THE JOURNAL OF THE ACOUSTICAL SOCIETY OF KOREA Vol.28, No.3E 2009. 9. pp. 109~117

Robust Person Identification Using Optimal Reliability in Audio-Visual Information Fusion

Md. Tariquzzaman*, Jin Young Kim*, Seung You Na*, Seung Ho Choi**
*School of Electronics & Computer Engineering, Chonnam National University
**Dept. of Computer Eng, Dongshin University
(Received August 31 2009; accepted September 5 2009)

Abstract

Identity recognition in real environment with a reliable mode is a key issue in human computer interaction (HCI). In this paper, we present a robust person identification system considering score-based optimal reliability measure of audio-visual modalities. We propose an extension of the modified reliability function by introducing optimizing parameters for both of audio and visual modalities. For degradation of visual signals, we have applied JPEG compression to test images. In addition, for creating mismatch in between enrollment and test session, acoustic Babble noises and artificial illumination have been added to test audio and visual signals, respectively. Local PCA has been used on both modalities to reduce the dimension of feature vector. We have applied a swarm intelligence algorithm, i.e., particle swarm optimization for optimizing the modified convection function's optimizing parameters. The overall person identification experiments are performed using VidTimit DB. Experimental results show that our proposed optimal reliability measures have effectively enhanced the identification accuracy of 7.73% and 8.18 % at different illumination direction to visual signal and consequent Babble noises to audio signal, respectively, in comparison with the best classifier system in the fusion system and maintained the modality reliability statistics in terms of its performance; it thus verified the consistency of the proposed extension.

Keywords: *Person Identification, Local PCA, Reliability Measures, Particle Swarm Optimization*

I. Introduction

Nowadays, daily communication and dealings between people and organizations such as government entities, public entities and private organizations are performed through human computer interaction where identification of the user is a vital concern to the organizations. Multimodal person identification, a goal of biometrics technology, is deployed increasingly in the secured organizations. Person Identification is a task of identifying a person using speaker behavioral/physiological traits such as speech signal and visual signal which are carrying linguistic/non-linguistic information and visual information, respectively. The development of the person recognition technology, particularly multimodal person recognition, is still an active area of research which has a range of applications from national security, communication system security, computer security, computer network security, e-commerce to forensics [2][4]. Besides, increasing demands of intelligence interface with systems like humanoid robots have created a need for automatic person recognition. With the current state of art, multimodal person identification system may perform well under controlled testing conditions. However, in real environment, voice information can easily be exposed to a different degree of noises due to the

Corresponding author: Jin Young Kim (beyondi@jnu.ac.kr)
School of Electronics & Computer Engineering, Chonnam National University, 300 Yongbong-Dong Buk-Gu, Gwangju 500-757, Republic of Korea

channel distortion of the various transmission medium, its surrounding environments and codec distortion meanwhile, the visual signal also can be degraded due to image quality, illumination change and occlusion. As a result, degradation in the performance of single modality based speaker/person identification as well as the multimodal person identification system occurs. Nonetheless, multimodal biometrics system performance is better than single modality based person identification system. To design a multimodal biometrics system, a number of issues are focused which are mainly categorized into three: which modality to be fused, where to be fused and how the different modalities information to be fused including its reliability. Besides the single modality based speaker identification /verification system [1] [4] [5] [8] [9] [11-13] [16-22] [25], there are different approaches existing in the literature [6-10] [15] [20] [23] for multimodal fusion where reliability measures are one of the key issues in the fusion process. The ultimate goal to measure the reliability is to reduce classification error rates and thus to increase the performance and robustness of a multimodal person identification system. Mostly, in the applied different approaches for multimodal fusion mainly the audio signal based expert's performance are relatively poor in comparison with the counterpart visual signal based expert's performance. Nevertheless, due to the uncertainty of the lighting condition to the visual signal in real world scenarios and consequently visual signal quality, the visual signal based expert performance may get relatively poor, which also leads to degradation in performance of a multimodal person identification system.

In this paper, we will extend the modified convection reliability function's optimizing parameter proposed in [21] to account for optimal reliability simultaneously through the audio and lip information based reliability measure in bimodal speaker identification system. In this system particularly the visual signal based expert misclassification rate is higher in comparison with the audio signal based expert for measuring the reliability statistics. For image quality degradation JPEG compression is performed to lip images to take poor quality image. In addition, Babble noises and artificial illumination are added to testing speech and visual signal, respectively. Local PCA [11] is adopted for both classifiers features dimension reduction. For individual experts' speaker model generation, we have implemented the classical GMM training. The proposed modified convection function optimizing factors are optimized by particle swarm optimization [7]. The entire experiments are performed using VidTimit DB [20].

II. System Architecture

In this section, we will describe the individual classifiers training, testing methodology and the overall baseline system.

2.1. Gaussian Mixture Model based Classifier

Among the different models [3] [5] [17] creation, Gaussian Mixture Model (GMM) is one that has merit by itself. GMMs model the static characteristics of the observation signal. The most basic form of a GMM consists of a single Gaussian probability density function:

$$b(o_t;\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^{D/2}|\Sigma|}}\exp\left\{-\frac{1}{2}(o_t-\mu)'\Sigma^{-1}(o_t-\mu)\right\}. \quad (1.a)$$

where μ and Σ_m represent the mean vector and covariance matrix respectively; D represents the dimensionality of o_t. In general a GMM consists of a weighted aggregation of Gaussians,

$$p(o_t|\lambda) = \sum_{m=1}^{N_m} c_m b_m(o_t;\mu_m,\Sigma_m) \quad (1.b)$$

where μ_m and Σ_m represent the mean and covariance matrix respectively of the m^{th} Gaussian b_m; and c_m is the weight of the m^{th} Gaussian such that $c_m \in [0,1]$ and $\sum_{m=1}^{M} c_m = 1$. N_m is the number of Gaussian mixtures in the GMM. The set of parameters for a given GMM,

λ, are denoted by

$$\lambda = \{c_m, \mu_m, \Sigma_m\}\Big|_{m=1}^{N_m} \quad (1.c)$$

A high degree of smooth densities can be computed in practice, if the number of GMM component densities is not restricted. EM Algorithm was applied to iterative training for the speaker model generation.

At the testing stage, the classification is performed by calculating the likelihood as follows while considering the features are statistically independent:

$$l(O|\lambda_n) = \prod_{t=1}^{T} p(o_t|\lambda_n) \quad (1.d)$$

where λ_n is the speaker/person model in the case of speaker/person identification, where $n = 1,2,...,N$. Finally, the identification is determined by:

$$\arg\max_n(\log[l(O|\lambda_n)])$$

In the experiments presented in this paper, MFCC and its delta features are extracted from speech signal and CMS [8] is performed on these features vector. Finally the local PCA is applied to these robust features to obtain the reduced dimension features vector which are used in audio based classifier system. Similarly, DCT features of visual signal are also reduced by local PCA which are used in the lip based classifier system. Manually we have taken lip ROI considering the lip center as the base point for ROI determination and thus created the lip database.

In individual classification process, each modality expert generates the log-likelihood score generally termed as $S(O_m, \lambda_n)$ where O is either voice or visual information. When m=1, it indicates that it is audio information, while at m=2 it is lip information and n of λ_n is the n-th speaker of the GMM model.

2.2. Reliability Measures and Audio-visual Information Integration

Broadly, the reliability measure uses the statistical nature of the experts which can be categorized into two, i.e., the reliability parameters can be measured at either the signal level or at the expert score level. Even though there are some methods considering the signal level confidence measure, the higher level confidence measure is more desirable. In the literature, there are different methods for score based reliability measure, which are mainly score entropy, dispersion, variance, cross classifier coherence and score difference. The score difference [15] [23] is our special interest which has been taken as the baseline method in this paper. The baseline method takes the statistical nature of observation probability based on the rank of information and the min-max normalized method is performed on all observation probabilities. The entire process for reliability parameters calculation is followed by the following sequences:

First, the individual modality generates the likelihood scores which are normalized using min-max normalization method differently. Mathematically we can express

$$\bar{S}(O_m|\lambda_j) = \frac{S(O_m|\lambda_j) - \mathrm{Min}P_m}{\mathrm{Max}P_m - \mathrm{Min}P_m} \quad (2)$$

where $\mathrm{Max}P_m = \max_j S(O_m|\lambda_j)$ and $\mathrm{Min}P_m = \min_j S(O_m|\lambda_j)$.

Secondly, the different modalities reliability parameters are calculated using the highest rank of the normalized score, i.e.,

$$\rho_m = \overline{\mathrm{Max}P_m} - \overline{\mathrm{Max}2P_m} \quad (3)$$

Here, $\overline{\mathrm{Max}P_m}$ is the highest rank of the expert score where m represent the modality either audio or video and $\overline{\mathrm{Max}2P_m}$ is the second highest rank of the individual expert score. After normalization $\overline{\mathrm{Max}P_m}$ becomes 1 if there is no preferred range and the above equation can be written as

$$\rho_m = 1 - \overline{\mathrm{Max}2P_m} \quad (4)$$

Thirdly, for mapping in between the reliability and the expert weighting factor, the weighting factor is

calculated from the reliability function as follows:

$$\alpha_1 = \frac{\rho_1}{\rho_1 + \rho_2}, \quad \alpha_2 = 1 - \alpha_2 \tag{5}$$

Fourthly, each modality score is integrated using the weighting factor as

$$S(O_1, O_2 | \lambda_n) = \alpha_1 \tilde{S}(O_1 | \lambda_n) + \alpha_2 \tilde{S}(O_2 | \lambda_n) \tag{6}$$

Then finally $n^* = \arg\max_n S(O_1, O_2 | \lambda_n)$ is the identified speaker.

III. Proposed Modified Reliability Function

3.1. Background of Introducing Optimization Factors on Reliability Function

From the above described section, we see that integrated scores of observation are calculated using equations (2) to (6). Specifically, the reliability value which is expressed by equation (4) is determined from the audio and visual modality information individually and it has an important role in enhancing the speaker recognition performance. For each expert, reliability value is expressed by equation (4) which was derived from the normalized observation probabilities and then the argument logic is applied for speaker identification. Explicitly, even though the convection function's, i.e., $\rho_m = f(\tilde{S}(O_m | \lambda_1), \tilde{S}(O_m | \lambda_2), \ldots, \tilde{S}(O_m | \lambda_n))$ final goal is recognition rate; nonetheless, there is no optimization parameter that maximizes the recognition rate in the bimodal speaker identification system. We can think two different possible conditions in the reliability measure regarding reliability function expressed by equation (4).

1. **Overestimation:** Reliability value is estimated in higher order rather than its optimum level; in this case, we should regulate the reliability function so that the reliability value reaches at the optimum point.
2. **Underestimation:** Reliability value is estimated poorly from its ground truth; in this case, we should also regulate the reliability function to raise the reliability value. Introducing an optimization factor in reliability function, we can control the reliability function so that an improvement in audio-visual speaker identification could be achieved thus for the ultimate goal of fusion to be fulfilled.

Thus, we have modified the convection function expressed by equation (4) by extending the proposed method in [21] through introducing optimization factors on both modalities, i.e., introducing the optimization factors let f_m on $\overline{\max 2P_m}$, so that the above two cases can be controlled. Mathematically, we can express the modified reliability function as follows:

$$\rho_m' = 1 - (\overline{\mathrm{Max2P}_m})^{f_m} \tag{7}$$

In equation (7) f_m i.e., f_1 and f_2 are the optimization variables. Considering $\overline{\max 2P}$ we can conclude the following condition for the modified reliability values for different limit values of f_m.

- $0 \le f_m \le 1 : \rho_m' < \rho_m$
- $f_m = 1 : \rho_m' = \rho_m$
- $f_m > 1 : \rho_m' > \rho_m$

Figure 1 shows the relationship and physical meaning

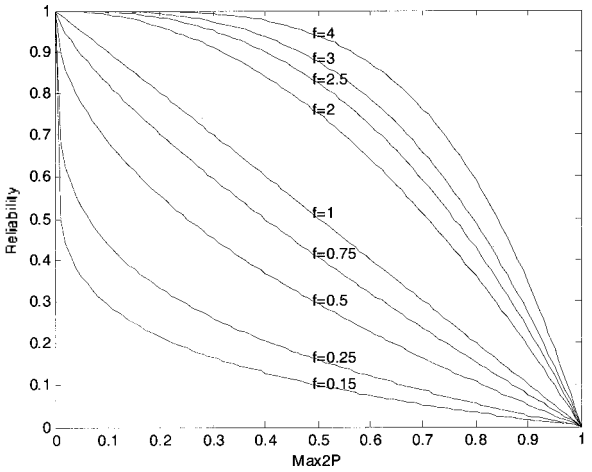


Figure 1. Graphical views of reliability values at Different f on $\overline{\mathrm{MAX2P}}$.

between $\overline{\max 2P}$ and reliability values for different values of the optimization factor f.

However, we cannot get the optimum values f_m in linear searching way and thus we need optimization. For optimization we need a target function which is shown in detail in the following section.

3.2. Optimization Target Function

The optimization object function can be defined with the help of speaker identification rate. The optimization object function is defined in the following:

$$x(f_1,f_2)=\frac{\sum_{k=1}^{K}\sum_{l=1}^{L_k}\delta(\arg\max_j(P_j(X_{kl})),k)}{K\sum_{k=1}^{K}L_k} \quad (8)$$

In the above function, $\delta(i,j)$ is the delta function, X_{kl} is the $l-th$ speech feature vectors of the $k-th$ speaker, and P_j is the observation probability of given feature sequence for $m-th$ speaker. The parameter $\arg\max_j P_j$ is theindex of the speaker having maximum probability and P_j is defined by equation (6). The optimization object function having optimization variable f_1 and f_2 expressed by equation (8) is a nonlinear function and it is impossible to find the closed solution of the function. Thus the PSO approach is applied for optimizing the object function. It is expected that the optimized parameter obtained by PSO enhances the performance of speaker recognition. Among various optimization methods, particle swarm optimization is adopted in our experiments due to its simplicity. In the next section the PSO algorithm is described in detail.

3.3. PSO Based Optimization

Particle swarm optimization (PSO) is a population based stochastic optimization technique developed in 1995 by James Kennedy and Russ Eberhart. PSO was devised by imitating movements of a bird flocks or insects. PSO performs with the renewal of each latent solution by combining previous latent solutions, which are called particle solution.

As a general problem, let's assume that a function $f(.)$ should be optimized with the parameter P. Then PSO method is described as follows:

1) Randomize each latent solution $\{P_{i0}\}$
2) For each $j-th$ iteration, perform the followings:
   2-1) calculate $f(P_{ij})$ for each P_{ij}
   2-2) Test the convergence. If converged, break the loop.
   2-3) For each i, calculate optimal solution on iterations $\{0,\ldots,j-1\}$.

Let us them be denoted by $pbest_j$.

   2-4) calculate the global optimal solution on $pbest_{ij}$. Let it be $gbest_j$.
   2-5) calculate the velocity of each particle as follows.

$$v_{ij}=v_{ij-1}+c_1r_1(pbest_j-P_{ij-1})+c_2r_2(gbest_j-P_{ij}) \quad (9)$$

where c_1 and c_2 are constants and r_1 and r_2 are random variables.

   2-5) Renew each particle.

$$P_{ij}=P_{ij-1}+v_{ij} \quad (10)$$

3) Determine the optimal solution as $gbest_j$.

PSO is a simple but powerful search technique which requires only primitive mathematical operators and has been applied successfully to a wide variety of search and optimization problems.

IV. Experiment and Investigation

4.1. Experimental Database and Specifications

For the experiments of this research work, the VidTIMIT audio-visual database was employed. The VidTIMIT database contains 43 speakers (19 female and 24 male), reciting short sentences selected from the NTIMIT corpus. The data were recorded in 3 sessions, with a mean delay of 7 days between

Session 1 and 2, and of 6 days between Session 2 and 3. The mean duration of each sentence is around 4 seconds, or approximately 106 video frames. A broadcast-quality digital video camera in a noisy office environment was used to record the data. The video of each person is stored as a sequence of JPEG images with a resolution of 512 × 384 pixels (columns × rows), with corresponding audio provided as a monophonic, 16 bit, 32 kHz WAV file.

4.2. Experiments and Results

For our proposed method validation, we have used VidTimit database which contains 10 sentences for each speaker in audio and visual signal level. However, we have taken only 9 utterances of each speaker and divided them into three groups: Group I (1-3 utterances), Group II (5-7 utterances) and Group III (8-10 utterances). Group I is used for speaker model generation while Group II and Group III are involved in PSO training and validation for f_1 and f_2 in testing stage for speaker identification. For the robustness test we have added the Babble noises to the testing dataset obtained from NoiseX-92 database. For degrading the image quality we have applied JPEG compression with a quality factor (QF) of 25. Moreover, we have added an artificial illumination to the respective testing visual image as follows:

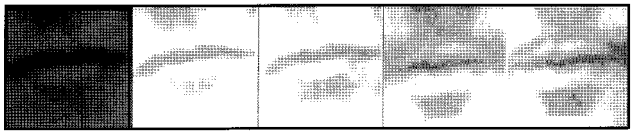
$$I(y,x) = w(y,x) + |\varphi| d + \delta \qquad (11)$$

where $y = 1,2,...,M_Y$, $x = 1,2,...,N_X$, and d is either y or x depending on the illumination direction, and $\varphi = -\delta/\delta$. In our experiment we have added the artificial illumination from down-to-up (DU) and left-to-right (LR) directions. Examples of the ROI lip images with artificial illumination in different directions are shown in Fig. 2. For features dimension reduction, we have applied local PCA to both the audio and visual feature vector. Before features extraction we have transformed the color image to gray image. Table 1 shows the overall specifications for our experiment.

The different experimental results are presented in Table 2 through Table 5. Table 2 shows the performance of audio based classifier with a high degree of signal distortion due to the babble noises. Only lip expert performances are depicted in Table 3. Due to the uncertainty of lighting condition on the comparatively

Table 1. GMM Based Experts Specifications.

Modality	Features	No. of Mixtures in GMM
Audio	Frame Level Features: MFCC + Delta Dimension reduction: local PCA Original Features Dimension: 38 No. of PC's: 10 Reduced Features Dimension: 10	3
Lip Image ROI	Input Image: 6464 pixel Compression: JPEG Quality factor (QF): 25 Features: DCT Dimension reduction: local PCA No. of PC's: 10	3



(a) (b) (c) (d) (e)

Figure 2. Different images view : (a) Train image; (b)-(e) Test images Light direction on different lip images:(b) & (d) Down-to-Up (DU) (c) & (e) Left-to-Right (LR).

Table 2. Only Audio Based Expert Performance.

Average SNR (dB)	Train DS Group	Test DS Group	IR (%)
1.8	I	II+III (5-10)	72.47

Table 3. Only Lip Based Expert Performances.

QF	δ	light Direc.	m	Train DS Gr.	Test DS Group	IR (%)
25	160	DU	-δ/δ	I	III (8-10)	49.61
25	160	DU	-δ/δ	I	II (5-7)	48.83
25	160	DU	-δ/δ	I	II+III	49.22
25	160	LR	-δ/δ	I	III (8-10)	51.93
25	160	LR	-δ/δ	I	II (5-7)	51.93
25	160	LR	-δ/δ	I	II+III	51.93

Table 4. Baseline Audio-visual Integrated System Performance.

Modality	Train DS Group	light Direc.	Test DS Group	IR (%)
Audio + Visual	I (1-3)	DU	II+III (5-10)	49.22
Audio + Visual	I (1-3)	LR	II+III (5-10)	51.93

Table 5. Proposed Method based Audio-visual Fused System Performances.

Light direction	Train DS Gr.	Test DS Gr.	Optimized f_1	Optimized f_2	Test DS Gr.	IR (%) using optim. f_1 & f_2	Avg. IR (%)
DU	I	III	2.565	0.0128	II	81.39	80.22
	I	II	1.987	0.0114	III	79.06	
LR	I	III	2.215	0.0156	II	82.17	80.65
	I	II	1.893	0.0137	III	79.06	

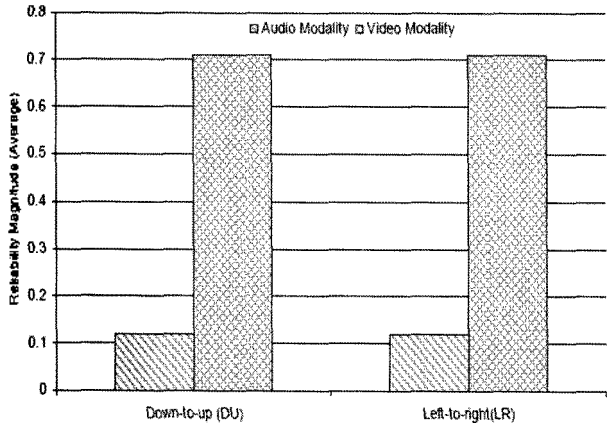


Figure 3. Graphical Views of Different Modalities Reliability Values in Average at Baseline System.

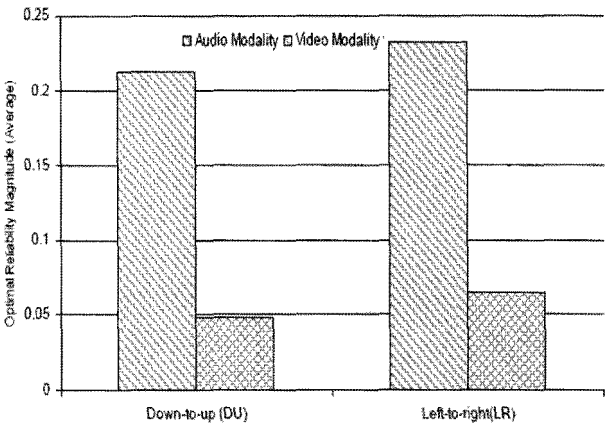


Figure 4. Graphical Views of Different Modalities Optimum Reliability Values in Average.

poor quality lip images, the lip based classifier system performances are degraded in comparison with its counterpart audio based classifier. Table 4 shows the experimental results of our adopted baseline system; its performance is very far from the fusion phenomena as it fails to reach higher or equal to the performance of best classifier system i.e., here the audio based classifier system. We have presented our proposed method based experimental results in Table 5. It is seen from the Table 5 that the proposed method based bimodal biometrics system fulfilled the desire of fusion strategy through an improvement of speaker identification rate of about 7.73% and 8.18% in average for down-to-up (DU) and left-to-right (LR) lighting direction, respectively, using the optimum reliability.

Figure 3 and Figure 4 show the reliability values in average of different modality obtained from the baseline system and proposed system respectively. Figure 3 shows that the audio based classifier system average reliability values are poor in comparison with that of its counterpart lip based classifier system even though the audio based classifier system performance is higher than the lip based classifier performances. It is seen from Figure 4 that the optimal reliability magnitude of audio classifier is higher than the lip based classifier and thus the proposed method also fulfilled the consistency of the current state of art reliability statistics.

V. Conclusion

In this paper we have presented a robust bimodal person identification system considering the optimal reliability in the integration process of audio and visual information. Introducing our proposed optimization factors in the existing convection function improved the performance of the bimodal robust person identification system significantly. Moreover, the proposed optimization factors clearly balanced the trend of reliability statistics between the modalities in terms of modalities performances. The optimization factors were optimized by PSO algorithm. The entire experiments were performed using the VidTimit database. With the bimodal person identification system, we

confirm that the proposed modified convection function could be a promising solution in multimodal biometrics technology. For further study, we will focus on multimodal based person identification, audio-visual person verification and particularly the audio-visual speech recognition in real environment.

Acknowledgements

This work was partially supported by National Research Foundation of Korea Grant funded by the Korean Government (2009-0077345). Also this research was partially supported by the MKE (The Ministry of Knowledge Economy), Korea, under the ITRC (Information Technology Research Center) support program supervised by the NIPA (National IT Industry Promotion Agency" (NIPA-2009-(C1090-0903-0008)).

References

1. B.S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Am.*, vol. 55, no. 6, pp. 1304-1312, 1974.
2. A.K. Jain, A. Ross, S. Prabhakar, "An introduction to biometric recognition," *IEEE Tran. Circuits Sys. Video Technol.*, vol. 14, no. 1, pp. 4-20, 2004.
3. M. Brand, N. Oliver, A. Pentland, "Coupled hidden Markov models for complex action recognition," In *Proc. of IEEE Internat. Conf. on Computer Vision and Pattern Recognition*, pp. 994-999, 1997.
4. J. P. Campbell, "Speaker recognition: a tutorial," In *Proc. IEEE*, vol. 85, no. 9, pp. 1437-1462, 1997.
5. R. Chengalvarayan and L. Deng, "A maximum a posteriori approach to speaker adaptation using the trended hidden Markov model," *IEEE Trans. Speech Audio Proc.* vol. 9, no. 6, pp. 549-557, 2001.
6. U. V. Chaudhari, et al., "Audio-visual speaker recognition using time-varying stream reliability prediction," *Proceeding of IEEE Int. Conference on Acoustics, speech and signal proc.* vol. 5, pp. 712-715, 2003.
7. R. Eberhart and J. Kennedy, "A new optimizer using particle swarm theory," In *Proc. Sixth Int. Symposium on Micro Machine and Human Science*, pp. 39-43, 1995.
8. S. Furui, "Cepstral Analysis technique for automatic speaker verification," *IEEE Trans. on Acoustics, Speech, and Signal Proc.*, vol. 29, no. 2, pp. 254-272, 1981.
9. H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. on Speech and Audio Proc.*, vol. 2, no. 4, pp. 578-589, 1994.
10. M. Heckmann, F. Berthommier, and K. Kristian, "Noise adaptive stream weighting in audio-visual speaker identification," *EURASIP J. Applied Signal Proc.* vol. 2002, pp. 1260-1273, 2002.
11. N. Kambhatla, and T.K. Leen, "Dimension reduction by local PCA," *Neural Computation*, vol. 9, no. 7, pp. 1493-1503, 1997.
12. C.H. Lee, C.H. Lin and B.H. Juang, "A Study on speaker adaptation on the parameters of continuous density hidden Markov models," *IEEE Trans. of Signal Proc.* vol. 39, no. 4, pp. 806-814, 1991.
13. R.J.Mammone, X. Zhang and R. P. Ramachandran, "Robust speaker recognition: a feature-based approach," *IEEE Signal Processing Magazine* vol. 13, no. 5, pp. 58-71, 1996.
14. E. Mengusoglu, "Confidence measure based model adaptation for speaker verification," In *Proc. 2nd IASTED Internat. Conf. on Communications, Internet and Information Technology*, pp. 408-411, 2000.
15. N. A. Fox, *Audio and video based person identification*, Ph.D. thesis, University College Dublin, 2005.
16. D. A. Reynolds, "An overview of automatic speaker recognition technology," *Proc. IEEE Internat. Conf. on Acoustics, Speech and Signal Processing*, vol. 4, pp. 4072-4075, 2000.
17. D. A. Reynolds, R. C. Ross, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech Audio Proc.* vol. 3, no.1, pp. 72-82, 1995.
18. D. Stephane, and R. Christophe, "Robust feature extraction and acoustic modeling at multitel: experiments on the Aurora databases," In *Proc. Eurospeech*, pp. 1789-1792, 2003.
19. C. H. Sit, M. W. Mak and S. Y. Kung, "Maximum likelihood and maximum a posteriori adaptation for distributed speaker recognition systems," In *Proc. of 1st Internat. Conf. on Biometric Authentication*, pp. 640-647, 2004.
20. C. Sanderson, Biometric Person Recognition: Face, Speech and Fusion, *VDM-Verlag*, 2008.
21. M. Tariquzzaman, Jin Young Kim and Joon-Hee Hong, "Improvement of reliability based information integration in audio-visual person identification," *J. Korean Soc. of Phonetic Sci. Speech Technol.* Vol. 62, pp. 149-161, 2007.
22. K. Yiu, M. .Mak and S. Kung, "Environment adaptation for robust speaker verification," In *Proc. EUROSPEECH*, pp. 2973-2976, 2003.
23. T. Wark, and S. Sridharan, "Adaptive Fusion of speech and lip information for robust speaker identification, *Digital Signal Processing*, vol. 11, pp.169-186, 2001.
24. M. Tariquzzaman, J. Y. Kim and S. Y. Na, "Robust Audio-visual Speaker Identification Using a Modified Score-based Reliability in Modality Integration," *(accepted) The International ACM Conference on Management of Emergent Digital EcoSystems*, 2009.
25. B. Zhen, X. Wu, Z. Liu and C. Huisheng, "An Enhanced RASTA processing for speaker identification," In *Proc. ICSLP*, pp. 251-254, 2000.

[Profile]

•Md. Tariquzzaman

Md. Tariquzzaman received the B.Sc. and M.Sc. degree in Applied Physics, Electronics and Communication Engineering (Erstwhile Electronics & Applied Physics) from Islamic University, Kushtia-7003, Bangladesh, in 2001 and 2003, respectively. He joined as a lecturer in the Dept. of Information and Communication Engineering, Islamic University, Kushtia-7003, Bangladesh, in September 2004 and on study leave since March 2006. He is currently pursuing PhD degree in the School of Electronics & Computer Engineering at Chonnam National University, Republic of Korea. His research interests are in the area of biometrics, computer vision and machine learning.

•Jin Young Kim

1986. 2 : Dept. of Electronics Eng. Seoul Nat'l Univ.(BS)
1988. 2 : Dept. of Electronics Eng. Seoul Nat'l Univ.(MS)
1993 .8 : Dept. of Electronics Eng. Seoul Nat'l Univ.(Ph.D)
1995 ~ : Chonnam Nat'l Univ., (professor)
Research Area: Audio-visual signal processing

•Seung You Na

1977.2 : Dept. of Electronics Eng. Seoul Nat'l Univ.(BS)
1986 : Dept. of ECE University of Iowa(Ph.D)
1987 ~ : Chonnam Nat'l Univ. (Professor)
Research Area: Intelligent control, Signal processing

•Seung Ho Choi

1981.2 : Dept. of Physics Chonnbuk Univ. (BS)
1984.8 : Dept. of Electronics Eng. Myungji Univ. (MS)
1992.2: Dept. of Electronics Eng. Myungju Univ (Ph.D)
Research Area : Multimedia signal processing