Hybrid Feature and Decision Fusion based Audio-Visual Speaker Identification in Challenging Environment

Md. Rabiul Islam
Assistant Professor
Department of Computer Science & Engineering
Rajshahi University of Engineering & Technology
Rajshahi-6204, Bangladesh.

Md. Fayzur Rahman
Professor
Department of Electrical & Electronic Engineering
Rajshahi University of Engineering & Technology
Rajshahi-6204, Bangladesh.

ABSTRACT
The contribution of this paper is to propose a novel approach of evaluating the performance of a noise robust audio-visual speaker identification system in challenging environment. Though the traditional HMM based audio-visual speaker identification system is very sensitive to the speech parameter variation, the proposed hybrid feature and decision fusion based audio-visual speaker identification is found to be stance and performs well for improving the robustness and naturalness of human-computer-interaction. Linear Prediction Cepstral Coefficients and Mel Frequency Cepstral Coefficients are used to extract the audio features and Active Appearance Model and Active Shape Model have been used to extract the appearance and shape based features for the facial image. Principal Component Analysis method has been used to reduce the dimensionality of large feature vector and to normalize, the vector normalization algorithm has been used. Features and decision both are fused in two different levels and finally four different classifier outputs are combined in parallel fashion to achieve the identification result. The performances of all these uni-modal and multi-modal system performance have been evaluated and compared with each other on VALID audio-visual multi-modal database, containing both vocal and visual biometric modalities.

General Terms
Hybrid Feature and Decision Fusion Based Speaker Identification, Human Computer Interaction, Biometrics.

Keywords
Hybrid Feature and Decision Fusion, Audio-Visual Speaker Identification, Cepstral Base Audio Features, Appearance and Shape Based Facial Features, Likelihood Ratio Based Score Fusion, Discrete Hidden Markov Model.

1. INTRODUCTION
Human speaker identification is bimodal in nature [1, 2]. Visual speech information can play a vital role for the improvement of natural and robust human-computer interaction [3, 4]. 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 [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].

Hybrid feature and decision fusion based audio-visual speaker identification system is proposed in this paper. RCC, LPCC, MFCC, ΔMFCC, ΔΔMFCC based audio feature extraction methods and for facial image, appearance and shape based feature extraction techniques have been applied to enhance the performance of the proposed scheme.

2. HYBRID FEATURE AND DECISION FUSION BASED AUDIO-VISUAL SPEAKER IDENTIFICATION MODEL
The block diagram for the proposed hybrid feature and decision fusion based audio-visual speaker identification system is shown in figure 1. MFCC and LPCC based audio features are extracted from the speech utterance and audio feature fusion is performed. Visual feature fusion is performed on appearance and shape based facial features. The audio feature fusion and visual feature fusion are fused again into audio-visual feature fusion method. On the decision fusion, audio reliability and visual reliability are measured according to the audio and visual HMM classifier and audio-visual likelihood ratio based score fusion are performed. Finally four separate identification output i.e. audio HMM classifier output, visual HMM classifier output, audio-visual feature fusion based classifier output and audio-visual likelihood ratio based score fusion are combined into parallel fashion by using majority vote decision fusion method to achieve the speaker identification result. Figure 2 shows the details working procedure of the proposed system.

3. AUDIO IDENTIFICATION
Sampling frequency of 11025 $H_{Z,}$ sampling resolution of 16-bits, mono recording channel and recorded file format of *.wav have been considered to capture the speech utterances. The speech preprocessing part has a vital role for the efficiency of learning. After acquisition of speech utterances, winner filter has been used to remove the background noise from the original speech utterances [15, 16]. 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 [17, 18]. To detect word boundary, the frame energy is computed using the sort-term log energy equation [18],

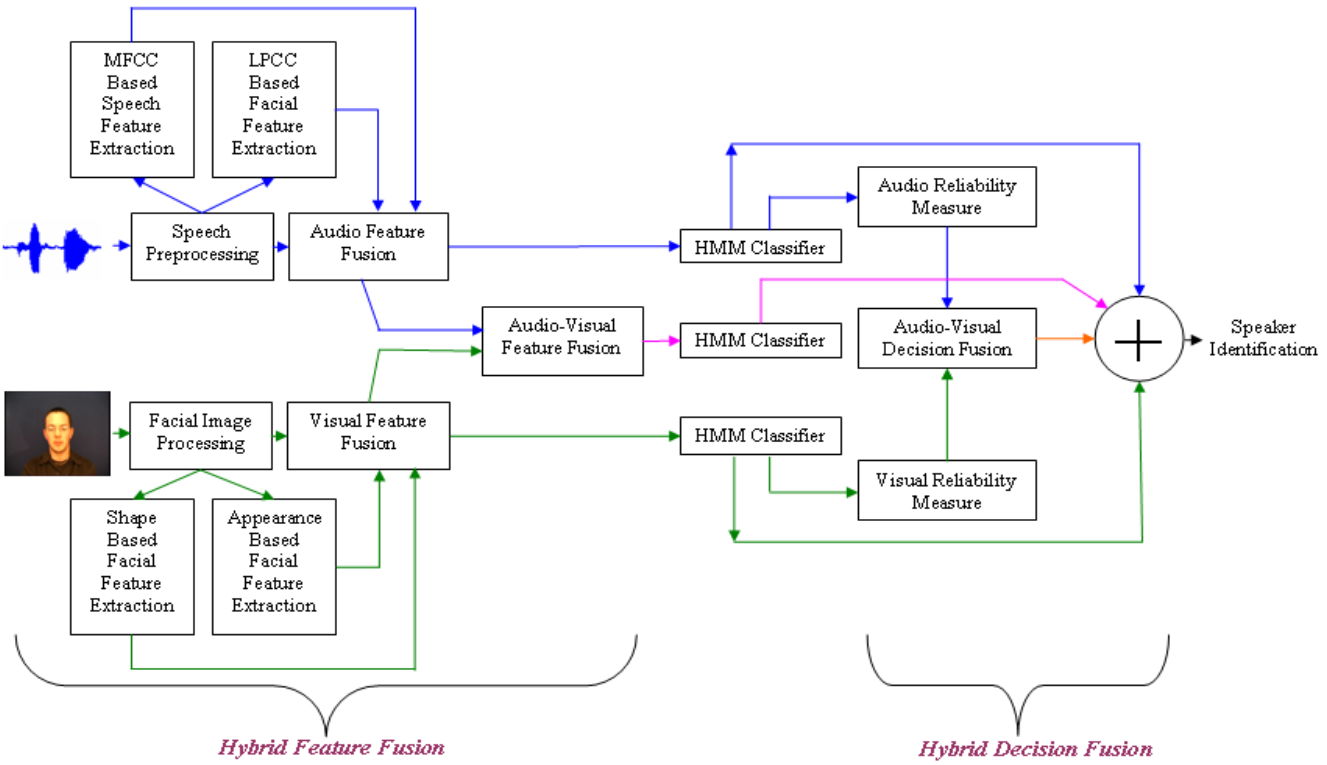

Figure 1: Paradigm of the proposed hybrid feature and decision fusion based audio-visual speaker identification

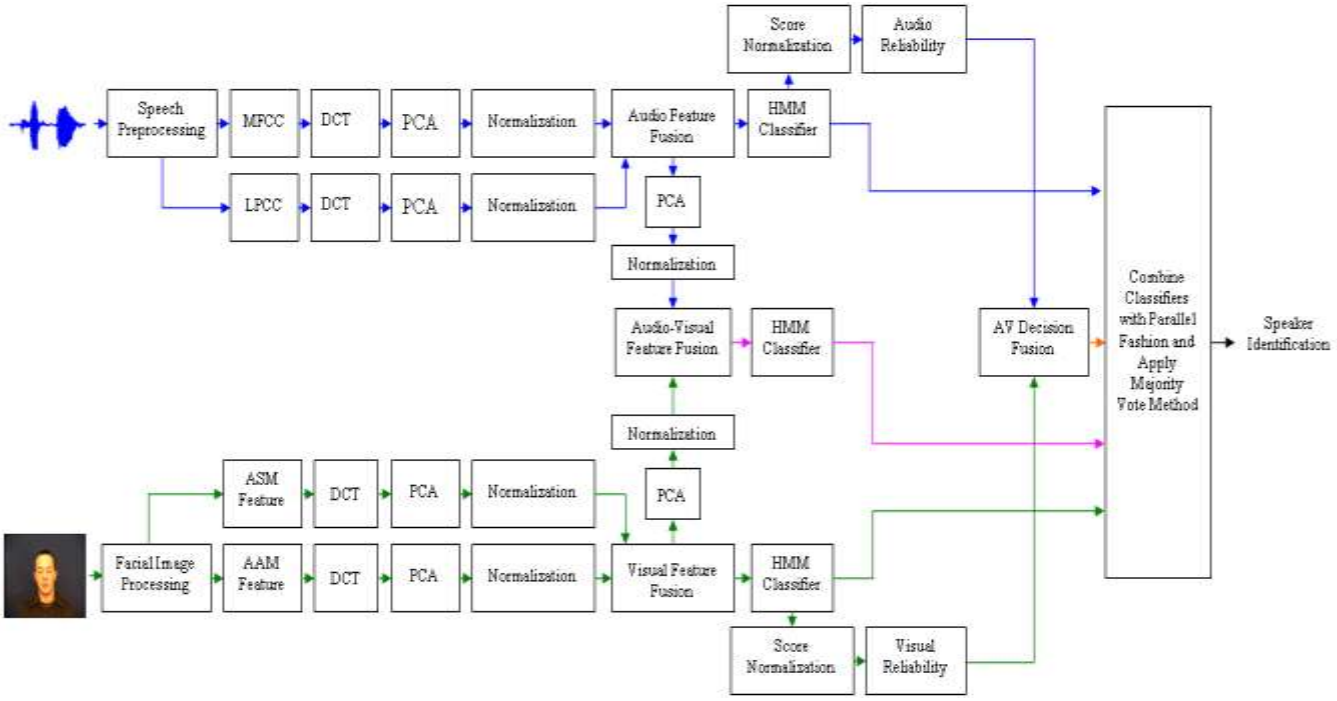

Figure 2: Details working procedure for the proposed hybrid feature and decision fusion based audio-visual speaker identification.

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

$$H(Z) = 1 - \alpha . z^{-1}, \; 0 \le \alpha \le 1 \quad (2)$$

Where α is the pre-emphasis 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 frame [21].

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 results from the framing process [22, 23]. The hamming window can be defined as follows [23]:

$$w(n)=\begin{cases}0.54-0.46\cos\frac{2\pi n}{N}, & -(\frac{N-1}{2})\le n\le(\frac{N-1}{2})\\ 0, & \text{Otherwise}\end{cases} \quad (3)$$

To extract the features from the speech utterances, various types of standard speech feature extraction techniques [24, 25, 26] such as RCC, MFCC, ΔMFCC, ΔΔMFCC, LPC, LPCC have been applied. Principal Component Analysis method has been used to reduce the dimensionality of the speech feature vector. Finally, HMM learning and classification algorithms [27, 28] have been applied to classify the speakers.

4. VISUAL IDENTIFICATION

The first step in image pre-processing is image acquisition. To do so, an imaging sensor along with signal digitization capability has been used so that captured image can be converted to digital form directly. 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 feature has been found. The procedure of the facial image pre-processing parts is shown in figure 3. To reduce the dimensionality of the facial feature vector, PCA and HMM training and testing algorithm have been used to classify the facial images.

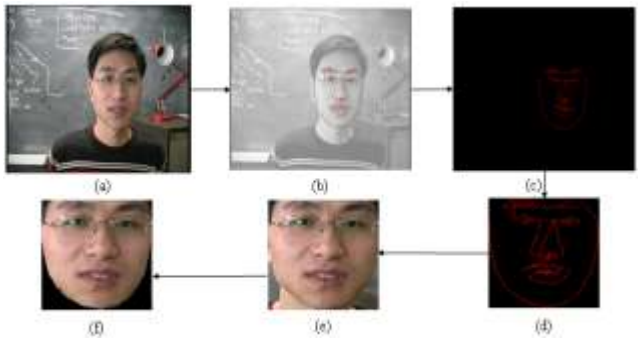

Figure 3: Facial image pre-processing for the proposed system (a) Original image (b) Output taken from Stams Active Appearance Model (c) Facial edges are extracted (d) Shape based features (e) Region Of Interest (ROI) selection with background noise (f) Appearance based facial features.

5. AUDIO-VISUAL FEATURE FUSION BASED IDENTIFICATION

The primary goal of the audio-visual feature fusion is when the noise level is low, the acoustic modality performs better than the visual one and, thus, the audio-visual identification performance should be at least as good as that of the acoustic speaker identification. When the noise level is high and the visual identification performance is better than the acoustic one, the integrated identification performance should be at least the same as or better than the performance of the visual-only identification [33]. Concatenation of two feature vectors produces result in feature vector with very large dimension. So, PCA has been used to reduce the dimension before using HMM classifier.

6. AUDIO-VISUAL LIKELIHOOD RATIO BASED SCORE FUSION

After the acoustic and visual sub-systems perform identification separately, their outputs are combined by a weighted sum rule to produce the final decision. Sensor level fusion and feature level fusion can be used before matching and after doing it, match score level, rank level and decision level fusion can be introduced. In this work, match score level was used to combine the audio and visual identification outputs. For a given audio-visual speaker test datum of O_A and O_V, the identification utterance C^* is given by [34],

$$C^*=\arg\max_i\{\gamma\log P(O_A/\lambda_A^i)+(1-\gamma)\log P(O_V/\lambda_V^i\} \quad (4)$$

Where λ_A^i and λ_V^i are the acoustic and the visual HMMs for the i^{th} utterance class respectively and $\log P(O_A/\lambda_A^i)$ and $\log P(O_V/\lambda_V^i)$ are there log likelihood against the i^{th} class.

Among various types of score fusion techniques, baseline reliability ratio-based integration has been used to combine the audio and visual identification results. The reliability of each modality can be measured from the outputs of the corresponding HMMs. When the acoustic speech is not corrupted by any noise, there are large differences between the acoustic HMMs output otherwise the differences become small. The reliability of each modality can be calculated by the most appropriate method which is best in performance [35],

$$S_m=\frac{1}{N-1}\sum_{i=1}^{N}(\max_j\log P(O/\lambda^j)-\log P(O/\lambda^i)) \quad (5)$$

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, $m\in\{A,V\}$.

Then the integrated weight of audio reliability measure γ_A can be calculated by [36],

$$\gamma_A=\frac{S_A}{S_A+S_V} \quad (6)$$

Where S_A and S_V are the reliability measure of the outputs of the acoustic and visual HMMs respectively.

The integrated weight of visual modality measure can be found as,

$$\gamma_V = (1-\gamma_A) \tag{7}$$

7. MULTIPLE CLASSIFIER FUSION

An effective way to combine multiple classifiers is required when a set of classifiers outputs are created. Various architectures and schemes have been proposed for combining multiple classifiers [37]. The majority vote [38, 39, 40, 41] is the most popular approach. Other voting schemes include the maximum, minimum, median [42], average [43] and product [44] schemes. Other approaches to combine classifiers include the rank-based methods such as the Borda count [45], the Bayes approach [40, 41], the Dempster-Shafer theory [41, 46, 47], the fuzzy integral [48], fuzzy connectives [49], fuzzy templates [50], probabilistic schemes [51], and combination by neural networks [52]. Majority vote approach has been used to combine four classifiers output in this work. The general voting routine can be defined as [53],

$$E(d) = \begin{cases} c_i & \underset{t\in\{1,..,m\}}{\forall} \sum_{j=1}^{n} B_j(c_t) \le \sum_{j=1}^{n} B_j(c_i) \ge \alpha.m + k(d) \\ r & \text{otherwise} \end{cases} \tag{8}$$

Where α is a parameter, k(d) is a function that provides additional voting constraints and the binary characteristics function can be defined as,

$$B_j(c_i) = \begin{cases} 1 \ if \ \ d_j = c_i \\ 0 \ if \ \ \mathrm{d_j} \neq c_i \end{cases} \tag{9}$$

Where the output of the classifiers from the decision vector, $d = [d_1, d_2, ...d_n]^T$ and $d_i \in \{c_1, c_2, ..., c_m, r\}$, c_i denotes the label of the i^{th} class and r denotes the rejection of assigning the input sample to any class.

8. EXPERIMENTALS RESULTS AND PERFORMANCE ANALYSIS

There are some critical parameters such as the number of frame length, frame increment, pre-emphasizing parameters for speech processing and cepstral coefficients, number of hidden states for HMM that affects the performance of the developed system. A trade off is made to explore the optimal values of the above parameters and experiments were performed using those parameters with clean speech utterances for both learning and identification. The optimal values of the above parameters were chosen and finally find out the results which are shown in the following subsections.

8.1 Optimum Parameter Selection for HMM

Various experiments have been performed for the selection of the optimum parameter on HMM. The highest identification of 98% has been achieved at the window length, N_L = 15 ms, frame increment, N_1 = 66%, pre-emphasizing parameter, α = 0.9, hidden states, N_H = 20 and the number of cepstral coefficients, N_{MC} =15. Figure 4 shows the results for various speech feature extraction technique i.e. MFCC, ΔMFCC, ΔΔMFCC, LPC and LPCC.

8.2 Performance Measurements of the Proposed System

VALID audio-visual database [54] has been used to measure the performance of the proposed speaker identification system. Artificial white Gaussian noise was added to the original clean speech utterances to simulate various SNR levels. The models were trained at clean speech utterances and tested under SNR level ranging from 0dB to 30dB at 5dB intervals. Table 1 shows the experimental results according to the VALID audio-visual database. Performance comparison among audio only, visual only, audio-visual feature fusion, audio-visual likelihood ratio based score fusion and combined classifiers i.e. majority vote output of the proposed system are shown in figure 5.

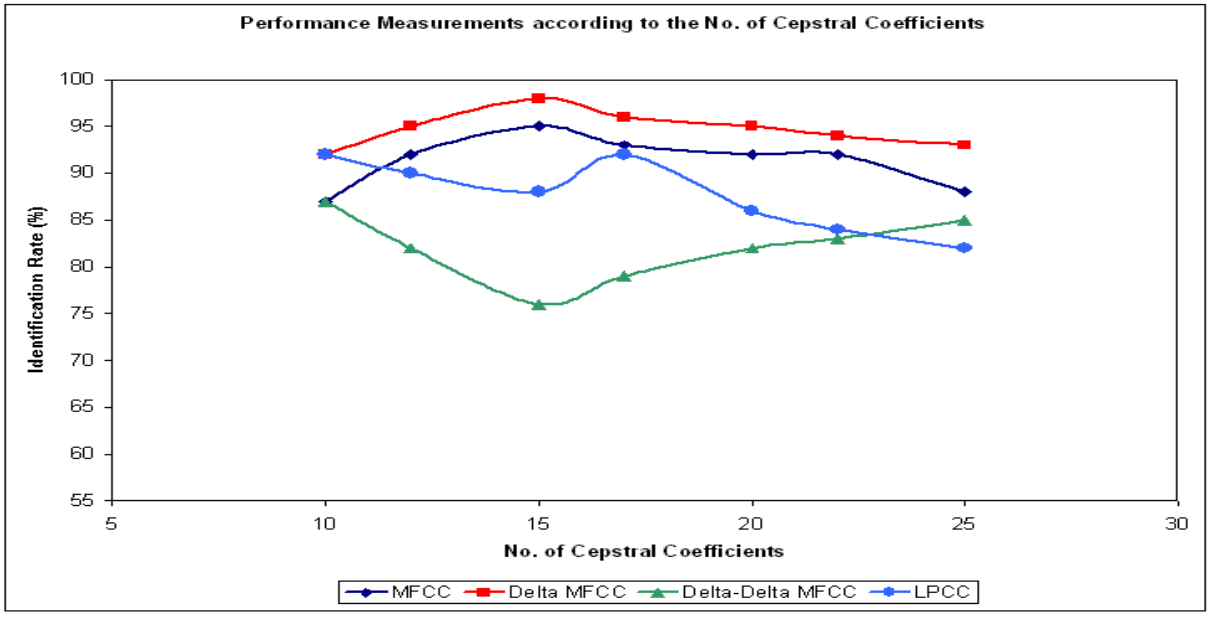

Figure 4: Speaker identification accuracy according to the number of cepstral coefficients.

Table 1. Performance measurements among Audio only, Visual only, AV feature fusion, AV score fusion and Majority vote approach according to various SNRs

SNR	Audio only (%)	Visual only (%)	AV feature fusion (%)	AV score fusion (%)	Majority Vote (%)
0	5.00	82.00	6.67	7.23	9.33
5	10.67	82.00	13.33	16.33	19.67
10	21.27	82.00	22.89	26.29	30.00
15	35.33	82.00	40.00	44.23	46.26
20	50.56	82.00	50.87	52.33	55.67
25	70.13	82.00	71.23	77.67	80.33
30	93.33	82.00	94.00	95.33	97.67
Average (%) (0dB ~ 30dB)	40.90	82.00	42.71	45.63	48.42

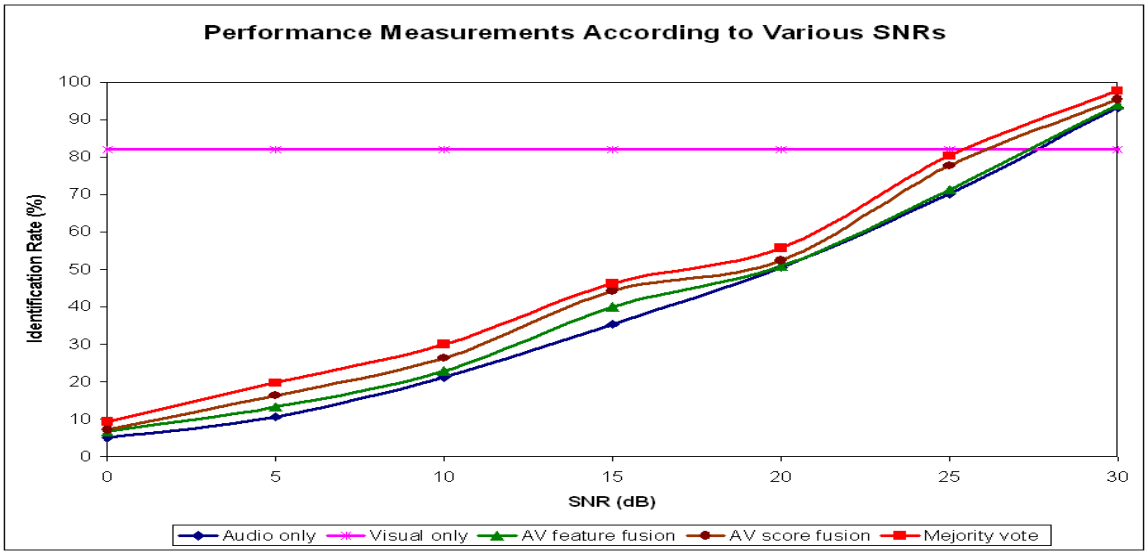

Figure 5: Performance comparison among audio only, visual only, audio-visual feature fusion, audio-visual likelihood ratio based score fusion and combined classifiers i.e. majority vote output of the proposed system.

The following observations have been accounted from the performance analysis of the proposed system.

- For visual only system the identification rate was found to be (82%) which remains constant regardless of acoustic SNR conditions. These values are larger than the acoustic only identification for noisy speech but smaller than for clean speech.
- The acoustic only identification rate degrades 93.33% to 5.00% with more artificially added white Gaussian noise.
- The average identification rate was found to be (42.71%) with AV feature fusion, (40.90%) with audio only and (45.63%) with AV score fusion.
- The majority vote approach (i.e. combined result of Audio only, Visual only, AV feature fusion and AV score fusion) achieves higher score than any other single and multimodal system.

9. CONCLUSIONS

We have proposed a novel architecture of introducing hybrid feature fusion and hybrid decision fusion for Audio-visual speaker identification system. This approach is general and is able to minimize the false rejection rate at a false acceptance rate. Experimental results according to the VALID database shows that the proposed hybrid feature and decision fusion based Audio-visual strategy achieves the best accuracies of speaker identification at all levels of acoustic signal-to-noise ratio, ranging from 0dB to 30dB. The identification rate of this system reveals that this proposed system can be used in various securities and access control purposes. The performance can also be populated according to large Audio-visual database.

10. REFERENCES

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

[2] R. Campbell, B. Dodd, and D. Burnham, Eds., *Hearing by Eye II*. Hove, United Kingdom: Psychology Press Ltd. Publishers, 1998.

[3] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition*," IEEE Trans. Multimedia,* 2(3):141–151, 2000.

[4] G. Potamianos, J. Luettin, and C. Neti, "Hierarchical discriminant features for audio-visual LVCSR," *Proc. Int. Conf. Acoust., Speech, Signal Process.*, pp. 165–168, 2001.

[5] Reynolds, D.A., "Experimental evaluation of features for robust speaker identification," *IEEE Transactions on SAP*, Vol. 2, 1994, pp. 639-643.

[6] Sharma, S., Ellis, D., Kajarekar, S., Jain, P. & Hermansky, H., "Feature extraction using non-linear transformation for robust speech recognition on the Aurora database," *Proc. ICASSP2000*, 2000.

[7] C. C. Chibelushi, F. Deravi, and J. S. D. Mason, "A review of speech-based bimodal recognition," *IEEE Trans. Multimedia*, vol. 4, pp. 23–37, Mar. 2002.

[8] D. N. Zotkin, R. Duraiswami, and L. S. Davis, "Joint audio-visual tracking using particle filters," *EURASIP J. Appl. Signal Processing*, vol. 2002, pp. 1154–1164, Nov. 2002.

[9] P. De Cuetos, C. Neti, and A. Senior, "Audio-visual intent to speak detection for human computer interaction," *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Istanbul, Turkey, June 5–9, 2000, pp. 1325–1328.

[10] D. Sodoyer, J.-L. Schwartz, L. Girin, J. Klinkisch, and C. Jutten, "Separatio of audio-visual speech sources: A new approach exploiting the audio-visual coherence of speech stimuli," *EURASIP J. Appl. Signal Processing*, vol. 2002, pp. 1165–1173, Nov. 2002.

[11] E. Foucher, L. Girin, and G. Feng, "Audiovisual speech coder: Using vector quantization to exploit the audio/video correlation," *Proc. Conf. Audio-Visual Speech Processing*, Terrigal, Australia, Dec. 4–6, 1998, pp. 67–71.

[12] J. Huang, Z. Liu, Y. Wang, Y. Chen, and E. Wong, "Integratio of multimodal features for video scene classification based on HMM," in *Proc. Works. Multimedia Signal Processing*, Copenhagen, Denmark, Sept. 13–15, 1999, pp. 53–58.

[13] E. Cosatto and H. P. Graf, "Photo-realistic talking-heads from image samples," *IEEE Trans. Multimedia*, vol. 2, pp. 152–163, Sept. 2000.

[14] Gerasimos Potamianos, Chalapathy Neti, and Sabine Deligne, "Joint Audio-Visual Speech Processing for Recognition and Enhancement," *Auditory-Visual Speech Processing Tutorial and Research Workshop (AVSP)*, pp. 95-104, St. Jorioz, France, September 2003.

[15] Simon Doclo and Marc Moonen, "On the Output SNR of the Speech-Distortion Weighted Multichannel Wiener Filter", *IEEE SIGNAL PROCESSING LETTERS*, VOL. 12, NO. 12, DECEMBER 2005.

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

[17] Koji Kitayama, Masataka Goto, Katunobu Itou and Tetsunori Kobayashi, "Speech Starter: Noise-Robust Endpoint Detection by Using Filled Pauses," *Eurospeech 2003*, Geneva, pp. 1237-1240.

[18] S. E. Bou-Ghazale and K. Assaleh, "A robust endpoint detection of speech for noisy environments with application to automatic speech recognition," *Proc. ICASSP2002*, vol. 4, 2002, pp. 3808–3811.

[19] Qi Li. Jinsong Zheng, Augustine Tsai, Qiru Zhou, "Robust Endpoint Detection and Energy Normalization for Real-Time Speech and Speaker Recognition," *IEEE Transaction on speech and Audion Processing*, Vol.10, No.3, March, 2002.

[20] Picone, J., "Signal modeling techniques in speech recognition," *Proceedings of the IEEE 81*, 9 (1993), pp. 1215–1247.

[21] L.P. Cordella, P. Foggia, C. Sansone, M. Vento., "A Real-Time Text-Independent Speaker Identification System", *Proceedings of 12th International Conference on Image Analysis and Processing*, IEEE Computer Society Press, Mantova, Italy, pp. 632 - 637 , September , 2003.

[22] F. Harris, "On the use of windows for harmonic analysis with the discrete fourier transform," *Proceedings of the IEEE 66*, vol.1 (1978), pp.51-84.

[23] J. Proakis and D. Manolakis, *Digital Signal Processing, Principles, Algorithms and Aplications.* Second edition, Macmillan Publishing Company, New York, 1992.

[24] D. Kewley-Port and Y. Zheng, "Auditory models of formant frequency discrimination for isolated vowels", *Journal of the Acostical Society of America*, 103(3), pp. 1654–1666, 1998.

[25] E. Zwicker., "Subdivision of the audible frequency band into critical bands (frequenzgruppen)", *Journal of the Acoustical Society of America*, 33:248–260, 1961.

[26] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", *IEEE Transactions on Acoustics Speech and Signal Processing*, 28, pp. 357–366, Aug 1980.

[27] M**.** Hwang**,** X. Huang**,** "Shared-Distribution Hidden**.** Markov Models for Speech Recognition", *IEEE. Trans. on. Speech and Audio Processing*, vol. 1, No. 4, pp. 414-420, April 1993.

[28] R.J. Elliott, L. Aggoun, and J.B. Moore, "Hidden Markov Models: Estimation and Control", *Applications of Mathematics: Stochastic Modeling and Applied Probability*, Vol. 29, Springer, Berlin, 1997.

[29] Stephen Milborrow and Fred Nicolls, "Locating Facial Features with an Extended Active Shape Model," available at http://www.milbo.org/stasm-files/locating-facial-features-with-an-extended-asm.pdf.

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

[31] J. Daugman, "Face detection: a survey," *Comput. Vis. Imag. Underst*, 83, 3, pp. 236- 274, 2001.

[32] Rafael C. Gonzalez and Richard E. Woods, *Digital Image Processing*. Addison-Wesley, 2002.

[33] Jong-Seok Lee and Cheol Hoon Park, *Speech Recognition, Technologies and Applications*, pp. 275-296, I-Tech, Vienna, Austria, 2008.

[34] A. Rogozan, P.S. Sathidevi, "Static and dynamic features for improved HMM based visual speech recognition," 1st

International Conference on Intelligent Human Computer Interaction, 9Allahabad, India, 20090, pp. 184-194.

[35] J. S. Lee, C. H. Park, "Adaptive Decision Fusion for Audio-visual speech Recognition", *Speech Recognition, Technologies and Applications*, ed. F. Mihelic, J. Zibert, (Vienna, Australia, 2008), pp. 550, 2008.

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

[37] Nayer Wanas, "Feature Based Architecture for Decision Fusion," Ph.D. thesis submitted to Systems Design Engineering, University of Waterloo, Ontario, 2003.

[38] R. Battiti and A. Colla. Democracy in neural nets: Voting schemes for classification. Neural Networks, vol. 7, no. 4, pp. 691–707, 1994.

[39] C. Ji and S. Ma. Combinations of weak classifiers. IEEE Transactions on Neural Networks, vol. 8, no. 1, pp. 32–42, 1997.

[40] L. Lam and C. Suen. Optimal combination of pattern classifiers. Pattern Recognition Letters, vol. 16, pp. 945–954, 1995.

[41] L. Xu, A. Krzy´zak, and C. 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.

[42] L. Kuncheva, "A theoritical study on six classifier fusion strategies," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 24, no. 2, pp. 281–286, 2002.

[43] P. Munro and B. Parmanto, *Competition among networks improves committee performance.* In Advances in Neural Information Processing Systems 9, pp. 592–598. MIT Press, Cambridge, 1997.

[44] D. Tax, M. Van Breukelen, R. Duin, and J. Kittler, "Combining multiple classifiers by averaging or by multiplying?", *Pattern Recognition*, vol. 33, pp. 1475–1485, 2000.

[45] T. Ho, J. Hull, and S. Srihari, "Decision combination in multiple classifier systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 1, pp. 66–75, 1994.

[46] T. Denouex, "A K-nearest neighbor classification rule based on Dempter-Shafer theory," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 25, no. 5, pp. 804–813, 1995.

[47] S. Le Hegarat-Mascle, I. Bloch, and D. Vidal-Madjar, "Introduction of neighborhood information in evidence theory and application to data fusion of radar and optical images with partial cloud cover," *Pattern Recognition*, vol. 31, no. 11, pp. 1811–1823, 1998.

[48] S. Cho and J. Kim, "Combining multiple neural networks by fuzzy integral for robust Classification," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 25, no. 2, pp. 380–384, 1995.

[49] L. Kuncheva. An application of owa operators to the aggregation of multiple classification decisions. In R. Yager and J. Kacprzyk, Eds., The Ordered Weighted Averaging Operators. Theory and Applications, pp. 330–343. Kluwer Academic Publishers, Dordrecht, 1997.

[50] L. Kuncheva, J. Bezdek, and M. Sutton, "On combining multiple classifiers by fuzzy templates," *Proceedings of the 1998 Annual Meeting of the North American Fuzzy Information Processing Society*, NAFIPS'98, pp. 193–197, Pensacola, FL, 1998.

[51] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, 1998.

[52] M. Ceccarelli and A. Petrosino, "Multi-feature adaptive classifiers for sar image segmentation," *Neurocomputing*, vol. 14, pp. 345–363, 1997.

[53] Dymitr Ruta and Bogdan Gabrys, "An Overview of Classifier Fusion Methods," *Computing and Information Systems*, 7 (2000), University of Paisley, p.1-10, 2002.

[54] N. A. Fox, B. A. O'Mullane and R. B. Reilly, "The Realistic Multi-modal VALID database and Visual Speaker Identification Comparison Experiments," *Proc. of the 5th International Conference on Audio- and Video-Based Biometric Person Authentication (AVBPA-2005)*, New York, 2005.

Authors Biographies

Md. Rabiul Islam was born in Rajshahi, Bangladesh, on December 26, 1981. He received his B.Sc. degree in Computer Science & Engineering and M.Sc. degrees in Electrical & Electronic Engineering in 2004, 2008, respectively from the Rajshahi University of Engineering & Technology, Bangladesh. From 2005 to 2008, he was a Lecturer in the Department of Computer Science & Engineering at Rajshahi University of Engineering & Technology. Since 2008, he has been an Assistant Professor in the Computer Science & Engineering Department, University of Rajshahi University of Engineering & Technology, Bangladesh. His research interests include bio-informatics, human-computer interaction, speaker identification and authentication under the neutral and noisy environments.

Md. Fayzur Rahman was born in 1960 in Thakurgaon, Bangladesh. He received the B. Sc. Engineering degree in Electrical & Electronic Engineering from Rajshahi Engineering College, Bangladesh in 1984 and M. Tech degree in Industrial Electronics from S. J. College of Engineering, Mysore, India in 1992. He received the Ph. D. degree in energy and environment electromagnetic from Yeungnam University, South Korea, in 2000. Following his graduation he joined again in his previous job in BIT Rajshahi. He is a Professor in Electrical & Electronic Engineering in Rajshahi University of Engineering & Technology (RUET). He is currently engaged in education in the area of Electronics & Machine Control and Digital signal processing. He is a member of the Institution of Engineer's (IEB), Bangladesh, Korean Institute of Illuminating and Installation Engineers (KIIEE), and Korean Institute of Electrical Engineers (KIEE), Korea.