Bimodal Biometric Person Authentication Using Speech and Face Under Degraded Condition

Soyuj Kumar Sahoo and S R Mahadeva Prasanna
Dept. of ECE, IIT Guwahati, Guwahati 781039, India
Email: {soyuj, prasanna}@iitg.ernet.in

Abstract—In this work, we present a bimodal biometric system using speech and face features and tested its performance under degraded condition. Speaker verification (SV) system is built using Mel-Frequency Cepstral Coefficients (MFCC) followed by delta and delta-delta for feature extraction and Gaussian Mixture Model (GMM) for modeling. A face verification (FV) system is built using the combination of Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). Sum rule is used for the fusion of the biometric scores. The performance of SV system under degraded condition is also checked. All the experimental results are shown upon a subset of IITG-DIT M4 multi-biometric database. The complementary information derived from the speech biometric at training stage is used to further decrease the FV error rate, which is termed as Cohort fed FV system. Finally we propose an improved bimodal person authentication system using SV and Cohort fed FV biometric systems.

Index Terms—Biometric, Speaker verification, Face verification, Multimodal system, score level fusion, Cohort fed.

I. INTRODUCTION

Person authentication can be performed by different methods like what we know (knowledge), what we have (token), and what we are (biometric) (e.g., face, gait). Password or card can be shared, forgotten or stolen, but not the biometric. Acquisition of biometric is more complex compared to making combinations of digits or stealing the card. In this way, biometric is more secure compared to PIN and password. Passwords are desirable to be different for different applications, but same biometric can be used for most of the applications and hence avoids book keeping. Any human physiological or behavioral characteristic can be used as a biometric characteristic (indicator) to make personal identification as long as it satisfies the requirements of the biometric like Universality, Distinctiveness, Permanence, Collectability, Performance, Acceptability and Circumvention [1]. But in practice, no single biometric can satisfy all the requirements of an ideal biometric system due to problems such as noisy data, intra-class variations, non-universality, spoof attacks, and unacceptable error rates. Here comes the use of multiple biometrics. Based on whether single or multiple biometrics are used for person authentication, biometric systems can be classified as Unimodal Biometric Systems and Multimodal Biometric Systems, respectively. The major limitation of biometrics is the effect of environment in which the system is operating. For instance, in case of poor illumination condition, face recognition system performance may degrade significantly.

In the literature multimodal biometric systems based on fingerprint, face and speech [2], face, fingerprint and hand geometry features [3], palmprint and hand geometry features [4], speech and signature [5], speech, face and signature features [6] have been described. In recent times much of the interest is concentrated in audio-visual multibiometric systems [7]. In this paper we describe a bimodal system using face and speech features under degraded condition. The main motivation behind the selection of face and speech as the biometric features for building a multimodal biometric system is that, face is the mostly used and speech is one of the features used by humans for person authentication with nominal error rate. Again both contain complimentary information about human characteristics with quite distinct causes of degradation. For instance, the effect of illumination in speech is nil compared to the effect on signature or palmprint, which are being used for person authentication in most of the practical applications since a long time. The other reason for the selection of these biometrics is the ease of collection of biometric data. That is, the sensors cost is very less and also set up for data collection will not need any special arrangement, a laptop with web camera and headset is sufficient to collect the data.

In the present work, first a multi-sensor database containing speech as well as face features for 94 speakers is collected. The database consists of simultaneous recording of speech data over multi-sensors. A speaker verification (SV) system is developed by using mel-frequency cepstral coefficients (MFCC) as feature and Gaussian Mixture Models (GMM) as the modeling technique [8]. A face verification (FV) is developed using subspace, principal component analysis (PCA) and linear discriminant analysis (LDA) techniques [9]. Score level fusion is implemented with well known z-score normalization technique to integrate the individual unimodal systems [10]. The system performance is compared against two different sensors to check the degraded environment performance. The cohort group information [11], which are quite different for these complementary biometrics is used to improve the bimodal performance further in both clean and degraded conditions.

The organization of the remaining work is as follows: Section II describes database used, Section III and IV describe each of the unimodal biometric systems, the SV system and the FV system, respectively. After the details of the unimodal systems, we describe the bimodal system in Section II.
V. Experimental results are presented in Section VI and finally summary and future work are provided in Section VII.

II. DATABASE

We have used a subset of IITG-DIT Multi- Biometric-Sensor- Environment- Language- Style (M4) database developed in house for these studies. IITG-DIT M4 database is collected in a setup having five different sensors, two different environments, different Indian languages and two different styles. The five different sensors include headphone microphone mounted close to the speaker, inbuilt tablet PC microphone, two mobile phones and one digital voice recorder. Except for the headphone microphone, all the other four sensors are placed at a distance of about 2 – 3 feet from the speaker. Speech was recorded simultaneously over these sensors. Speech recorded in headphone microphone and inbuilt tablet PC microphone are at 16 kHz and stored with 16 bits/sample resolution. Speech recorded in digital voice recorder is at 44.1 kHz and stored with 16 bits/sample, which is later resampled to 16 kHz and stored at 16 bits/sample. The speech recorded in two mobiles are at 8 kHz and sampled at 16 bits/sample. The recording was done in two different environments, namely, office and hostel rooms. The recording was done in two languages, namely, English and favorite language of the speaker which happens to be one of the Indian languages like Hindi, Telugu, Kannada, Oriya and so on.

Suitable positioning of the headset was decided such that video as well as speech data recording can be performed at the same time. Video streams are recorded using Logitech CMOS Webcam with well illuminated room in different sessions. No restriction is put on users about their sitting, dressing, pose and hair styles. Recording is done for 30 – 60 seconds having frame rate of 15 fps (frames per second) with 640 × 480 pixel resolution per frame. Finally the video files are stored as Windows Media Video (WMV) format. Some of the practical degraded environment effects form session to session that exists in our database as shown in Fig. 1.

III. SPEAKER VERIFICATION SYSTEM

A. Feature Extraction

In the training and testing process, the speech signal is processed in frames of 20 ms at 10 ms frame rate. For each 20 ms Hamming windowed frame, MFCC are calculated using 22 mel-frequency band filters. Excluding the zeroth coefficient remaining 13 coefficient are taken. Delta (Δ) and delta-delta (ΔΔ) of MFCC are also computed using two preceding and two succeeding feature vectors from the current feature vector. Thus the feature vector will be of 39 dimension with 13 MFCC, 13 ΔMFCC and 13 ΔΔMFCC. The silence frames are then removed based on energy threshold (0.06 × E_{avg}, where E_{avg} refers to average energy) of speech file. Cepstral mean subtraction (CMS) and variance normalization is implemented over feature sets before applying training and testing to handle the channel mismatch case. As CMS reduces the performance when there is not much variability in the recording sensor and environment, and it improves the performance when there is variation [8]. In the present study both channel match and mismatch conditions are considered.

B. Training

GMM, the most successful model for text-independent speaker verification, can be viewed as the combination of parametric (Hidden Markov Model, i.e HMM) and nonparametric (k-mean clustering) models. In training phase, the maximum likelihood model parameters are estimated using the expectation-maximization (EM) algorithm in an iterative manner for a given training set. From the observed training vectors the GMM parameters \( \{w_i, \mu_i, \Sigma_i\}, \ i = 1, 2, ..., M \) are optimized to increase the likelihood \( p(X_{\text{train}}/\lambda) \) of the estimated user model, where \( w_i, \mu_i, \Sigma_i, M, X_{\text{train}} \) and \( \lambda \) refers to the weight, mean, variance, no of Gaussians mixed in each speaker’s GMM, training feature vector set and the training model, respectively. Hence \( N \) no. of GMMs each per speaker is developed.

\[
\lambda_j = \{w_i, \mu_i, \Sigma_i\}, \ i = 1, 2, ..., M
\]

where \( j = 1, 2, ..., N \) and \( \lambda_j \) refers to the model corresponding to \( j^{th} \) speaker.

C. Testing

In testing stage, each incoming test signal is tested against a claimed model. The likelihood scores of all models are further used to obtained the performance.

IV. FACE VERIFICATION SYSTEM

Face is one of the important biometrics for authentication and hence can add a second level of verification, in case, speaker verification system is indecisive. Face Recognition is one of the most successful applications of image processing. Face recognition is the problem of identifying three-dimensional (3D) face objects from two dimensional (2D) face...
images. Face recognition approaches can be widely classified as feature-based approaches and appearance-based approaches. However research is mainly concentrated in appearance-based approaches. This is because even though feature-based approaches are less sensitive to variations in illumination and viewpoint and to inaccuracy in face localization, the feature extraction techniques needed for feature-based approaches are not still reliable or accurate enough. Among the appearance-based methods, we use subspace methods like PCA or eigenfaces followed by LDA or fisherfaces. The different stages in the face recognition are carried out in the following way:

A. Preprocessing

In the preprocessing stage the images are manually cropped and resized to a fixed dimension. The images are cropped in such a way that only the facial regions are captured, although the alignment of the faces need not be uniform. Before finding feature vectors for each user, histogram equalization is done both in training and testing.

B. Training

In this stage, we apply these dimensionality reduction techniques to the images that are selected for training using PCA. PCA is based on information theory approach that decomposes face images into a small set of characteristic feature images called eigenfaces, which may be thought of as the principal components of the initial training set of face images [9]. Then LDA is used to find an optimal linear transformation that maximizes the class separability. LDA does so by maximizing the ratio of between-class scatter to the within-class scatter [9]. Then the weight vector (a set of \( N - 1 \) weights) for each of the training face images using the \( N - 1 \) fisherfaces by projecting each image is computed.

C. Testing

In testing stage, each incoming test image is tested against a claimed template. After dimension reduction using PCA, the weight vector (a set of \( N - 1 \) weights) for the testing image is calculated using the \( N - 1 \) fisherfaces. The Euclidean distances between the weight vector of each face class and the weight vector of the testing images are calculated. These scores are further used to obtain the performance of the systems.

V. BIMODAL BIOMETRIC SYSTEM

In a bimodal biometric system, two biometrics may be from two different sensors for the same biometric, two different representations for the same biometric, or two different biometric traits. A bimodal biometric system based on different traits is expected to be more robust to noise, address the problem of non-universality, improve the matching accuracy, and provide reasonable protection against spoof attacks. This is the motivation for selecting a bimodal biometric system using face and speech as traits for our work. The four important modules in a biometric system are sensor module, feature extraction module, matching module, decision module. The biometric data can be fused at any of these levels. Prior to classification, integration of information can take place either at the sensor level or at the feature level. Biometric systems that integrate information at an early stage of processing are believed to be more effective than those systems which perform integration at a later stage, since the features contain richer information about the input biometric data than the matching score or the output decision of a classifier. However, integration at the feature level is difficult to achieve in practice. Schemes for integration of information after the classification stage can be divided into four categories: dynamic classifier selection, fusion at the abstract, rank, and matching score levels. Next to the feature vectors, the matching scores contain the richest information about the input pattern. Also, it is relatively easy to access and combine the scores generated by the different matchers. Consequently, integration of information at the matching score level is the most common approach in multimodal biometric systems. This is the reason for selecting score level fusion in our work [10].

Two factors important to the accuracy of multimodal biometric system using score level fusion are the choice of the technique used for score normalization and also for data fusion. There are many ways in which the biometric data can be normalized, and the biometric scores can be fused. Experiments indicate that min-max normalization or z-score normalization followed by sum rule performs better than any other combination [10]. So, in this work we use z-score normalization for score normalization and sum rule for biometric data fusion. Again if the two scores are in different scales; like in our case, one is in similarity scale (i.e. the likelihood scores got from SV system) and another one is in the error scale (i.e. the minimum distance scores got from FV system), then after normalization simple addition will not work. Hence we require a score transformation method to convert all these scores to a single scale. After normalization we transfer individual system scores to error scale by using following method,

\[
x' = \begin{cases} \frac{x - \mu_x}{\sigma_x} & \text{scores in error scale} \\ \frac{x'}{x_{\text{max}}} & \text{scores in similarity scale} \end{cases}
\]

where \( x' \) is the new transformed score, \( x \) is the old score, \( \mu_x \) mean of the scores, \( \sigma_x \) variance of the scores, \( x_{\text{max}} \) and \( x_{\text{min}} \) are the maximum and minimum score, respectively.

A. Cohort fed Unimodal system

All the individual unimodal systems in a multimodal system doesn’t give high performance always. Hence we choose complimentary biometric characteristics to improve the multimodal performance. At the training stage, performance of individual unimodal systems are calculated using some portions of training data not used in training the models. Then the individual model information collected from the highly performing unimodal system, like cohort group or closely scoring group information [11], is fed to the low performing unimodal system in testing stage to further improve the performance. Hence
the bimodal system, performance increases. The following are the steps involved in the implementation of bimodal biometric system using face, speech features and the cohort group information collected in training stage from speech (which is the high performing unimodal system in our case).

1) Take the multiple biometrics from an individual.
2) Extract the set of features from each biometric using corresponding feature extraction methods.
3) Compare these feature vectors with corresponding system’s database and get matching scores.
4) Cohort group information collected at training stage is fed to the FV system to improve the performance.
5) Normalize the matching scores obtained from each biometric feature using z-score normalization.
6) Transform the scores to a common scale before adding.
7) Find the sum of matching scores of each unimodal corresponding to every individual.
8) If the final score crosses the threshold, then the claim is declared as a positive claim.

VI. EXPERIMENTS AND OUTCOMES

A robust verification system should distinguish clearly between a positive claim and negative claim. In the mismatched condition the performance of a verification system is measured in terms of false rejection rate (FRR) and false acceptance rate (FAR). Equal error rate (EER) is defined as the error rate at which both FAR and FRR are equal. Hence the multimodal verification system performance is calculated in terms of EER using detection error trade-off (DET) curve plot [12]. In case of SV system, keeping the language as English and conversational style, experiments are conducted on IITG-DIT M4 database as follows:

1) Sensor matched condition: Training and testing speech are collected over the same sensor.
2) Sensor mismatched condition: Training and testing speech are collected over different sensors.

In order to compare the performance obtained using bimodal system, the sensor matched SV system is termed as baseline system. The FV, Cohort fed FV, sensor mismatched SV, bimodal and Cohort fed bimodal system performances are compared against the baseline system using DET curve plot.

A. Experimental details

For the present work, we consider 94 speakers set of IITG-DIT M4 database which include 74 male speakers and 20 female speakers. The speech recorded in the headphone microphone (H01) is more clean compared to the speech recorded in inbuilt tablet PC microphone (T01), which is more affected by environmental environmental noise like air conditioner, fan sound and reverberation. The initial 2 minutes of first session speech data recorded in H01 is used for building the models. We have considered fixed no of Gaussians, i.e. 128 size GMM for each speaker. Remaining first session speech data of H01 is used in testing the built system performance in identification mode to compute the 10 user cohort groups per each user. For each speaker, 10 speech segments between 30-45 sec duration from the second session are taken as test utterances. Therefore for 94 speakers set, there are in total 940 test trials. In the testing process, each test segment is tested against 11 models mentioned in detection list, out of which one is genuine model and rest are impostor models.

In case of FV system, the 20 frames collected from the different segments of first session video streams are used in training stage. Out of 20, first 16 images are used to built the FV system using LDA followed by PCA. The remaining 4 images are used to compute the threshold using verification through identification mode, which is used by Cohort fed FV system to improve the FV performance. 10 different frames from the second session video streams of each user are used for testing. Hence, there are in total 940 test images and each test image is tested against 11 models out of which one is genuine model like the SV test.

1) Cohort fed FV experiment: In testing stage while a test image is tested against any model, it is further tested against its cohort group computed in training stage by other biometric system, i.e. SV system. If the claimed model’s score qualifies for the best score among its cohort group, the score is then pushed towards the FV system’s threshold which is obtained during training phase. In this case we add a positive or negative bias (i.e. \( \pm \delta \), where \( \delta \) varies from 0.3-0.5, an experimental value) to the scores according to weather they qualify for best score among the cohort or not. This procedure is known as score tuning. In both training and testing stage scores are normalized using z-score normalization technique before computing the FV threshold and score tuning.

B. Experimental results and Discussions

Individual unimodal system (i.e. SV (for both H01 and T01), FV, Cohort fed FV system) performances in terms of EER are given in Table I and their corresponding DET curves are shown in Fig. 2(a). In case of SV system the clean speech, i.e. the speech recorded in H01 gives the best performance due to sensor match, while the degraded environment speech, i.e. the speech recorded in T01 gives poor performance due to sensor mismatch. The FV system gives poor performance due lot of variations from training to testing data with respect to illuminance, pose, facial expression, camera distance etc.

From the DET curve plots, as shown in Fig. 2(a), we can observe that the SV under degraded condition (i.e. sensor mismatch) and FV gives poor performance due to two complementary error rates. The FV system gives more false alarm rate where as the SV under degraded condition gives more miss probability rate. Hence both the biometric contains complementary information about a persons characteristics in degraded environments. Hence the Binodal system using these unimodal systems may give better performance. However as shown in Table II corresponding to Fig. 2(b) and (c), the error rates of Bimodal system using clean/degraded speech and face features decreased, hence performance increased. Again as the speech and face contain complementary information, the cohort groups for each of them are different, which are more competing models for a negative claim. So, when the
FV system is tested against the SV system’s cohort groups the false alarm rate decreases and hence the performance increases as shown in Table I and Fig. 2(a). Therefore, combination of Cohort fed FV system and SV further decreases the bimodal verification error rate, as shown in Table II and Fig. 2(b) and (c). Hence FV biometric can add a second level of verification, in case, SV system is indecisive.

**TABLE I**

**PERFORMANCE OF INDIVIDUAL UNIMODAL SYSTEMS IN TERMS OF EER.**

<table>
<thead>
<tr>
<th>Unimodal System</th>
<th>EER</th>
</tr>
</thead>
<tbody>
<tr>
<td>SV against H01</td>
<td>8.08</td>
</tr>
<tr>
<td>SV against T01</td>
<td>22.65</td>
</tr>
<tr>
<td>FV</td>
<td>22.87</td>
</tr>
<tr>
<td>Cohort fed FV</td>
<td>19.89</td>
</tr>
</tbody>
</table>

**TABLE II**

**PERFORMANCE OF BIMODAL SYSTEMS WITH COMBINING DIFFERENT UNIMODAL SYSTEMS IN TERMS OF EER.**

<table>
<thead>
<tr>
<th>FV</th>
<th>SV against H01</th>
<th>SV against T01</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original FV</td>
<td>7.44</td>
<td>17.23</td>
</tr>
<tr>
<td>Cohort fed FV</td>
<td>5.85</td>
<td>14.04</td>
</tr>
</tbody>
</table>

**VII. SUMMARY AND CONCLUSIONS**

We have presented a bimodal person verification system using speech and face as biometrics. Experimental results clearly indicate the efficacy of bimodal system even under degraded environment. We can see from the Table II, that the performance of the bimodal system increases in both clean and degraded case compared to their individual unimodal performances. The cohort fed experiment shows that the complementary information collected from other biometric can be used to improve performance. From the results and discussions of Sec. VI-B, the proposed cohort fed bimodal system using two different biometrics gives the best performance in both clean and noisy conditions.

In future we may try to make individual unimodal systems more robust and test the robustness of proposed cohort fed system. Future work may include exploring the concept of cohort fed in case of other biometrics.

**REFERENCES**