Hindawi Publishing Corporation
ISRN Artificial Intelligence
Volume 2014, Article ID 737814, 13 pages
http://dx.doi.org/10.1155/2014/737814

Research Article

BPN Based Likelihood Ratio Score Fusion for Audio-Visual Speaker Identification in Response to Noise

Md. Rabiul Islam[1] and Md. Abdus Sobhan[2]

[1] *Department of Computer Science & Engineering, Rajshahi University of Engineering & Technology, Rajshahi 6204, Bangladesh*
[2] *School of Engineering & Computer Science, Independent University, Dhaka 1229, Bangladesh*

Correspondence should be addressed to Md. Rabiul Islam; rabiul_cse@yahoo.com

Received 25 September 2013; Accepted 3 November 2013; Published 8 January 2014

Academic Editors: J. Molina, M. Monti, M. Ture, and J. M. Usher

This paper deals with a new and improved approach of Back-propagation learning neural network based likelihood ratio score fusion technique for audio-visual speaker Identification in various noisy environments. Different signal preprocessing and noise removing techniques have been used to process the speech utterance and LPC, LPCC, RCC, MFCC, ΔMFCC and ΔΔMFCC methods have been applied to extract the features from the audio signal. Active Shape Model has been used to extract the appearance and shape based facial features. To enhance the performance of the proposed system, appearance and shape based facial features are concatenated and Principal Component Analysis method has been used to reduce the dimension of the facial feature vector. The audio and visual feature vectors are then fed to Hidden Markov Model separately to find out the log-likelihood of each modality. The reliability of each modality has been calculated using reliability measurement method. Finally, these integrated likelihood ratios are fed to Back-propagation learning neural network algorithm to discover the final speaker identification result. For measuring the performance of the proposed system, three different databases, that is, NOIZEUS speech database, ORL face database and VALID audio-visual multimodal database have been used for audio-only, visual-only, and audio-visual speaker identification. To identify the accuracy of the proposed system with existing techniques under various noisy environment, different types of artificial noise have been added at various rates with audio and visual signal and performance being compared with different variations of audio and visual features.

1. Introduction

Biometric authentication [1] has grown in popularity as a way to provide personal identification. Person's identification is crucially significant in many applications and the hike in credit card fraud and identity thefts in recent years indicate that this is an issue of major concern in wider society. Individual passwords, pin identification, or even token based arrangement all have deficiencies that restrict their applicability in a widely networked society. Biometrics is used to identify the identity of an input sample when compared to a template, used in cases to identify specific people by certain characteristics. No single biometrics is expected to effectively satisfy the needs of all identification applications. A number of biometrics have been proposed, researched and evaluated for authentication applications. Each biometrics has its strengths and limitations, and accordingly, each biometrics appeals to a particular identification application [2]. Biometric characteristics can be divided in physiological and behavioral classes [3]. Physiological characteristics are related to the shape of the body and thus it varies from person to person. Fingerprints, face recognition, hand geometry, and iris recognition are some examples of this type of Biometrics. Behavioral characteristics are related to the behavior of a person. Some examples in this case are signature, keystroke dynamics, voice, and so on.

The Audio-Visual speaker identification system combines the speech and face biometric characteristics which mix the physiological and behavioral characteristics. There are different levels where the audio and visual features can be concatenated. Preclassification and post-classification are the two broad categories for information fusion in biometric

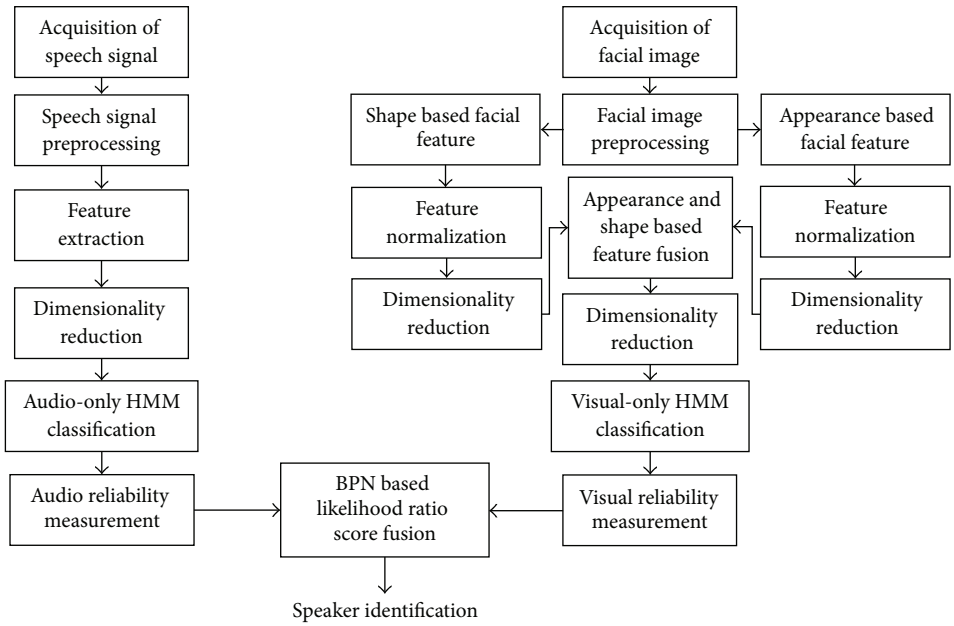


Figure 1: Paradigm of the proposed BPN based likelihood ratio score fusion for audio-visual speaker identification system.

system [4]. In preclassification, multimodal information is fused before going to the classifier decision. But, in post-classification, information is combined after the decision of multiple classifiers. In this proposed system, feature level fusion has been performed under preclassification approach. Appearance and shape based facial features are combined for visual identification result. Decision level fusion has been applied in the proposed system where audio and visual speaker identification decisions are concatenated to find out the final identification result. However, both feature level and decision level fusion are performed for the proposed system.

The rest of the paper is constructed as follows: Section 2 describes the literature review and the proposed system architecture, Section 3 focuses on the audio-only identification performance, visual-only identification has been elaborated in Section 4, Audio and visual reliability measurement techniques are focused on Section 5 and back-propagation learning neural network based likelihood ratio score fusion technique has been shown in Section 6. NOIZEUS speech database has been used to measure the performance of the Audio-Only speaker identification, visual-only performance has been populated using ORL database and overall system performance that is, back-propagation learning neural network score fusion based performance has been counted by applying VALID audio-visual multimodal database.

2. Literature Review and Proposed System Architecture

Since human speech is bimodal in nature [5, 6], visual speech information can play a vital role for the improvement of natural and robust human-computer interaction [7–11]. Most published works in the areas of speech recognition and speaker recognition focus on speech under the noiseless environments and few published works focus on speech under noisy conditions [12–15]. Indeed, various important human-computer components, such as speaker identification, verification [16, 17], localization [18], speech event detection [19], speech signal separation [20], coding [21], video indexing and retrieval [22] and text-to-speech [23, 24] have been shown to benefit from the visual channel [25]. Adaptive weighting in decision fusion with acoustic and visual features from a given Audio-Visual speech datum, the recognized utterance class has been proposed [26]. The reliability of each audio and visual modality can be measured in various ways such as average absolute difference of loglikelihood [27], variance of loglikelihood [28], average difference of log-likelihood from the maximum [29], and inverse entropy of posterior probability [30]. Decision level information integration techniques have been developed where each biometric matcher individually decides on the best match based on the input presented to it. Methods like majority voting [31], behavior knowledge space [32], weighted voting based on the Dempster-Shafer theory of evidence [33], AND rule and OR rule [34], and so forth are some of the decision level fusion techniques proposed by different researchers.

The proposed architecture of the audio-visual speaker identification system is shown in Figure 1. Signal preprocessing and noise removing techniques have been applied after acquisition of the speech utterances. Then features are extracted using various standard speech feature extraction methods such as LPC, LPCC, RCC, MFCC, ΔMFCC, and ΔΔMFCC. Principal Component Analysis (PCA) has been used to reduce the dimensionality of the extracted feature vector. Now the reduced feature vector is feed to Discrete Hidden Markov Model (DHMM) to get the log likelihood of each speech modality. Reliability measurement method

has been used to measure the reliability for audio signal. For visual identification, captured faces are preprocessed using different noise removing techniques and Active Shape Model (ASM) is used to extract the appearance and shape based features. These two different types of features are fused after applying feature normalization and PCA based dimensionality reduction techniques. The concatenations of these features are important in the sense when the appearance based feature is captured with noise (i.e., light variations) then shape based features can retain the performance on a satisfied level. This is also true when the shape based feature is captured by noise highly. By combining this approach, the proposed system performs very well especially in various lighting environmental conditions. Finally, log likelihood of visual modality has been evaluated using DHMM classification and reliability has been measured using the same reliability measurement technique like audio modality. Integrated weights of audio and visual reliability measurement are fed to the Back-propagation learning neural network algorithm to calculate the final speaker identification result.

Rogozan and Deléglise [26] developed a technique for combining different likelihoods of multilevel biometric identification. In this proposed system, BPN algorithm has been used to combine the likelihood of audio and visual modality to enhance the performance of audio-visual speaker identification. This is the main contribution of the proposed system. Experimental results show the superiority of BPN based approach over the Rogozan and Deléglise [26] method in terms of audio-visual speaker identification system.

3. Audio-Only Speaker Identification

3.1. Speech Signal Preprocessing and Feature Extraction. Speech signal preprocessing plays an important role for the efficiency of speaker identification. After capturing the speech utterances, wiener filter has been used to remove the background noise from the original speech utterances [35]. The wiener filter is a noise removing filter based on Fourier iteration. Its main advantage is the short computational time it takes to find a solution [36].

Let $s(t)$ be the smear signal let and $r(t)$ be the known response that causes the convolution. Then $s(t)$ is related to $u(t)$ by

$$s(t) = \int_{-\infty}^{\infty} r(t-\tau)\, u(\tau)\, d\tau \tag{1}$$

or

$$S(f) = R(f)\, U(f), \tag{2}$$

where S, R, U are Fourier Transform of s, r, and u. Consider the following.

The second source of signal corruption is the unknown background noise $n(t)$. Therefore the measured signal $c(t)$ is a sum of $s(t)$ and $n(t)$

$$c(t) = s(t) + n(t). \tag{3}$$

To deconvolve s to find u, simply divide $S(f)$ by $R(f)$ that is, $U(f) = S(f)/R(f)$ in the absence of noise n. To deconvolve c where n is present then one needs to find an optimum filter function $\phi(t)$ or $\phi(f)$ which filters out the noise and gives a signal $\tilde{u}$ by

$$\tilde{U}(f) = \frac{C(f)\,\phi(f)}{R(f)}, \tag{4}$$

where $\tilde{u}$ is as close to the original signal as possible.

For $\tilde{u}$ to be similar to u, their differences square is as close to zero as possible; that is,

$$\int_{-\infty}^{\infty} |\tilde{u}(t) - u(t)|^2 dt \tag{5}$$

or

$$\int_{-\infty}^{\infty} |\tilde{u}(f) - u(f)|^2 df \tag{6}$$

is minimized.

Substituting the above three equations, the Fourier version becomes:

$$\int_{-\infty}^{\infty} |R(f)|^{-2} |S(f)|^2 |1-\phi(f)|^2 + |N(f)|^2 |\phi(f)|^2 df \tag{7}$$

after rearranging. The best filter is one where the above integral is a minimum at every value of f. This is, when

$$\phi(f) = \frac{|S(f)|^2}{|S(f)|^2 + |N(f)|^2}. \tag{8}$$

Now, $|S(f)|^2 + |N(f)|^2 \approx |C(f)|^2$, where $|C(f)|^2$, $|S(f)|^2$, and $|N(f)|^2$ are the power spectrum of C, S, and N. Therefore,

$$\phi(f) \approx \frac{|S(f)|^2}{|C(f)|^2}. \tag{9}$$

Figure 2(a) shows a sample signal with background noise and Figure 2(b) shows the signal after applying the wiener filter.

Speech end points detection and silence part removal algorithm have been used to detect the presence of speech and to remove pulse and silences in the speech utterance [37, 38] which is shown in Figure 3.

To detect word boundary, the frame energy is computed using the short-term log energy equation [39]

$$E_i = 10 \log \sum_{t=n_i}^{n_i+N-1} S^2(t). \tag{10}$$

Preemphasis has been used to balance the spectrum of voiced sounds that have a steep roll-off in the high frequency region [38]. The transfer function of the FIR filter in the z-domain is [40]

$$H(Z) = 1 - \alpha \cdot z^{-1}, \quad 0 \le \alpha \le 1, \tag{11}$$

where α is the preemphasis parameter.

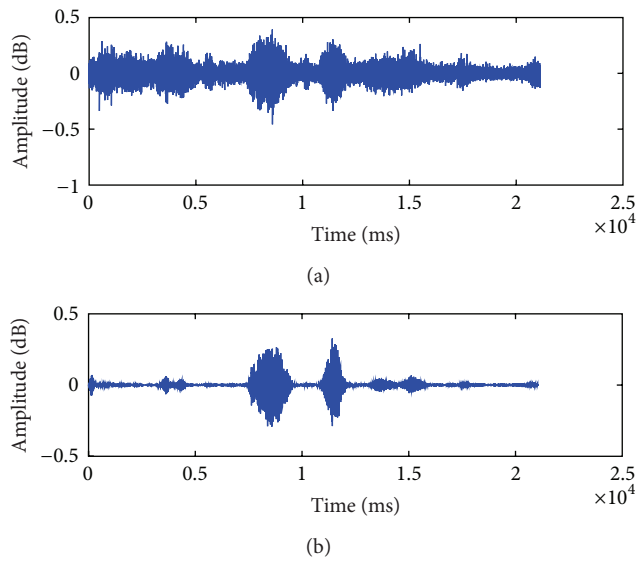


Figure 2: Effects of filtering technique (a) signal with noise (b) signal after applying Wiener filtering technique.

Frame blocking has been performed with an overlapping of 25% to 75% of the frame size. Typically a frame length of 10–30 milliseconds has been used. The purpose of the overlapping analysis is that each speech sound of the input sequence would be approximately centered at some frame [41].

From different types of windowing techniques, a hamming window has been used for this system. The purpose of using windowing is to reduce the effect of the spectral artifacts that results from the framing process [42]. The hamming window can be defined as follows [43]:

$$w(n) = \begin{cases} 0.54 - 0.46\cos\dfrac{2\pi n}{N}, & -\left(\dfrac{N-1}{2}\right) \le n \le \left(\dfrac{N-1}{2}\right) \\ 0 & \text{Otherwise.} \end{cases} \tag{12}$$

To extract the audio features, RCC, LPCC, MFCC, ΔMFCC, and ΔΔMFCC based various standard speech feature extraction techniques [44, 45] have been used to enhance the efficiency of the system because the quality of the system depends on the proper feature extracted values.

3.2. Experimental Results according to NOIZEUS Speech Database. NOIZEUS speech corpus [46, 47] has been used to calculate the accuracy of the audio-only speaker identification system which contains 30 IEEE sentences (produced by three male and three female speakers) corrupted by eight different real-world noises at different SNRs. The noise was taken from the AURORA database and includes suburban train noise, babble, car, exhibition hall, restaurant, street, airport and train-station noise. The sentences were originally sampled at 25 kHz and downsampled to 8 kHz [48].

To measure the performance of the system according to NOIZEUS speech database, one clean speech utterance has been used for learning and four different noisy speeches

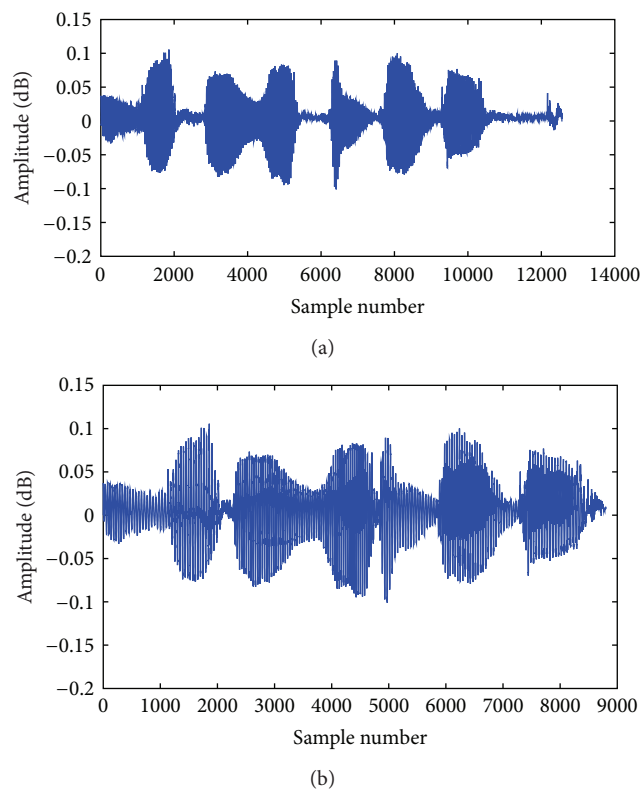


Figure 3: (a) Speech with silence parts (b) Results after applying silence parts removal algorithm.

ranging from 0 dB to 20 dB with 5 dB interval are used for testing purpose. Tables 1, 2, 3, 4, 5, 6, 7, 8, and 9 show the results of audio-only speaker identification rate at different types of noisy environments with different SNRs.

Table 9 shows the overall average speaker identification rate for NOIZEUS speech corpus. From the table, it is easy to compare the performance among MFCC, ΔMFCC, ΔΔMFCC, RCC and LPCC methods for DHMM based audio-only speaker identification system. It is shown that ΔMFCC has greater performance 48.85% than any other methods such as MFCC, ΔΔMFCC, RCC and LPCC. It also shows that ΔMFCC feature can perform better than any other feature extraction method in all of eight different environmental conditions.

4. Visual-Only Speaker Identification

4.1. Facial Feature Extraction and Dimensionality Reduction. After acquisition of a face image, Stams [49] Active Shape Model (ASM) has been used to detect the facial features. Then the binary image has been taken. The Region Of Interest (ROI) has been chosen according to the ROI selection algorithm [50, 51]. Lastly the background noise has been eliminated [52] and finally appearance based facial feature has been found. The procedure of the facial image preprocessing parts is shown in Figure 4 where Figures 4(d) and 4(e) shows the shape based and appearance based facial feature respectively.

Table 1: Airport Noise Average Identification Rate (%) for NOIZEUS Speech Corpus.

SNR	Method				
	MFCC	ΔMFCC	ΔΔMFCC	RCC	LPCC
15 dB	60.67	63.33	53.67	53.67	56.67
10 dB	53.33	58.67	43.33	43.67	53.33
5 dB	50.00	53.33	30.00	33.33	43.33
0 dB	17.67	17.67	10.67	12.33	15.00
Average	45.42	48.25	34.42	35.75	42.08

Table 2: Babble Noise Average Identification Rate (%) for NOIZEUS Speech Corpus.

SNR	Method				
	MFCC	ΔMFCC	ΔΔMFCC	RCC	LPCC
15 dB	62.67	65.33	53.67	56.67	60.00
10 dB	56.33	60.33	43.33	43.67	56.67
5 dB	43.33	53.33	33.33	43.33	50.00
0 dB	17.67	18.00	11.33	13.33	12.00
Average	45.00	49.25	35.42	39.25	44.67

Table 3: Car Noise Average Identification Rate (%) for NOIZEUS Speech Corpus.

SNR	Method				
	MFCC	ΔMFCC	ΔΔMFCC	RCC	LPCC
15 dB	56.67	60.33	43.33	50.00	53.33
10 dB	56.67	56.67	33.33	40.67	53.33
5 dB	46.33	46.67	30.00	40.33	42.33
0 dB	15.33	17.33	10.00	12.33	15.00
Average	43.75	45.25	29.17	35.83	41.00

Table 4: Exhibition Hall Noise Average Identification Rate (%) for NOIZEUS Speech Corpus.

SNR	Method				
	MFCC	ΔMFCC	ΔΔMFCC	RCC	LPCC
15 dB	65.67	67.33	50.00	53.33	60.33
10 dB	60.00	63.33	43.33	46.67	56.33
5 dB	53.67	56.67	33.33	43.33	50.00
0 dB	18.33	18.33	10.33	13.33	16.33
Average	49.42	51.42	34.25	39.17	45.75

Table 5: Restaurant Noise Average Identification Rate (%) for NOIZEUS Speech Corpus.

SNR	Method				
	MFCC	ΔMFCC	ΔΔMFCC	RCC	LPCC
15 dB	60.00	63.33	43.33	53.33	56.67
10 dB	56.67	60.00	40.00	46.67	53.33
5 dB	53.33	56.67	33.33	40.00	46.67
0 dB	17.33	18.67	13.67	15.00	15.67
Average	46.83	49.67	32.58	38.75	43.09

Table 6: Street Noise Average Identification Rate (%) for NOIZEUS Speech Corpus.

SNR	Method				
	MFCC	ΔMFCC	ΔΔMFCC	RCC	LPCC
15 dB	63.33	67.67	53.67	56.67	60.00
10 dB	53.33	60.00	46.67	43.33	50.00
5 dB	50.00	50.00	30.00	43.33	43.67
0 dB	17.33	18.67	11.33	13.67	16.00
Average	46.00	49.09	35.42	39.25	42.42

Table 7: Train Noise Average Identification Rate (%) for NOIZEUS Speech Corpus.

SNR	Method				
	MFCC	ΔMFCC	ΔΔMFCC	RCC	LPCC
15 dB	63.33	67.67	50.00	53.33	58.33
10 dB	60.00	63.33	43.33	46.67	53.33
5 dB	46.67	50.00	30.00	43.33	43.67
0 dB	18.33	18.67	12.33	13.67	15.33
Average	47.08	49.92	33.92	39.25	42.67

Table 8: Train Station Noise Average Identification Rate (%) for NOIZEUS Speech Corpus.

SNR	Method				
	MFCC	ΔMFCC	ΔΔMFCC	RCC	LPCC
15 dB	63.67	63.33	48.33	50.00	58.33
10 dB	60.67	56.67	43.33	46.67	53.33
5 dB	56.00	53.33	33.33	46.67	46.67
0 dB	16.67	18.33	12.67	13.33	18.00
Average	49.25	47.92	34.42	39.17	44.08

To improve the performance of the face recognition system and since we want to compare the proposed technique with the appearance and shape based feature fusion method, we have to combine the appearance and shape based features. The concatenation procedure of two different features is shown in Figure 5. Initially raw 5000 dimension appearance based features and 176 dimension shape based features are extracted. The Principal Component Analysis method [53, 54] has been used to reduce the dimension of appearance and shape based features into 192 and 14, respectively. Two different features are added and produced 206 dimension features. Finally, PCA has been used again to resize from 206 dimensional to 130 dimensional appearance-shape based facial feature vector.

4.2. Experimental Results according to ORL Facial Database. Olivetti Research Laboratory (ORL) face database [55] produced by AT&T Laboratories has been used for measuring the performance of the proposed system. The database contains 10 different images of 40 distinct subjects. For some of the subjects, the images were taken at different times, varying lighting slightly, facial expressions (open/closed eyes, smiling/nonsmiling), and facial details (glasses/no-glasses).

Table 9: Overall Average Speaker Identification Rate (%) for NOIZEUS Speech Corpus.

Various Noises	Method				
	MFCC	ΔMFCC	ΔΔMFCC	RCC	LPCC
Airport Noise	45.42	48.25	34.42	35.75	42.08
Babble Noise	45.00	49.25	35.42	39.25	44.67
Car Noise	43.75	45.25	29.17	35.83	41.00
Exhibition Hall Noise	49.42	51.42	34.25	39.17	45.75
Restaurant Noise	46.83	49.67	32.58	38.75	43.09
Street Noise	46.00	49.09	35.42	39.25	42.42
Train Noise	47.08	49.92	33.92	39.25	42.67
Train Station Noise	49.25	47.92	34.42	39.17	44.08
Average Identification Rate (%)	46.59	48.85	33.70	38.30	43.22

All the images are taken against a dark homogeneous background and the subjects are in upright, frontal position (with tolerance for some side movement). The size of each face image is 92×112 and 8-bit grey levels. Experiment results are evaluated according to various dimensions such as optimum value selection of the number of hidden states of DHMM, response of the system based on noisy facial images and the system accuracy based on appearance, shape and combined appearance and shape based facial features.

4.2.1. System Response for Noisy Facial Images. The facial identification performance has been tested with the variations of different noises. Filtering is used for modifying or enhancing an image. To emphasize certain features or remove other features from an image, different filtering techniques are used. Filtering is a neighbourhood operation in which the value of any given pixel in the output image is determined by applying some algorithm to the values of the pixels in the neighbourhood of the corresponding input pixel. A pixel's neighbourhood is some set of pixels defined by their locations relative to that pixel. To remove the noise from the facial images, wiener filtering technique has been used. Wiener filtering technique has been used to remove or reduce white Gaussian noise from the facial image. Wiener filter can be used adaptively to an image where the variance is large, wiener filter performs little smoothing and where the variance is small, wiener filter performs more smoothing. Wiener filtering technique performs selective operation compared with other filters, preserving edges and other high-frequency parts of an image.

For measuring the accuracy of the face system, noise has been added in various rates for appearance based, shape based and appearance-shape based feature fusion technique with PCA based dimensionality reduction where Euclidian distance has been used as a classifier. Table 10 shows the response of applying Wiener filtering technique.

4.2.2. Performance Measurements between Single and Multiple Feature Fusion Based Techniques. Facial identification performance has been measured according to individual feature based technique such as appearance based feature, shape based feature and appearance-shape based feature fusion based technique. Receiver Operating Characteristics (ROC) curve is generated for the above mentioned techniques where a tradeoff is made between security and user friendness. The performance graph is shown in Figure 6. From the graph, it is shown that the appearance-shape based feature fusion can achieve compared with highest accuracy individual appearance based and shape based technique. For example, at a FRR = 30%, the appearance based, shape based and appearance-shape feature fusion FAR are 42%, 30%, and 28% respectively.

5. Audio and Visual Reliability Measurements

Since DHMM learning and testing models have been adopted for the audio and visual system, an ergodic discrete HMM (DHMM), θ_k [56], has been built in DHMM training phase for each face k. The model parameters (A, B, and θ) have been estimated to optimize the likelihood of the training set observation vector for the kth face by using the Baum-Welch algorithm. The Baum-Welch reestimation formula has been considered as follows [57, 58]:

$$\overline{\Pi}_i = \gamma_1(i), \qquad \overline{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)},$$
$$\overline{b}_j(\vec{k}) = \frac{\sum_{t=1(s,t,\vec{o}_t=\vec{v}_k)}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)}. \tag{13}$$

In the DHMM testing phase, for each unknown face to be recognized which includes

(i) measurement of the observation sequence, $O = \{o_1, o_2, \ldots, o_n\}$, via a feature analysis of the corresponding face,

(ii) transformation of the continuous values of O into integer values,

(iii) calculation of model likelihood for all possible models, $P(O \mid \theta_k)$, $1 \le k \le K$,

(iv) declaration of the face as k^* person whose model likelihood is highest—that is,

$$k^* = \arg\max_{1 \le k \le K} \left[P(O \mid \theta_k)\right]. \tag{14}$$

In this work, the probability computation step has been performed using Baum's Forward-Backward algorithm [58, 59]. By applying HMM as a learning phase, the log likelihood of each appearance and shape based feature of each person face have been captured. After getting the log likelihood of each modality separately, their outputs are combined by a weighted sum rule to produce the final decision. In this work, match score level is used to combine the appearance and shape based outputs. For a given appearance-shape test datum of O_A and O_S, the final recognition C^* is given by [60]

$$C^* = \arg\max_i \left\{\gamma \log P\left(\frac{O_A}{\lambda_A^i}\right) + (1-\gamma) \log P\left(\frac{O_S}{\lambda_S^i}\right)\right\}, \tag{15}$$

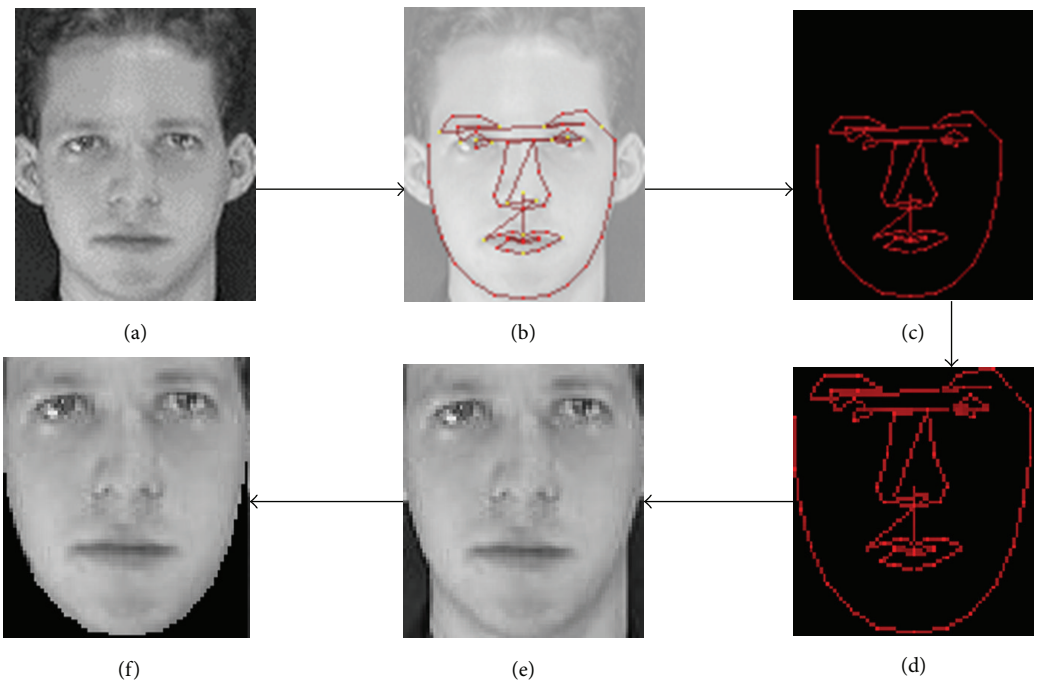


Figure 4: Facial image preprocessing for the proposed system (a) original image: (b) Output taken from Stams Active Shape Model (c) extracted facial edges (d) shape based features (e) region Of interest (ROI) selection with background noise, and (f) appearance based facial features.

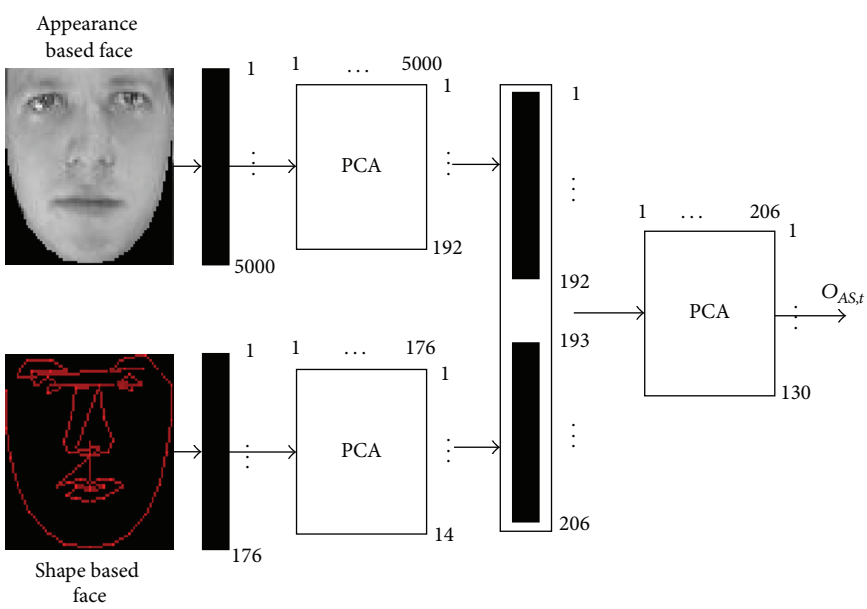


Figure 5: Process of appearance and shape based feature fusion.

where λ_A^i and λ_S^i are the appearance and the shape HMMs for the ith utterance class, respectively, and $\log P(O_A/\lambda_A^i)$ and $\log P(O_S/\lambda_S^i)$ are their log-likelihood against the ith class.

Among various types of score fusion techniques, baseline reliability ratio-based integration has been used to combine the appearance and shape recognition results. The reliability of each modality can be measured from the outputs of the corresponding HMMs. The reliability of each modality can be calculated by the most appropriate method which is the best in performance [61],

$$S_m = \frac{1}{N-1}\sum_{i=1}^{N}\left(\max_j \log P\left(\frac{O}{\lambda^j}\right) - \log P\left(\frac{O}{\lambda^i}\right)\right). 0 \quad (16)$$

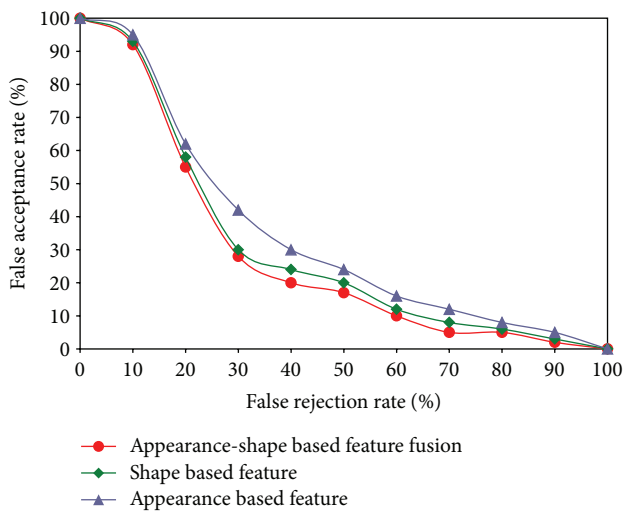


Figure 6: Performance comparison among individual modality and multimodal feature fusion techniques.

Which means the average difference between the maximum log-likelihood and the other ones and N is the number of classes being considered to measure the reliability of each modality, $m \in \{A, S\}$.

Then the integrated weight of appearance based reliability measure γ_A can be calculated by [62]

$$\gamma_A = \frac{S_A}{S_A + S_S}, \tag{17}$$

where S_A and S_S are the reliability measures of the outputs of the appearance and shape HMMs, respectively.

The integrated weight of shape modality measure can be found as

$$\gamma_S = (1 - \gamma_A). \tag{18}$$

6. BPN Based Likelihood Ratio Score Fusion

A Back-propagation learning feed-forward neural network [63] with tan-sigmoid transfer functions has been used in both the hidden layer and the output layer which is shown in Figure 7. Three-layer Back-propagation learning neural network algorithm has been used to classify the visual speech features [64].

If the input vector is $I = [p_1, p_2, \ldots, p_n]$, then the output of hidden layer has been calculated as follows:

$$n = IW + b,$$
$$a = f(n) = \frac{2}{\left(1 + e^{(-2*n)}\right)^{-1}}, \tag{19}$$

where, W is weight vector and b is bias input. The error is calculated as the difference between the target output and the network actual output. The goal is to minimize the average of the sum of these errors. Consider the following:

$$\text{mse} = \frac{1}{M}\sum_{k=1}^{M} e(k)^2 = \frac{1}{M}\sum (t(k) - a(k))^2. \tag{20}$$

Here, mse means mean square error, $t(k)$ represents the target output, and $a(k)$ represents the network output. The weights and bias values are updated based on the goal average error value.

In the proposed audio-visual system, the final weights and bias values are calculated in the training stage. In test phase, the output of the network has been calculated for the new input and compared with the target output to select the class of the input. The numbers of input layers, hidden layers and output layers nodes are 2, 100, and 8, respectively. The overall procedure for the proposed system with likelihood ratio based score fusion with Back-propagation learning neural network is shown in Figure 8.

The major drawbacks of Back-propagation learning neural network algorithm are the training time and local minima. Convergence time of the Back-propagation algorithm is inversely proportional to the error tolerance rate. In learning, effective use of error rate can decrease the convergence time. At first, select the final error rate. Then converge the weights such that all the patterns overcome some of the percentage error of the total system (the error must be higher than the final error rates). Finally, converge the system to the next lower error rate until crossing the final targeted error. For example, if the error rate of the system is 0.001, first the converged error rate for all of the patterns is 0.009, then 0.005, 0.003, and finally 0.001. This process is known as SET-BPL [65].

100 speech utterances are trained in Back-propagation learning neural network and the effects of applying SET-BPL of the proposed system areshown in Figure 9. Sometimes local minima problem occurs in a Back-propagation learning neural network algorithm. As a result, some precautions such as addition of internal nodes and lowering the gain term have been considered to set the learning parameters. The addition of internal nodes and lowering the gain term can increase the convergence time. To overcome these learning difficulties, a momentum term has been used to speed up the convergence process for this proposed speaker identification system.

Table 10: Results after Applying Wiener Filter.

Noise addition rate (in Variance)	Recognition rate based on PCA		
	Appearance based feature	Shape based feature	Appearance-shape based feature fusion
0.01	89%	91%	93%
0.05	88%	88%	90%
0.08	82%	85%	87%
0.1	80%	81%	82%
0.4	75%	77%	80%

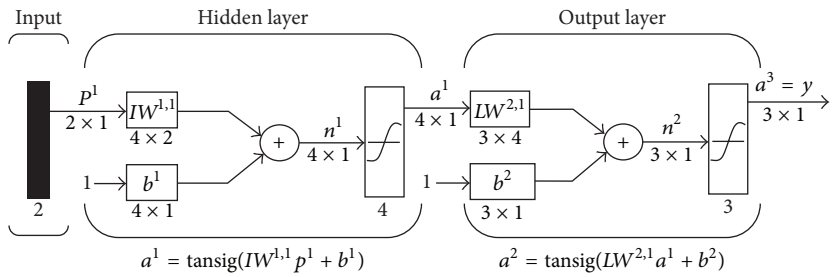


Figure 7: Architecture of Back-propagation neural network.

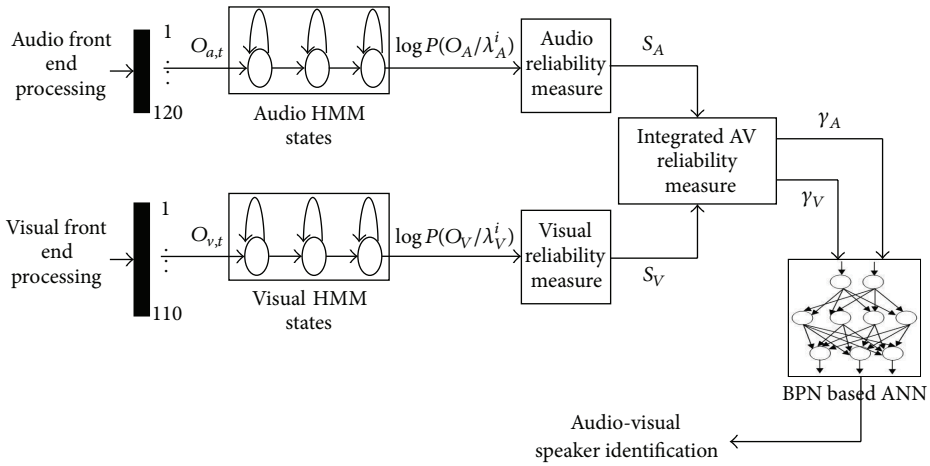
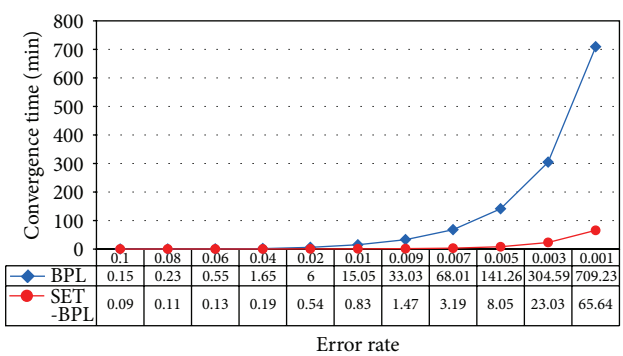


Figure 8: BPN based score fusion technique for proposed audio-visual speaker identification system.



	0.1	0.08	0.06	0.04	0.02	0.01	0.009	0.007	0.005	0.003	0.001
BPL	0.15	0.23	0.55	1.65	6	15.05	33.03	68.01	141.26	304.59	709.23
SET-BPL	0.09	0.11	0.13	0.19	0.54	0.83	1.47	3.19	8.05	23.03	65.64

Figure 9: Effects of applying SET-BPL process of the proposed system.

6.1. Experimental Evaluation according to VALID Audio-Video Database. VALID audio-visual multimodal database [66] has been used to measure and compare the accuracy between the proposed and existing system. For visual features, database contains 106 subjects each with four office lighting conditions, gathered periodically over one month giving some temporal variation and one studio session with controlled lighting. The 576×720 stills were extracted from the video segments. Three sets of these are offered, the 1st, 10th, and 50th frames for each of the 106×5 sessions. The five sessions were recorded over a period of one month, allowing for variation in the voice, clothing, facial hair, hairstyle and cosmetic appearance of the subjects and also variation of the visual background, illumination, and acoustic noise. The first session was recorded under controlled acoustic/illumination conditions, that is, controlled environment. The database is designed to be realistic and challenging; hence the other four sessions were recorded in noisy real-world scenarios with no control on illumination or acoustic noise that is, uncontrolled environment. Some processed facial images of VALID database are shown in Figure 10.

For the speech wave, the database contains 106 subjects with one studio and four office conditions recordings for each person corresponding to the facial images where the following two different speech utterances are found:

Uttered sentence 1:

"Joe Took Father's Green Shoe bench Out,"

Uttered sentence 2: "5 0 6 9 2 8 1 3 7 4."

From the above two sentences, second sentence has been used for learning and testing operations. Out of five facial images and speeches, the neutral face and corresponding speech utterance have been used for learning and other four official noisy images and speeches have used for testing.

6.2. Performance Analysis between Existing and Proposed Method. To evaluate the performance of BPN based likelihood ratio based score fusion technique, different variations

Figure 10: Some processed facial images of VALID database for the proposed system.

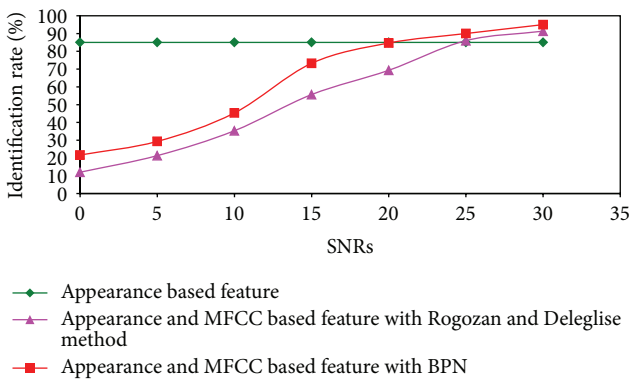


Figure 11: Results of score fusion for appearance based facial feature with MFCC based feature of audio feature.

of audio and visual features are combined and results are taken according to various SNRs of audio signal which are shown in the following subsections.

6.2.1. Experiment of Appearance Based Facial Feature with Audio Feature. Appearance based facial features are concatenated with MFCC based audio feature to populate the performance of the proposed score fusion based speaker identification. Results are shown in Figure 11 where the highest speaker identification rate has been found to be 95% at SNR of 30 dB for proposed BPN score fusion approach compared with existing Rogozan and Deléglise method of 91.33%.

6.2.2. Experiment of Shape Based Facial Feature with Audio Feature. Figure 12 shows the results of shape based facial feature and MFCC based audio feature. At SNR of 30 dB, the speaker identification rate of Rogozan and Deléglise method and proposed BPN score fusion approach has been achieved with 93.33% and 96.33%, respectively.

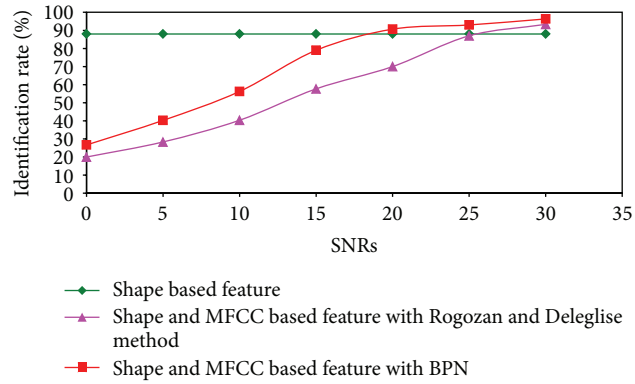


Figure 12: Shape based facial feature and MFCC based audio features for score fusion technique.

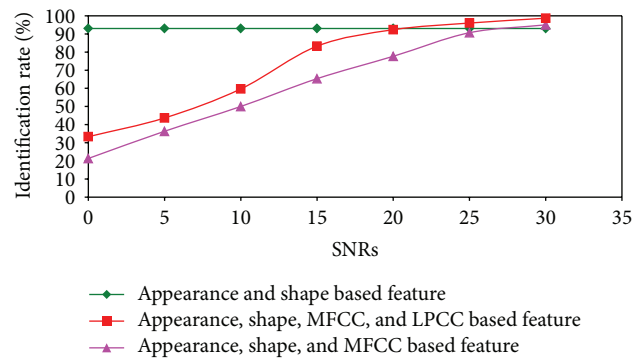


Figure 13: Result of the combinations of audio features with combined appearance and shape based facial feature for score fusion technique.

6.2.3. Experiment of Combined Appearance-Shape Based Facial Feature with Audio Feature. Results of appearance and shape based facial features with MFCC based audio feature for score fusion technique are shown in Figure 13. The highest speaker identification rate of 98.67% has been found at SNR of 30 dB with proposed BPN score fusion approach where existing Rogozan and Deléglise method achieves 95% at the same SNR.

Form the above experimental results, it has been shown that Back-propagation learning network based score fusion approach gives greater performance than any combination of audio and visual features compared with existing Rogozan and Deléglise method of score fusion. Here, it has also been focused on that combined appearance and shape based facial feature achieves higher accuracies than any individual facial feature based technique which is shown in Table 11.

7. Conclusions and Observations

In this work, proposed system performance has been evaluated in various levels with various dimensions. Two different types of facial features are combined with audio feature with various artificial noise addition rates. NOIZEUS speech database has been used to evaluate the performance of

Table 11: Performance Comparison of Various Combination of Facial Features with Audio Feature for Proposed BPN Approach.

SNRs	Type		
	Appearance and ΔMFCC based feature with BPN approach (in %)	Shape and ΔMFCC based feature with BPN approach (in %)	Appearance, shape and ΔMFCC based feature with BPN approach (in %)
0 dB	21.67	26.67	33.33
5 dB	29.33	40.25	43.67
10 dB	45.33	56.25	59.67
15 dB	73.25	79.00	83.25
20 dB	84.67	90.67	92.33
25 dB	90.00	93.00	96.00
30 dB	95.00	96.33	98.67

the Audio-Only speaker identification system whereas ORL facial database has been used for visual-only identification system. Finally, overall performance that is, audio-visual speaker identification has been measured according to VALID audio-visual database. Noise removing techniques have used to reduce or eliminate the noises from speech utterances and facial images. Experimental results and performance analysis shows the versatility of the proposed BPN score fusion approach over the existing Rogozan and Deléglise method for audio-visual speaker identification system which can be effectively used in various real life access control and authentication purposes.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] J. D. Woodward, "Biometrics: Privacy's foe or Privacy's friend?" *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1480–1492, 1997.

[2] A. K. Jain, R. Bolle, and S. Pankanti, "Introduction to biometrics," in *Biometrics, Personal Identification in Networked Society*, A. K. Jain, R. Bolle, and S. Pankanti, Eds., pp. 1–41, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1999.

[3] D. Bhattacharyya, R. Rahul, A. A. Farkhod, and C. Minkyu, "Biometric authentication: a review," *International Journal of U- and E- Service, Science and Technology*, vol. 2, no. 3, 2009.

[4] C. Sanderson and K. K. Paliwal, "Information fusion and person verification using speech and face information," Research Paper IDIAP-RR 02-33, IDIAP, 2002.

[5] D. G. Stork and M. E. Hennecke, *Speechreading by Humans and Machines*, Springer, Berlin, Germany, 1996.

[6] R. Campbell, B. Dodd, and D. Burnham, *Hearing by Eye II*, Psychology Press, Hove, UK, 1998.

[7] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Transactions on Multimedia*, vol. 2, no. 3, pp. 141–151, 2000.

[8] G. Potamianos, J. Luettin, and C. Neti, "Hierarchical discriminant features for audio-visual LVCSR," in *Proceedings of the IEEE Interntional Conference on Acoustics, Speech, and Signal Processing*, pp. 165–168, May 2001.

[9] G. Potamianos and C. Neti, "Automatic speechreading of impaired speech," in *Proceedings of the Conference on Audio-Visual Speech Processing*, pp. 177–182, 2001.

[10] F. J. Huang and T. Chen, "Consideration of lombard effect for speechreading," in *Proceedings of the IEEE 4th Workshop on Multimedia Signal Processing*, pp. 613–618, October 2001.

[11] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306–1325, 2003.

[12] D. A. Reynolds, "Experimental evaluation of features for robust speaker identification," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 639–643, 1994.

[13] S. Sharma, D. Ellis, S. Kajarekar, P. Jain, and H. Hermansky, "Feature extraction using non-linear transformation for robust speech recognition on the aurora database," in *Proceedings of the IEEE Interntional Conference on Acoustics, Speech, and Signal Processing (ICASSP '00)*, pp. 1117–1120, June 2000.

[14] D. Wu, A. C. Morris, and J. Koreman, "MLP internal representation as disciminant features for improved speaker recognition," in *Proceedings of the International Conference on Non-Linear Speech Processing (NOLISP '05)*, pp. 25–33, Barcelona, Spain, 2005.

[15] Y. Konig, L. Heck, M. Weintraub, and K. Sonmez, "Nonlinear discriminant feature extraction for robust text-independent speaker recognition," in *Proceedings of the RLA2C, ESCA Workshop on Speaker Recognition and Its Commercial and Forensic Applications*, pp. 72–75, 1998.

[16] C. C. Chibelushi, F. Deravi, and J. S. D. Mason, "A review of speech-based bimodal recognition," *IEEE Transactions on Multimedia*, vol. 4, no. 1, pp. 23–37, 2002.

[17] X. Zhang, C. C. Broun, R. M. Mersereau, and M. A. Clements, "Automatic speechreading with applications to human-computer interfaces," *EURASIP Journal on Applied Signal Processing*, vol. 2002, no. 11, pp. 1228–1247, 2002.

[18] D. N. Zotkin, R. Duraiswami, and L. S. Davis, "Joint audio-visual tracking using particle filters," *EURASIP Journal on Applied Signal Processing*, vol. 2002, no. 11, pp. 1154–1164, 2002.

[19] P. de Cuetos, C. Neti, and A. W. Senior, "Audio-visual intent-to-speak detection for human-computer interaction," in *Proceedings of the IEEE Interntional Conference on Acoustics, Speech, and Signal Processing*, pp. 2373–2376, June 2000.

[20] D. Sodoyer, J.-L. Schwartz, L. Girin, J. Klinkisch, and C. Jutten, "Separation of audio-visual speech sources: a new

approach exploiting the audio-visual coherence of speech stimuli," *EURASIP Journal on Applied Signal Processing*, vol. 2002, no. 11, pp. 1165–1173, 2002.

[21] E. Foucher, L. Girin, and G. Feng, "Audiovisual speech coder: using vector quantization to exploit the audio/video correlation," in *Proceedings of the Conference on Audio-Visual Speech Processing*, pp. 67–71, Terrigal, Australia, 1998.

[22] J. Huang, Z. Liu, Y. Wang, Y. Chen, and E. Wong, "Integration of multimodal features for video scene classification based on HMM," in *Proceedings of the Workshop on Multimedia Signal Processing*, pp. 53–58, Copenhagen, Denmark, 1999.

[23] M. M. Cohen and D. W. Massaro, "What can visual speech synthesis tell visual speech recognition?" in *Proceedings of the Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, Calif, USA, 1994.

[24] E. Cosatto and H. P. Graf, "Photo-realistic talking-heads from image samples," *IEEE Transactions on Multimedia*, vol. 2, no. 3, pp. 152–163, 2000.

[25] G. Potamianos, C. Neti, and S. Deligne, "Joint audio-visual speech processing for recognition and enhancement," in *Proceedings of the Auditory-Visual Speech Processing Tutorial and Research Workshop (AVSP '03)*, pp. 95–104, St. Jorioz, France, 2003.

[26] A. Rogozan and P. Deléglise, "Adaptive fusion of acoustic and visual sources for automatic speech recognition," *Speech Communication*, vol. 26, no. 1-2, pp. 149–161, 1998.

[27] A. Adjoudani and C. Benoit, "On the integration of auditory and visual parameters in an HMM-based ASR," in *Humans and Machines: Models, Systems, and Applications*, D. G. Stork and M. E. Hennecke, Eds., pp. 461–472, Springer, Berlin, Germany.

[28] T. W. Lewis and D. M. W. Powers, "Sensor fusion weighting measures in audio-visual speech recognition," in *Proceedings of the Conference on Australasian Computer Science*, pp. 305–314, Dunedine, New Zealand, 2004.

[29] G. Potamianos and C. Neti, "Stream confidence estimation for audio-visual speech recognition," in *Proceedings of the International Conference on Spoken Language Processing*, pp. 746–749, Beijing, China, 2000.

[30] I. Matthews, J. A. Bangham, and S. Cox, "Audiovisual speech recognition using multiscale nonlinear image decomposition," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP '96)*, pp. 38–41, October 1996.

[31] L. Lam and C. Y. Suen, "Application of majority voting to pattern recognition: an analysis of its behavior and performance," *IEEE Transactions on Systems, Man, and Cybernetics A*, vol. 27, no. 5, pp. 553–568, 1997.

[32] L. Lam and C. Y. Suen, "Optimal combinations of pattern classifiers," *Pattern Recognition Letters*, vol. 16, no. 9, pp. 945–954, 1995.

[33] L. Xu, A. Krzyzak, and C. Y. Suen, "Methods of combining multiple classifiers and their applications to handwriting recognition," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 22, no. 3, pp. 418–435, 1992.

[34] J. Daugman, "Biometric decision landscapes," Tech. Rep. TR482, University of Cambridge Computer Laboratory, 2000.

[35] S. Doclo and M. Moonen, "On the output SNR of the speech-distortion weighted multichannel Wiener filter," *IEEE Signal Processing Letters*, vol. 12, no. 12, pp. 809–811, 2005.

[36] R. Wang and W. Filtering, "PHYS, 3301, scientific computing," Project Report for NOISE Group, May 2000.

[37] K. Kitayama, M. Goto, K. Itou, and T. Kobayashi, "Speech starter: noise-robust endpoint detection by using filled pauses," in *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech '03)*, pp. 1237–1240, Geneva, Switzerland, September 2003.

[38] Q. Li, J. Zheng, A. Tsai, and Q. Zhou, "Robust endpoint detection and energy normalization for real-time speech and speaker recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 3, pp. 146–157, 2002.

[39] N. Wiener and R. E. A. C. Paley, *Fourier Transforms in the Complex Domains*, American Mathematical Society, Providence, RI, USA, 1934.

[40] J. W. Picone, "Signal modeling techniques in speech recognition," *Proceedings of the IEEE*, vol. 81, no. 9, pp. 1215–1247, 1993.

[41] L. P. Cordella, P. Foggia, C. Sansone, and M. Vento, "A real-time text-independent speaker identification system," in *Proceedings of the 12th International Conference on Image Analysis and Processing*, pp. 632–637, IEEE Computer Society Press, 2003.

[42] F. J. Harris, "On the use of windows for harmonic analysis with the discrete Fourier transform," *Proceedings of the IEEE*, vol. 66, no. 1, pp. 51–83, 1978.

[43] J. Proakis and D. Manolakis, *Digital Signl Processing, Principles, Algorithms and Applications*, Macmillan, New York, NY, USA, 2nd edition, 1992.

[44] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[45] T. F. Li and S.-C. Chang, "Speech recognition of mandarin syllables using both linear predict coding cepstra and Mel frequency cepstra," in *Proceedings of the 19th Conference on Computational Linguistics and Speech Processing*, pp. 379–390, 2007.

[46] Y. Hu and P. C. Loizou, "Subjective comparison of speech enhancement algorithms," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '06)*, pp. I153–I156, Toulouse, France, May 2006.

[47] Y. Hu and P. C. Loizou, "Evaluation of objective measures for speech enhancement," in *Proceedings of the 9th International Conference on Spoken Language Processing (INTERSPEECH '06)*, pp. 1447–1450, September 2006.

[48] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 1, pp. 229–238, 2008.

[49] S. Milborrow, *Locating facial features with active shape models [Masters dissertation]*, Faculty of Engineering, University of Cape Town, Cape Town, South Africa, 2007.

[50] R. Herpers, G. Verghese, K. Derpains, and R. McCready, "Detection and tracking of face in real environments," in *Proceedings of the International IEEE Workshop Recognition, Analysis and Tracking of Face and Gesture in Real- Time Systems*, pp. 96–104, Corfu, Greece, 1999.

[51] J. Daugman, "Face detection: a survey," *Computer Vision and Image Understanding*, vol. 83, no. 3, pp. 236–274, 2001.

[52] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, Addison-Wesley, 2002.

[53] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[54] T. Matthew and A. Pentland, *Face Recognition Using Eigenfaces*, Vision and Modeling Group, The Media Laboratory, Massachusetts Institute of Technology, 1991.

[55] F. S. Samaria and A. C. Harter, "Parameterisation of a stochastic model for human face identification," in *Proceedings of the 2nd IEEE Workshop on Applications of Computer Vision*, pp. 138–142, Orlando, Fla, USA, December 1994.

[56] R. O. Duda, P. E. Hart, and D. G. Strok, *Pattern Classification*, A Wiley-Interscience Publication, John Wiley & Sons, 2nd edition, 2001.

[57] V. Sarma and D. Venugopal, "Studies on pattern recognition approach to voiced-unvoiced-silence classification," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '78)*, vol. 3, pp. 1–4, 1978.

[58] L. R. Rabiner, "Tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[59] P. A. Devijver, "Baum's forward-backward algorithm revisited," *Pattern Recognition Letters*, vol. 3, no. 6, pp. 369–373, 1985.

[60] A. Rogozan and P. Deléglise, "Adaptive fusion of acoustic and visual sources for automatic speech recognition," *Speech Communication*, vol. 26, no. 1-2, pp. 149–161, 1998.

[61] J. S. Lee and C. H. Park, "Adaptive decision fusion for audio-visual speech recognition," in *Speech Recognition, Technologies and Applications*, F. Mihelic and J. Zibert, Eds., p. 550, 2008.

[62] A. Adjoudant and C. Benoit, "On the integratio of auditory and visual parameters in an HMM-based ASR," in *Humans and Machines: Models, Systems, and Speech Recognition, Technologies and Applications*, D. G. Strok and M. E. Hennecke, Eds., pp. 461–472, Springer, Berlin, Germany, 1996.

[63] J. A. Freeman and D. M. Skapura, *Neural Networks, Algorithms, Applications and Programming Techniques*, Addison-Wesley, 1991.

[64] W. C. Yau, D. K. Kumar, and S. P. Arjunan, "Visual recognition of speech consonants using facial movement features," *Integrated Computer-Aided Engineering*, vol. 14, no. 1, pp. 49–61, 2007.

[65] M. R. Islam and M. A. Sobhan, "Improving the convergence of backpropagation learning Neural Networks Based Bangla Speaker Identification System for various weight update frequencies, momentum term and error rate," in *Proceedings of the International Conference on Computer Processing of Bangla (ICCPB '06)*, pp. 27–34, Independent University, 2006.

[66] A. F. Niall, A. O. Brian, and B. R. Richard, "VALID: a new practical audio-visual database, and comparative results," in *Audio- and Video-Based Biometric Person Authentication*, vol. 3546 of *Lecture Notes in Computer Science*, pp. 201–243, 2005.