

Research Article

Feature Fusion Based Audio-Visual Speaker Identification Using Hidden Markov Model under Different Lighting Variations

Md. Rabiul Islam¹ and Md. Abdus Sobhan²

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 31 August 2013; Revised 14 January 2014; Accepted 16 January 2014; Published 5 March 2014

Academic Editor: T. Warren Liao

Copyright © 2014 Md. R. Islam and Md. A. Sobhan. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The aim of the paper is to propose a feature fusion based Audio-Visual Speaker Identification (AVSI) system with varied conditions of illumination environments. Among the different fusion strategies, feature level fusion has been used for the proposed AVSI system where Hidden Markov Model (HMM) is used for learning and classification. Since the feature set contains richer information about the raw biometric data than any other levels, integration at feature level is expected to provide better authentication results. In this paper, both Mel Frequency Cepstral Coefficients (MFCCs) and Linear Prediction Cepstral Coefficients (LPCCs) are combined to get the audio feature vectors and Active Shape Model (ASM) based appearance and shape facial features are concatenated to take the visual feature vectors. These combined audio and visual features are used for the feature-fusion. To reduce the dimension of the audio and visual feature vectors, Principal Component Analysis (PCA) method is used. The VALID audio-visual database is used to measure the performance of the proposed system where four different illumination levels of lighting conditions are considered. Experimental results focus on the significance of the proposed audio-visual speaker identification system with various combinations of audio and visual features.

1. Introduction

Human speaker identification is bimodal in nature [1, 2]. In a face-to-face conversation, we listen to what others say and at the same time observe their lip movements, facial expressions, and gestures. Especially, if we have a problem in listening due to environmental noise, the visual information plays an important role for speech understanding [3]. Even in the clean environment, speech recognition performance is improved when the talking face is visible [4]. Generally, it is true that audio-only speaker identification system is not sufficiently adequate to meet the variety of user requirements for person identification. The AVSI system promises to alleviate some of the drawbacks encountered by audio-only identification. Visual speech information can play an important role in the improvement of natural and robust human-computer interaction [5, 6]. Indeed, various important human-computer components, such as speaker

identification, verification [7], localization [8], speech event detection [9], speech signal separation [10], coding [11], video indexing and retrieval [12], and text-to-speech [13], have been shown to benefit from the visual channel [14].

Audio-visual identification system can significantly improve the performance of a biometric system besides improving population coverage, deterring spoof attacks, increasing the degrees of freedom, and reducing the failure-to-enroll rate. Although the storage requirements, processing time, and the computational demands of audio-visual system are much higher than audio-only system for speaker identification, effective integration of audio and visual features can remove or reduce most of the mentioned problems. Fusion of audio and visual features is an important fusion strategy which can improve system performance of AVSI system. In feature level fusion, sufficient information can exist compared to match score, rank, and decision level fusions [15]. As a result, it can be expected that feature level

fusion can achieve greater performance over other fusion strategies which have been incorporated in the proposed AVSI system.

Most published works in the areas of AVSI system focus on decision level fusion strategy for noiseless environments [16–19] and very few research works introduce noisy environmental conditions [20, 21]. The aim of this work is to use feature fusion for the AVSI system under different illumination levels of lighting conditions. Different audio and visual features are combined in multiple levels in the proposed system. The subsequent sections of the paper focus on the proposed block diagram, feature extraction of the speech and facial features, fusion of multimodal audio and visual feature vectors, dimensionality reduction of multiple features, classification by using HMM, and performance analysis of the proposed AVSI system.

2. Paradigm of the Proposed Audio-Visual Speaker Identification System

The architecture of the proposed audio-visual feature fusion based speaker identification system is shown in Figure 1. The MFCC and LPCC based audio features and Appearance and Shape based facial features are concatenated separately, and finally the audio and visual based features are fused to feed the HMM classifier. Normalization technique has been applied at different times to normalize the features. PCA has been used to reduce the dimension of the speech and facial feature vector in such a way that the principal components of the original features retain their attributes sufficiently.

3. Audio Feature Extraction and Fusion

To capture the speech signal, sampling frequency of 11025 Hz, sampling resolution of 16 bits, monorecording channel, and recorded file format = *.wav have been considered. Wiener filter has been used to remove the background noise from the original speech utterances [22, 23]. 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 a background noise [24, 25]. To detect word boundary, the frame energy is computed using the sort-term log energy equation [23]:

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

Preemphasis has been used to balance the spectrum of voiced sounds that have a steep roll-off in the high frequency region. The transfer function of the FIR filter in the z -domain is [26]

$$H(z) = 1 - \alpha z^{-1}, \quad 0 \leq \alpha \leq 1, \quad (2)$$

where α is the preemphasis parameter.

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 frames [27].

From different types of windowing techniques, Hamming window has been used for this system. The purpose of using windowing is to reduce the effect of the spectral artifacts that result from the framing process [28]. The Hamming window can be defined as follows:

$$w(n) = \begin{cases} 0.54 + 0.46 \cos \left[\frac{2\pi}{N} n \right], & n = -\frac{N}{2}, \dots, -1, 0, 1, \dots, \frac{N}{2} \\ 0.54 - 0.46 \cos \left[\frac{2\pi}{N} n \right], & n = 0, 1, 2, \dots, N-1. \end{cases} \quad (3)$$

To extract the features from the speech utterances, various types of standard speech feature extraction techniques [26] such as RCC, MFCC, Δ MFCC, $\Delta\Delta$ MFCC, LPC, and LPCC have been applied. After computation of the speech features, 12 frames are extracted where each frame contains 36 features. To normalize the feature value of each frame, feature mean normalization method has been applied. In this way, $(12 \times 36) = 432$ dimensional features are extracted from each speech utterance. Finally, 432 dimensional MFCC and 432 dimensional LPCC based speech features are extracted from the speech utterances. PCA based dimension reduction has been performed where 432 dimensional feature vectors have been converted into 120 dimension ones. Lastly, MFCC and LPCC based feature vectors are concatenated to produce 240 dimensional MFCC-LPCC based audio features.

4. Visual Feature Extraction and Fusion

High-quality digital camera has been used to capture the face images. After acquisition of face image, Stams [29] Active Appearance 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 [30, 31]. Lastly, the background noise has been eliminated [32] and finally appearance based facial features have been found. The procedure of the facial image preprocessing parts is shown in Figure 2. To reduce the dimensionality of the facial feature vector, PCA has been used.

5000 dimensional appearance based and 176 dimensional shape based features have been computed by the above mentioned process. Raw facial features, that is, 5000 dimensional appearance feature vectors and 176 dimensional shape feature vectors have been converted into 192 and 14 dimension ones, respectively. Before applying PCA, min-max normalization technique has been used using the following equation:

$$x' = \frac{x - \min(F_x)}{\max(F_x) - \min(F_x)}, \quad (4)$$

where F_x is the function which generates x .

The min-max technique is effective when the minimum and the maximum values of the component feature values

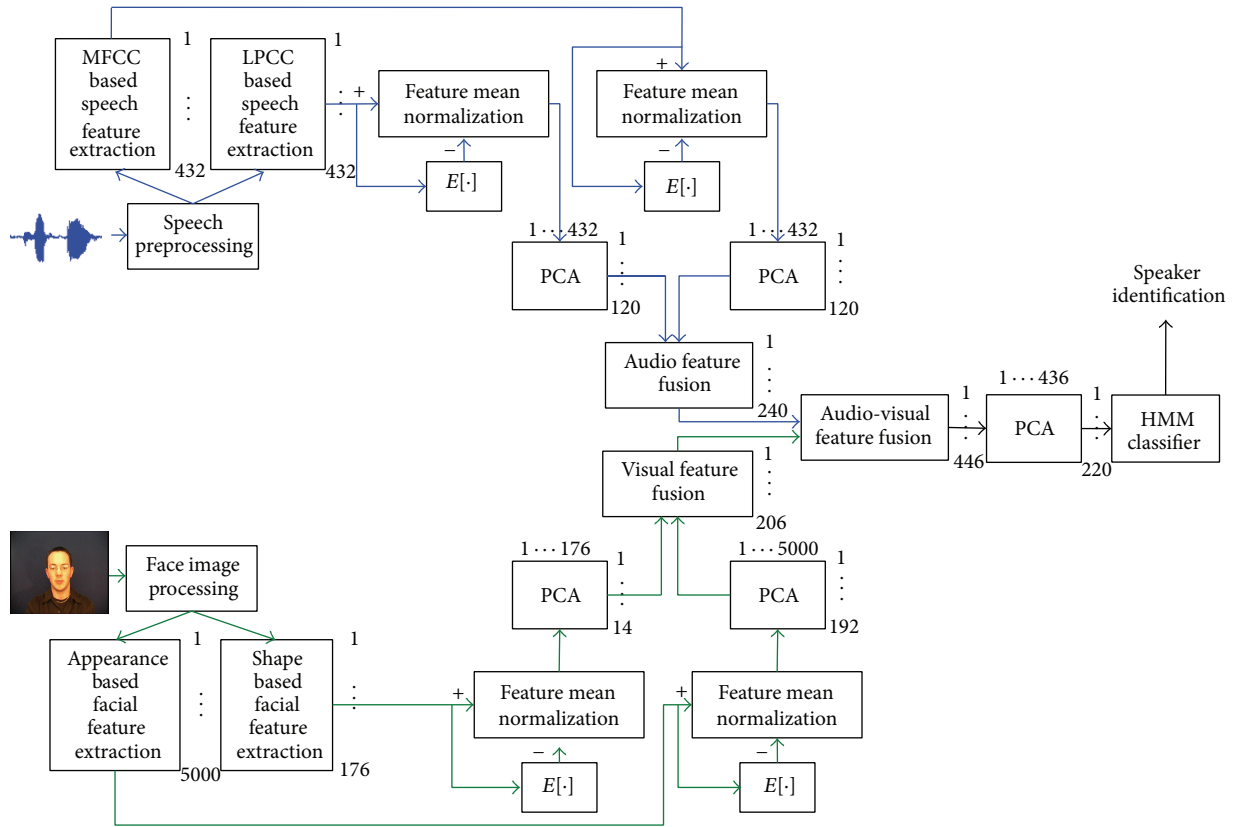


FIGURE 1: Paradigm of the proposed feature fusion based audio-visual speaker identification system.

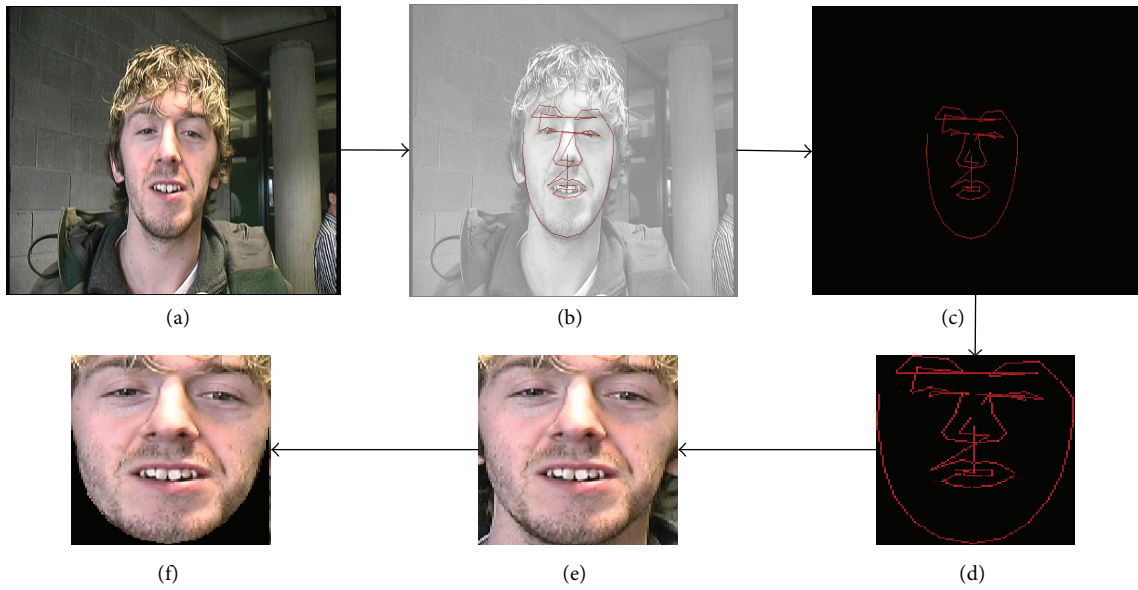


FIGURE 2: Facial image before processing for the proposed systems: (a) original image, (b) output taken from Stams Active Appearance Model, (c) extracted facial edges, (d) shape based features, (e) region of interest (ROI) selection with background noise, and (f) appearance based facial features.

are known previously. In cases where such information is not available, an estimate of these parameters has to be obtained from the available sample training data [33]. Finally, 192 dimensional appearances and 14 dimensional shape features are concatenated which is mentioned as visual feature fusion.

5. Audio-Visual Feature Fusion and HMM Classification

240 dimensional audio feature vectors and 206 dimensional visual feature vectors are fused to produce 446 dimensional audio-visual feature vectors. Since the dimension of audio-visual feature vector is large enough, PCA has been used to reduce the dimension to 220. The reduced audio-visual features are finally fed to HMM learning and classification model.

In HMM training phase, for each face k , an ergodic discrete HMM (DHMM), θ_k has been built [34]. The model parameters (A, B, θ) have been estimated to optimize the likelihood of the training set observation vectors for the k th face by using Baum-Welch algorithm. The Baum-Welch reestimation formula has been considered as follows [34, 35]:

$$\tilde{\pi}_i = \gamma_1(i), \quad (5)$$

$$\tilde{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i, j)}{\sum_{t=1}^{T-1} \gamma_t(i)}, \quad (6)$$

$$\tilde{b}_j(\vec{k}) = \frac{\sum_{t=1}^T \gamma_t(j)}{\sum_{t=1}^T \gamma_t(j)}, \quad (7)$$

where

$$\xi_t(i, j) = \frac{\alpha_t(i) a_{ij} b_j(\tilde{o}_{t+1}) \beta_{t+1}(j)}{\sum_{i=1}^N \sum_{j=1}^N \alpha_t(i) a_{ij} b_j(\tilde{o}_{t+1}) \beta_{t+1}(j)}, \quad (8)$$

$$\gamma_t(i) = \sum_{j=1}^N \xi_t(i, j).$$

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

- (i) measurement of the observation sequence, $O = \{o_1, o_2, \dots, o_n\}$, via a feature analysis of the speech corresponding to a face,
- (ii) transformation of the continuous values of O into integer values,
- (iii) calculation of model likelihoods for all possible models, $P(O | \theta_k)$, $1 \leq k \leq K$,
- (iv) declaration of the face as k^* person whose model likelihood is highest; that is,

$$k^* = \arg \max_{1 \leq k \leq K} [P(O | \theta_k)]. \quad (9)$$

In this proposed work, the probability computation step has been performed using Baum's Forward-Backward algorithm [35].

6. Performance Analysis of the Proposed System

VALID audio-visual database [36] has been used to measure the performance of the proposed speaker identification system. Artificial white Gaussian noise has added to the original clean speech utterances to simulate various SNR levels. The models have been trained at clean speech utterances and tested with utterances under SNR level ranging from 0 to 30 dB at an interval of 5 dB. VALID database contains one neutral and four different office environmental noisy speeches of each person. Out of the five speech utterances, one clean speech is used for learning and the four others are used for testing purpose where noises are artificially added from 0 to 30 dB an interval of 5 dB.

Experimental results of the audio-visual feature fusion based identification have been captured in various dimensions. Since MFCC and LPCC based audio features and appearance and shape based visual features are considered for the audio-visual feature fusion based speaker identification system, experimental results are examined according to various combinations of MFCC, LPCC, and MFCC-LPCC based features of audio modality and appearance, shape, and appearance-shape based features of visual modality.

6.1. Optimum Value Selection on the Number of Hidden States of DHMM, N_H . Since the number of hidden states of DHMM affects the performance of the audio-visual feature fusion based speaker identification system, results are taken according to different combinations of the audio and visual features which are shown in Figure 3. In Figure 3(a), optimum value of the number of hidden states is shown according to appearance based facial features with the combination of MFCC, LPCC, and both MFCC and LPCC based audio features separately. The highest speaker identification rate of 94% has been found at 15 numbers of hidden states when using appearance, MFCC, and LPCC based features. The identification rate of shape based facial features with MFCC, LPCC, and combined MFCC and LPCC based audio features are shown in Figure 3(b) according to the number of hidden states. When the number of hidden states is 13, the highest identification rate of 95% has been found. Figure 3(c) shows the results after setting up the number of hidden states of different combination of appearance and shape based features with MFCC and LPCC.

6.2. Performance Analysis of Different Variations of Audio-Visual Feature Fusion. To measure the performance of the audio-visual feature fusion based speaker identification system, different experiments have been performed, and we measured the performance from different angles which are shown in the following subsections.

6.2.1. Experiment of Individual Feature Based Performance. Individual performances of each of the feature extraction techniques are examined, which are shown in Table 1. Results are shown according to different audio SNRs rates where the visual feature is not affected by the audio noise. From the

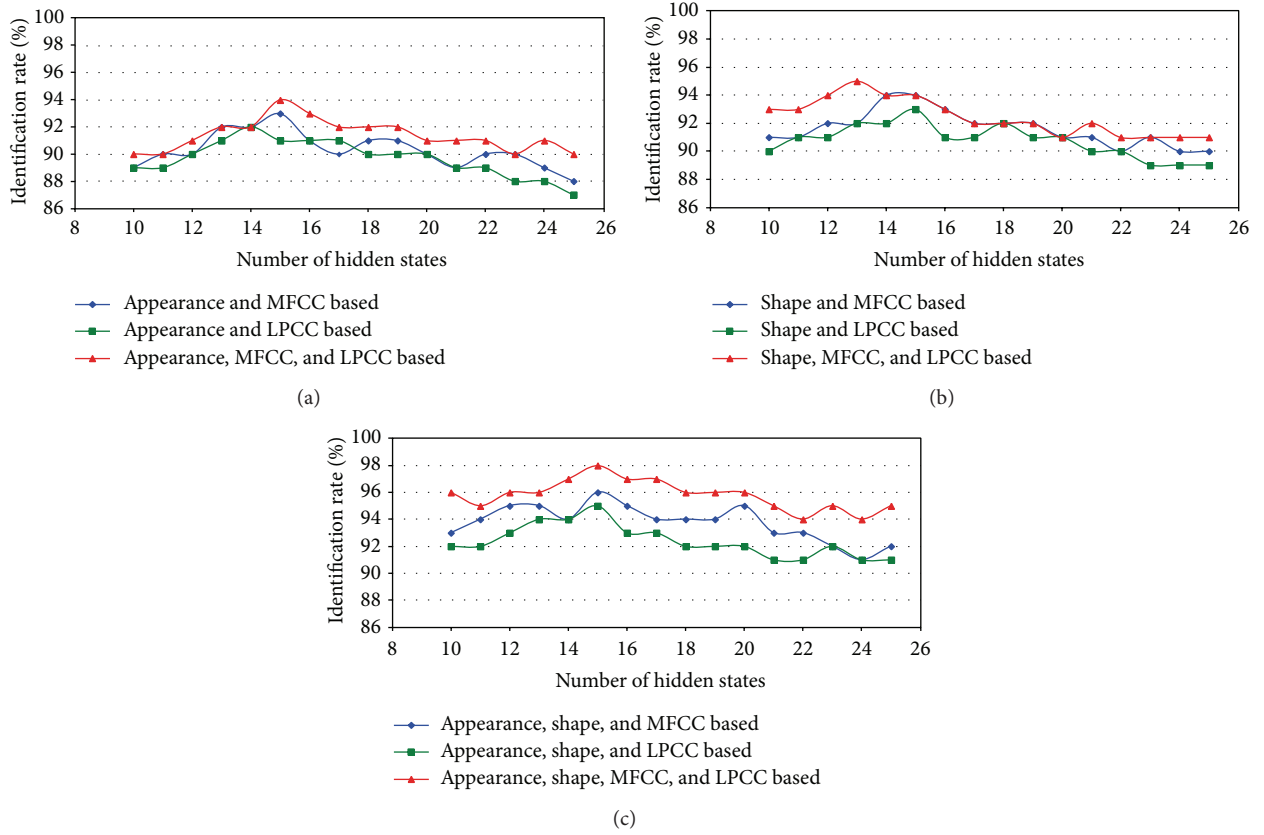


FIGURE 3: Results after setting up the number of hidden states of DHMM with various dimensions.

TABLE 1: Performance measurements among various combinations of appearance based, shape based, and appearance-shape based visual features and MFCC based, LPCC based, and MFCC-LPCC based audio features according to various SNRs.

SNR (dB)	Appearance based feature (%)	Shape based feature (%)	Appearance-shape based feature fusion (%)	MFCC based feature (%)	LPCC based feature (%)	MFCC-LPCC based feature fusion (%)
0	85.00	88.00	93.00	5.00	4.00	8.33
5	85.00	88.00	93.00	10.67	8.33	13.33
10	85.00	88.00	93.00	21.33	18.67	28.33
15	85.00	88.00	93.00	35.33	30.00	50.00
20	85.00	88.00	93.00	50.67	43.33	72.33
25	85.00	88.00	93.00	70.00	63.67	80.00
30	85.00	88.00	93.00	90.00	86.33	93.67
Average (%) (0 dB~30 dB)	85.00	88.00	93.00	40.42	36.33	49.42

table, it is shown that combined appearance and shape based facial feature can give greater identification rate than individual visual features. But the shape based features have better identification than appearance based for face identification. On the other hand, for audio identification performance, rate is varied according to different SNRs. Identification rate increases with the decrease of SNRs rate. At 0 dB noise, the identification rate is lower and the rate is higher at 30 dB audio noise. From the result, it has been shown that the MFCC based result is better than LPCC based feature, and

combined MFCC and LPCC based features can give much better performance than MFCC only and LPCC only features.

6.2.2. Experiment on Different Audio Features with Appearance Based Facial Features. Various audio features such as MFCC, LPCC, and combined MFCC and LPCC features are concatenated with the appearance based facial feature. HMM learning and classification techniques are used to measure the performance of the system. The results are examined according to various SNRs ranging from 0 to 30 dB which

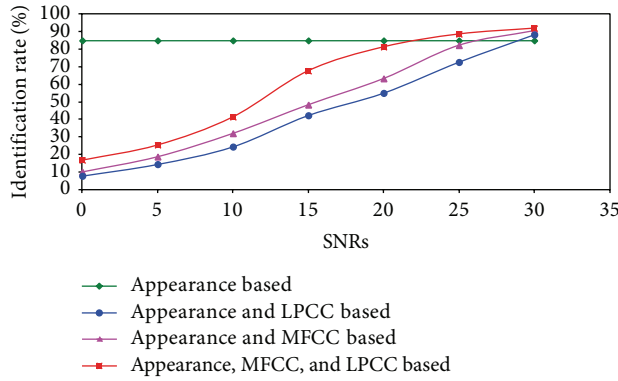


FIGURE 4: Result of various combinations of audio features with appearance based facial feature.

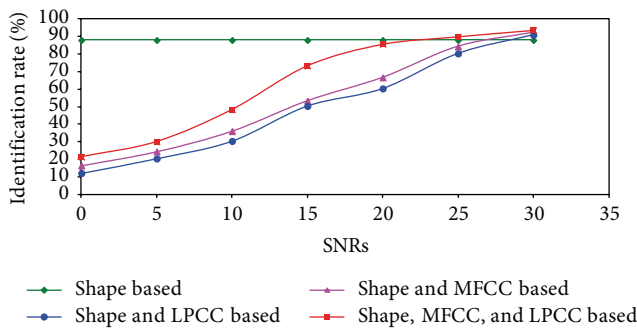


FIGURE 5: Result of various combinations of audio features with shape based facial feature.

are shown in Figure 4. The highest speaker identification rate of 92% has been found at appearance based features with combined MFCC and LPCC based audio features according to SNR rate of 30 dB.

6.2.3. Experiment on Different Audio Features with Shape Based Facial Features. Performance measurements between different audio features with shape based facial features are shown in Figure 5. It shows that the combination of MFCC and LPCC based audio features can give best result with shape based visual feature. The highest identification rate was found to be 93.67% at signal-to-noise ratio of 30 dB.

6.2.4. Experiment on Audio Features with Combined Appearance and Shape Based Facial Feature. Finally, the appearance and shape based facial features are combined with the MFCC and LPCC based audio features and the corresponding measurements of the performances are shown in Figure 6. In this experiment, we observe that the appearance and shape based combined facial features give constant identification rate, and it is 93%. The identification rate of other combinations of audio and visual features has different identification rates with the variation of various SNRs. In all of the cases, the audio-visual feature fusion based identification rate increases with the decreases of audio SNRs. At SNR 30 dB, the identification rate was 93.67%, 92%, and 95% when applying MFCC,

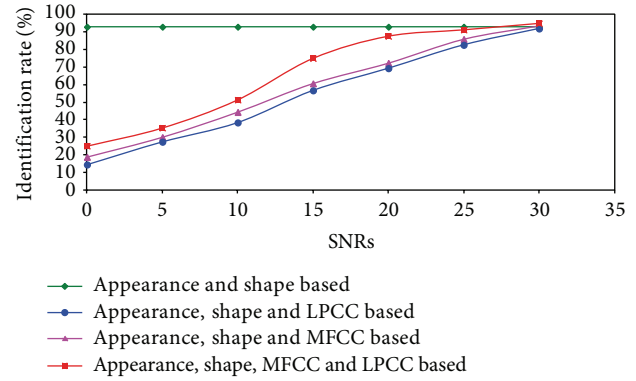


FIGURE 6: Result of various combinations of audio features with combined appearance and shape based facial feature.

LPCC, and MFCC-LPCC with combined appearance and shape based facial feature, respectively.

7. Conclusions and Observations

Experimental results and performance analysis reveal that the proposed audio-visual feature fusion based speaker identification system can perform better compared to any other single feature based approach. Four different varied light environments are considered for facial image. Artificial white Gaussian audio noises of 0 to 30 dB at an interval of 5 dB are added to the clean speech signals. The proposed system is capable enough to work at different variations of lighting environments. But the system performance may degrade at other noisy environment, which may be covered in a future research paper of the authors. In that case, the proposed feature-fusion approach may be applied before the decision-fusion approach which may further increase the efficiency of the audio-visual speaker identification rate.

Conflict of Interests

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

References

- [1] D. G. Stork and M. E. Hennecke, Eds., *Speechreading by Humans and Machines*, Springer, Berlin, Germany, 1996.
- [2] R. Campbell, B. Dodd, and D. Burnham, Eds., *Hearing by Eye II*, Psychology Press, Hove, UK, 1998.
- [3] L. A. Ross, D. Saint-Amour, V. M. Leavitt, D. C. Javitt, and J. J. Foxe, "Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments," *Cerebral Cortex*, vol. 17, no. 5, pp. 1147–1153, 2007.
- [4] P. Arnold and F. Hill, "Bisensory augmentation: a speechreading advantage when speech is clearly audible and intact," *British Journal of Psychology*, vol. 92, no. 2, pp. 339–355, 2001.
- [5] S. Dupont and J. Luetin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Transactions on Multimedia*, vol. 2, no. 3, pp. 141–151, 2000.

- [6] G. Potamianos, J. Luettin, and C. Neti, "Hierarchical discriminant features for audio-visual LVCSR," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 165–168, May 2001.
- [7] 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.
- [8] 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.
- [9] P. de Cuetos, C. Neti, and A. W. Senior, "Audio-visual intent-to-speak detection for human-computer interaction," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 2373–2376, Istanbul, Turkey, June 2000.
- [10] 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.
- [11] 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, December 1998.
- [12] 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 IEEE 3rd Workshop on Multimedia Signal Processing*, pp. 53–58, Copenhagen, Denmark, September 1999.
- [13] 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.
- [14] 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, Saint-Jorioz, France, September 2003.
- [15] A. Ross and R. Govindarajan, "Feature level fusion using hand and face biometrics," in *Biometric Technology for Human Identification II*, Proceedings of SPIE, pp. 196–204, Orlando, Fla, USA, March 2005.
- [16] 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.
- [17] 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. 1306–1326, 2003.
- [18] J.-S. Lee and C. H. Park, *Adaptive Decision Fusion for Audio-Visual Speech Recognition*. Speech Recognition, 2008.
- [19] K. Nandakumar, Y. Chen, S. C. Dass, and A. K. Jain, "Likelihood ratio-based biometric score fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 342–347, 2008.
- [20] Md. Rabiul Islam and Md. Fayzur Rahman, "Likelihood ratio based score fusion for audio-visual speaker identification in challenging environment," *International Journal of Computer Applications*, vol. 6, no. 7, pp. 6–11, 2010.
- [21] L. Girin, G. Feng, and J. L. Schwartz, "Fusion of auditory and visual information for noisy speech enhancement: a preliminary study of vowel transitions," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '98)*, pp. 1005–1008, 1998.
- [22] 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.
- [23] N. Wiener and R. E. A. C. Paley, *Fourier Transforms in the Complex Domains*, American Mathematical Society, Providence, RI, USA, 1934.
- [24] K. Kitayama, M. Goto, K. Itou, and T. Kobayashi, "Speech starter: noise-robust endpoint detection by using filled pauses," in *Proceedings of the Eurospeech*, pp. 1237–1240, Geneva, Switzerland, 2003.
- [25] S. E. Bou-Ghazale and K. Assaleh, "A robust endpoint detection of speech for noisy environments with application to automatic speech recognition," in *Proceedings of the IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP '02)*, vol. 4, pp. 3808–3811, May 2002.
- [26] J. W. Picone, "Signal modeling techniques in speech recognition," *Proceedings of the IEEE*, vol. 81, no. 9, pp. 1215–1247, 1993.
- [27] 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, Mantova, Italy, September 2003.
- [28] 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.
- [29] S. Milborrow, *Locating facial features with active shape models [dissertation]*, Faculty of Engineering, University of Cape Town, Cape Town, South Africa, 2007.
- [30] R. Herpers, G. Verghese, K. Derpains, and R. McCready, "Detection and tracking of face in real environments," in *Proceedings of the IEEE International Workshop on Recognition, Analysis and Tracking of Face and Gesture in Real-Time Systems*, pp. 96–104, Corfu, Greece, 1999.
- [31] J. Daugman, "Face detection: a survey," *Computer Vision and Image Understanding*, vol. 83, no. 3, pp. 236–274, 2001.
- [32] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, Addison-Wesley, 2002.
- [33] A. Ross and R. Govindarajan, "Feature level fusion using hand and face biometrics," in *Biometric Technology for Human Identification II*, vol. 5779 of *Proceedings of SPIE*, pp. 196–204, Orlando, Fla, USA, March 2005.
- [34] J.-S. Lee and C. H. Park, *Speech Recognition, Technologies and Applications*, I-Tech, Vienna, Austria, 2008.
- [35] P. A. Devijver, "Baum's forward-backward algorithm revisited," *Pattern Recognition Letters*, vol. 3, no. 6, pp. 369–373, 1985.
- [36] N. A. Fox, B. A. O'Mullane, and R. B. Reilly, "VALID: a new practical audio-visual database, and comparative results," in *Audio - and Video-Based Biometric Person Authentication*, Lecture Notes in Computer Science, pp. 777–786, 2005.

