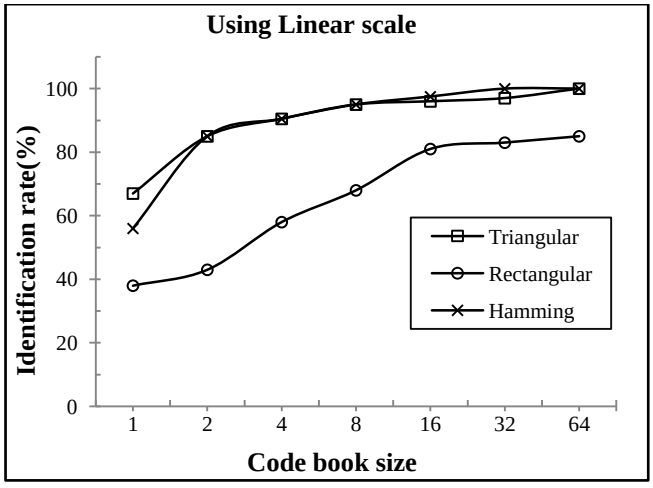

Fig.6. Identification rate (in %) for different windows (using Linear scale)

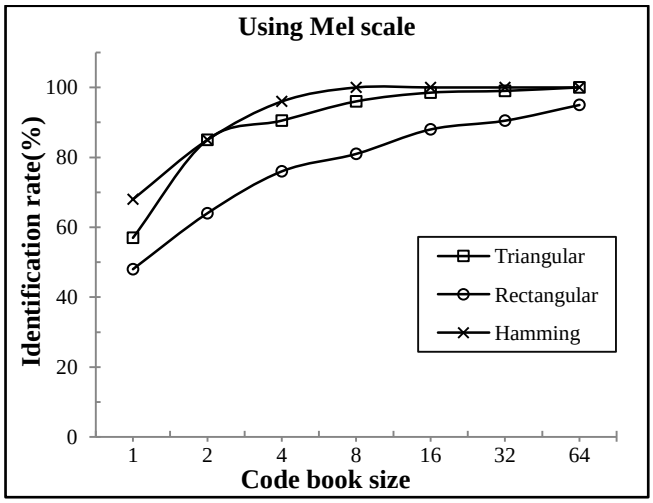

Fig.7. Identification rate (in %) for different windows (using Melscale)

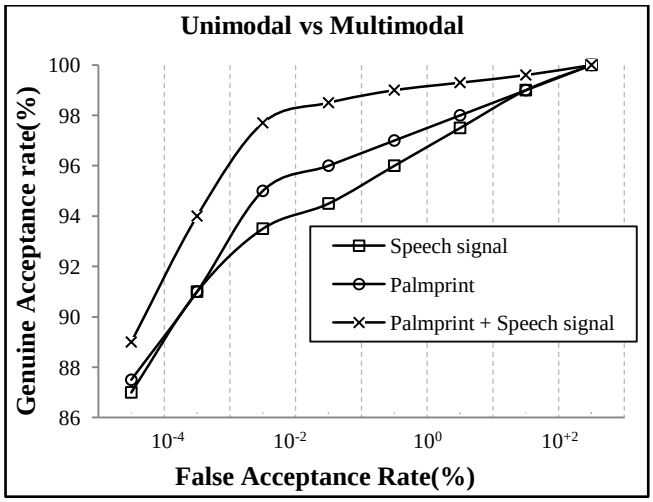

Fig.8. Unimodal vs Multimodal

Table.1. Accuracy, FAR, FRR of individual recognition and after Fusion

Trait	Algorithm	FAR (%)	FRR (%)	Accuracy (%)
Palmprint	Haar Wavelet	4.2	1.3	95.8
Speech Signal	MFCC	5.3	7.4	94.1
Palmprint+ Speech Signal	Weighted sum of score techniques	1.8	0.8	98.2

7. CONCLUSION

Biometric systems are widely used to overcome the traditional methods of authentication. But the unimodal biometric system fails in case of biometric data for particular trait. Thus the individual score of two traits (speech signal & palmprint) are combined at classifier level and trait level to develop a multimodal biometric system. The performance table shows that multimodal system performs better as compared to unimodal biometrics with accuracy of more than 98%.

REFERENCES

[1] A. Ross, K. Nandakumar, and A. K. Jain, "*Handbook of Multibiomtrics*", Springer-Verlag, 2006.

[2] Mahesh P.K. and M.N. Shanmukhaswamy, "Comprehensive Framework to Human Recognition Using Palmprint and Speech Signal", *Communications in Computer and Information Science In Springer-Verlag Berlin Heideberg*, Vol. 131, pp. 368-377, 2011.

[3] Mahesh P.K. and M.N. Shanmukhaswamy, "Integration of multiple cues for human authentication system", *Proceedings of the International Conference and Exhibirion on Biometrics Technology, In Procedia Computer Science*, Vol. 2, pp.188-194, 2010.

[4] Jr., J. D., Hansen, J., and Proakis, J, "Discrete Time Processing of Speech Signals", *second edition IEEE Press, New York*, 2000.

[5] Comp.speech Frequently Asked Questions WWW site, http://svr-www.eng.cam.ac.uk/comp.speech/.

[6] D. A. Reynolds, "Experimental Evaluation of Features for Robust Speaker Identification," *IEEE Transactions on Speech and Audio Processing (SAP)*, Vol. 2, pp. 639-643, 1994.