IEICE TRANS. INF. & SYST., VOL.E94–D, NO.10 OCTOBER 2011

LETTER

A Visual Signal Reliability for Robust Audio-Visual Speaker Identification

Md. TARIQUZZAMAN[†], *Member*, **Jin Young KIM**[†a)], **Seung You NA**[†], **Hyoung-Gook KIM**[††], *Nonmembers*, *and* **Dongsoo HAR**[†††], *Member*

SUMMARY In this paper, a novel visual signal reliability (VSR) measure is proposed to consider video degradation at the signal level in audio-visual speaker identification (AVSI). The VSR estimation is formulated using a Gaussian fuzzy membership function (GFMF) to measure lighting variations. The variance parameters of GFMF are optimized in order to maximize the performance of the overall AVSI. The experimental results show that the proposed method outperforms the score-based reliability measuring technique.

key words: *visual signal reliability measure, Gaussian fuzzy membership function, audio-visual speaker identification*

1. Introduction

Multi-modal integration for audio-visual speaker identification (AVSI) is one of the robust approaches [1]–[5] in noisy environments, where speech signals have relatively high levels of distortion. The main issues concerning AVSI involve an integration structure and reliability decision-making. In state-of-the-art bimodal biometrics technology, a number of approaches have been proposed for reliability measures, which can be broadly classified into signal-based measures (SIM) and score-based measures (SCM). SCMs utilize the distribution of the model probabilities [1]–[3], and in SIM, SNR and the voicing index are the representative parameters for audio signal [4], [5] only. However, until now there has been no SIM for the corresponding video signal in AVSI. Thus, SCMs are generally used for AV integration, or SIM is adopted with the assumption that video signal is not corrupted or degraded by variation of lighting conditions and camera setup between installment and testing image collection. Therefore, measuring the visual reliability at the signal level is an important issue in dealing with the lighting variation in real-environment images for the robust AVSI. This letter proposes a novel visual signal reliability (VSR) estimation technique at the signal level formulated by a Gaussian fuzzy membership function (GFMF). The proposed VSR method employs lighting change indicators that measure the global and directional lighting variations in visual signals. The proposed method is evaluated by performing AVSI experiments with the VidTimit database [6].

Manuscript received November 18, 2010.
Manuscript revised April 20, 2011.

[†]The authors are with the Dept. of Electronics Engineering, Chonnam National University, 500–757, South Korea.
[††]The author is with the Dept. of Wireless Communications Eng., Kwangwoon University, 139–701, South Korea.
[†††]The author is with the Dept. of Information & Communications, GIST, South Korea.
a) E-mail: beyondi@jnu.ac.kr

DOI: 10.1587/transinf.E94.D.2052

2. Baseline AVSI System: Score-Based Fusion

Based on the classical Gaussian mixture model [7], we implemented the classifiers of the individual modalities in parallel, and the late integration (LI) approach [2] is adopted for integrating the audio and visual information. In the AVSI system lip information is selected as visual modality. Figure 1 shows the baseline AVSI system. The main features of the implemented system are included in Table 1. For audio-visual fusion, we adopted the score dispersion as an SCM approach in the baseline fusion scheme. Assuming K speakers, the fusion procedure is as follows [1], [2].

a) Generate the audio and video log-likelihood scores through individual classifiers for the input of AV fusion; $\{S_k^A = log\, P(O_A \mid \lambda_k^A)\}$ and $\{S_k^V = log\, P(O_V \mid \lambda_k^V)\}$, for $k = 1, 2 \cdots, K$, where O_A and O_V are audio and video observations, respectively, and λ_k^A and λ_k^V are the audio and video GMMs for the *k-th* subject.

b) Normalize the audio and video scores based on min-max rule, which shifts and scales the scores into the range $[0, 1]$; $\tilde{S}_k^m = (S_k^m - S_{\min}^m)/S_{\max}^m - S_{\min}^m$ for $m = \{A, V\}$, where $S_{\min}^m$ and $S_{\max}^m$ are the minimum and maximum of S_k^m, respectively.

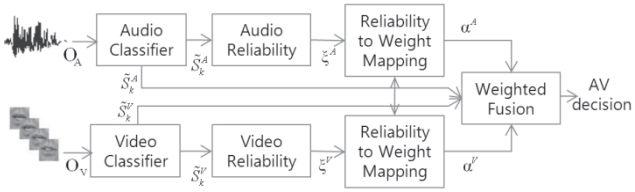

Fig. 1 Implemented baseline AVSI system.

Table 1 AVSI system features.

Modality	Feature parameters	Classifier
Audio (speech signal)	Pre-processing: Energy-based voice activity detection[8] Features: 17 Mel-Cepstrums with the cepstral mean subtraction [8]	GMM 3 mixtures [7]
Video (lip Image)	Input Image: 64×64 pixel Feature: Local principle component analysis 10 principle components [9]	GMM 3 mixtures [7]

c) Calculate the audio and video reliability values ξ^A and ξ^V based on the score dispersion [1], [2];

$$\xi^m = \frac{\text{Max}\,\tilde{S}_k^m - \text{Max2}\,\tilde{S}_k^m}{\text{Mean}\,\tilde{S}_k^m} \tag{1}$$

for $m = \{A, V\}$, where Max, Max2 and Mean are the maximum, second maximum and mean values of the normalized scores $\{\tilde{S}_k^m\}$, respectively.

d) Calculate the weighting values, α^A and α^V, for audio and video;

$$\alpha^m = \frac{\xi^m}{\xi^A + \xi^V}, \quad \text{for} \quad m = \{A, V\}. \tag{2}$$

e) Calculate the integrated score using the weighting values, α^A and α^V;

$$\tilde{S}_k^{AV} = \alpha^A \tilde{S}_k^A + \alpha^V \tilde{S}_k^V. \tag{3}$$

f) Finally, the identification process is completed by $arg\max_k \tilde{S}_k^{AV}$, where k is the speaker model index.

3. Proposed GFMF-VSR

Generally, in digital image processing, the grayscale image is obtained from the RGB color space in the form of the following luminance (Y) expression:

$$Y = 0.2989R + 0.587G + 0.114B.$$

Let us suppose that we have a grayscale image $f(y, x)$ typically termed as the intensity image, which has a 256 possible different shades of gray from black to white; where y and x represent the spatial co-ordinates, $y = 0, 1, 2, \ldots, M-1$ and $x = 0, 1, 2, \ldots, N-1$.

In a video-based speaker identification system, the identification process uses the images' characteristics of each speaker including the average intensity of the faces or lips. Therefore, in the installment stage, we can mathematically express the average intensity (μ) of an image as follows:

$$\mu_{kl} = \frac{1}{MN} \sum_{y=0}^{M-1} \sum_{x=0}^{N-1} f_{kl}^{tr}(y, x); \quad l = 1, 2, \ldots, L_k \tag{4}$$

$$\bar{\mu}_k = E(\mu_{kl}), \tag{5}$$

where f_{kl}^{tr} indicates the l-th training (tr) image of the k-th speaker in the installment/training stage, μ_{kl} represents the average intensity at the l-th image of the k-th speaker, and L_k is the total number of *k-th* speaker's utterances. $\bar{\mu}_k$ is the average image intensity for the k-th speaker. Similarly, we can express the average intensity of the left-half (LH) and right-half (RH) image as:

$$\mu_{kl}^{LH} = \frac{2}{MN} \sum_{y=0}^{M-1} \sum_{x=0}^{\frac{N}{2}-1} f_{kl}^{tr}(y, x); \quad l = 1, 2, \ldots, L_k \tag{6}$$

$$\mu_{kl}^{RH} = \frac{2}{MN} \sum_{y=0}^{M-1} \sum_{x=\frac{N}{2}}^{N-1} f_{kl}^{tr}(y, x); \quad l = 1, 2, \ldots, L_k \tag{7}$$

$$\bar{\mu}_k^{LHRH} = E(\mu_{kl}^{LH}) - E(\mu_{kl}^{RH}), \tag{8}$$

where μ_{kl}^{LH} and μ_{kl}^{RH} indicate the average intensity of the left-half and right-half images at the l-th image of the k-th speaker, respectively, and $\bar{\mu}_{kl}^{LHRH}$ denotes the average intensity difference between μ_{kl}^{LH} and μ_{kl}^{RH} of the L_k images. Subsequently, for measuring the intensity difference between the upper-half (UH) and down-half (DH) images we can write:

$$\mu_{kl}^{UH} = \frac{2}{MN} \sum_{y=\frac{M}{2}}^{M-1} \sum_{x=0}^{N-1} f_{kl}^{tr}(y, x); \quad l = 1, 2, \ldots, L_k \tag{9}$$

$$\mu_{kl}^{DH} = \frac{2}{MN} \sum_{y=0}^{\frac{M}{2}-1} \sum_{x=0}^{N-1} f_{kl}^{tr}(y, x); \quad l = 1, 2, \ldots, L_k \tag{10}$$

$$\bar{\mu}_k^{UHDH} = E(\mu_{kl}^{UH}) - E(\mu_{kl}^{DH}), \tag{11}$$

where μ_{kl}^{UH} and μ_{kl}^{DH} indicate the average intensity of the upper-half and down-half images at the l-th utterance of the *k-th* speaker, respectively and $\bar{\mu}_{kl}^{UHDH}$ denotes the average intensity difference between μ_{kl}^{UH} and μ_{kl}^{DH} of the L_k images. In the same way, we can write the following expression for an input image $f^{te}(y, x)$ of individual utterance at the testing (te) stage:

$$\mu_{te} = \frac{1}{MN} \sum_{y=0}^{M-1} \sum_{x=0}^{N-1} f^{te}(y, x) \tag{12}$$

$$\mu_{te}^{LH} = \frac{2}{MN} \sum_{y=0}^{M-1} \sum_{x=0}^{\frac{N}{2}-1} f^{te}(y, x) \tag{13}$$

$$\mu_{te}^{RH} = \frac{2}{MN} \sum_{y=0}^{M-1} \sum_{x=\frac{N}{2}}^{N-1} f^{te}(y, x) \tag{14}$$

$$\mu_{te}^{LHRH} = \mu^{LH} - \mu^{RH} \tag{15}$$

$$\mu_{te}^{UH} = \frac{2}{MN} \sum_{y=\frac{M}{2}}^{M-1} \sum_{x=0}^{N-1} f^{te}(y, x) \tag{16}$$

$$\mu_{te}^{DH} = \frac{2}{MN} \sum_{y=0}^{\frac{M}{2}-1} \sum_{x=0}^{N-1} f^{te}(y, x) \tag{17}$$

$$\mu_{te}^{UHDH} = \mu^{UH} - \mu^{DH} \tag{18}$$

In Eq. (12), f^{te} indicates the input image at the testing stage and μ_{te} represents the input image average intensity. In Eqs. (13), (14) and (15), μ_{te}^{LH} and μ_{te}^{RH} indicate the average intensity of the left-half and right-half input image, respectively, and μ_{te}^{LHRH} denotes the average intensity difference between μ_{te}^{LH} and μ_{te}^{RH} of the input image. In Eqs. (16), (17) and (18), μ_{te}^{UH} and μ_{te}^{DH} indicate the average intensity of the upper-half and down-half input image, respectively, and μ_{te}^{UHDH} denotes the average intensity difference between μ_{te}^{UH} and μ_{te}^{DH} of the input image.

The proposed VSR measure employs the lighting change indicators, i.e., the global and directional lighting variation using each set of intensity calculations noted in the above expressions and, consequently, the VSR values can be separately determined for each speaker. That is, the VSR measure is formulated using GFMF as in Eq. (19).

$$\xi_k^V = e^{-\left\{ \frac{(\mu_{te} - \bar{\mu}_k)^2}{2\sigma_1^2} + \frac{\left(\mu_{te}^{LHRH} - \bar{\mu}_k^{LHRH}\right)^2}{2\sigma_2^2} + \frac{\left(\mu_{te}^{UHDH} - \bar{\mu}_k^{UHDH}\right)^2}{2\sigma_3^2} \right\}}, \tag{19}$$

where ξ_k^V is the visual reliability for the k-th speaker's model, σ_1, σ_2 and σ_3 are the control parameters of the GFMF-VSR measure and should be optimized for the SI

performance maximization. According to Eq. (19), the VSR value is not unique through the speakers. It depends on the speaker index. Hence, the visual weighting value is determined separately for each speaker using the reliability as follows.

$$\alpha_k^A = \frac{\xi^A}{\xi^A + \xi_k^V} \tag{20a}$$

$$\alpha_k^V = \frac{\xi_k^V}{\xi^A + \xi_k^V} \tag{20b}$$

Finally, the integrated AV score is typically expressed as

$$\tilde{S}(O_A, O_V \mid \lambda_k) = \alpha_k^A \tilde{S}_k^A + \alpha_k^V \tilde{S}_k^V, \tag{21}$$

To optimize the parameters of σ_1, σ_2 and σ_3 in Eq. (19), we need to set an object function. In this paper, we adopt the identification rate as the target function in the following Eq. (22) to optimize the variances.

$$g(\sigma_1, \sigma_2, \sigma_3) = \frac{\sum_{k=1}^{K}\left(\sum_{l=1}^{L_k} \delta(arg\max_i(\tilde{S}_i(O_{Akl}, O_{Vkl} \mid \lambda_i)), k)\right)}{\sum_{k=1}^{K} L_k} \tag{22}$$

In the above function expressed by Eq. (22), $\delta(i, j)$ is the delta function, K is the number of the speakers, k is the speaker index, L_k is the total number of *k-th* speaker's utterances, l is the utterance index. And O_{Akl} or O_{Vkl} is the l-th audio or visual observations of the k-th speaker, i is the speaker model index and $\tilde{S}_i$ is the integrated score at given model λ_i for the given observations of O_{Akl} and O_{Vkl}. $arg\max_i \tilde{S}_i$ means the index of the speaker with the maximum score value. In Eq. (22), the integrated probability $\tilde{S}_i$ is the function of α_i^A and α_i^V, and α_i^V is related to σ_1, σ_2 and σ_3 by Eqs. (20) and (19). So, the target function defined through the identification rate is controlled by σ_1, σ_2 and σ_3. However, optimizing the object function is not a linear problem. Thus, we adopt the particle swarm optimization (PSO) algorithm [10], which is one of the well-known approaches for solving the nonlinear optimization problem. The PSO procedure is as follows:

a) Initialize S particle vectors randomly;

$$\Sigma_j = (\sigma_{1j}, \sigma_{2j}, \sigma_{3j}), \quad j = 1 \ldots S.$$

b) Calculate $g(\Sigma_j)$ and set the temporary best particle, Σ_{TB}, so that $g(\Sigma_{TB}) > g(\Sigma_j)$ and set the global best particle, $\Sigma_{GB} = \Sigma_{TB}$.

c) Update the particles $\{\Sigma_j\}$ according to the PSO rules [10].

d) Calculate Σ_{TB}, and if $g(\Sigma_{TB}) < g(\Sigma_{GB})$, then $\Sigma_{GB} = \Sigma_{TB}$.

e) Until a termination criterion is met, repeat c) and d).

After finishing the PSO process, Σ_{GB} are the optimization parameters for maximizing Eq. (22).

4. Experiments and Results

We performed the AVSI experiments using the VidTimit database containing 43 speakers and 10 utterances (U1∼U10) for each speaker. The lip ROI (region of interest) RGB color pixel blocks were converted to gray scale [0, 255] images of 64×64 pixel from the lip database. The lip database was manually created from the VidTimit video database based on the lip center. We partitioned the database into the following groups:

(a) DS1A: Audio U1∼U4
(b) DS1V: Video U1∼U3
(c) DS2: U5∼U7 (Audio & Video)
(d) DS3: U8∼U10 (Audio & Video)

DS1A & DS1V are used for training audio & video GMMs, respectively. Note that in the training stage, the data in audio and video are asynchronously selected. DS2 and DS3 are used for creating the GFMF-VSR model and evaluating the proposed method alternately.

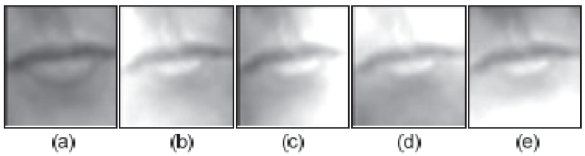

Fig. 2 Examples of lip image degradation due to lighting changes. (a) original image (b) illumination change from LR (c) illumination change from RL (d) illumination change from UD (e) illumination change from DU.

In order to create the global and directional lighting variations in the installment and testing images, an artificial illumination was added to the testing lip images in different directions. Suppose we have a lip image $w(y, x)$ to which illumination will be added. Using Eq. (23), an illuminated image $F(y, x)$ is obtained [11], [12].

$$F(y, x) = w(y, x) + \frac{-\gamma}{D} z + \gamma, \quad \text{for } y = 0, 1, 2, \ldots, M-1 \text{ and } x = 0, 1, 2, \ldots, N-1, \tag{23}$$

where z is either y or x axis direction depending on the illumination direction, γ controls the 'strength' of the artificial illumination, and D is the total length through x-axis ($N = 64$) or y-axis ($M = 64$) of a lip image. Examples of the ROI lip images with artificial illumination in different directions at $\gamma = 150$ and $D = 64$ are shown in Fig. 2. Four directions of the lighting variation i.e., left-to-right (LR), right-to-left (RL), up-to-down (UD) and down-to-up (DU) are taken into account; as a result, there were twelve testing video utterances at the testing stage. To synchronize with these visual data, we have shaped the audio utterances accordingly.

In our experiments, the values of L_k in Eqs. (4) to (11) are equals to three, since we have taken only three utterances in the training stage for visual model (λ^V) creation. For the VSR measure, we have employed only the first frame of each utterance at the training and test stages.

Table 2 shows the audio-based, lip-based and audio-visual score-based speaker identification performances while the testing images have illumination variations in the visual classifier system. In the experiments AV-SCM was implemented based on A.F. Niall's approach [2]. As

Table 2 Audio and video SI performance with the illumination change.

Modality	γ	Light Direction	Train data	Test data	Average SI rate (%)
Audio	-	-	DS1A	DS2,DS3	86.04
Video	150	LRRLUDDU	DS1V	DS2,DS3	91.08
AV-SCM	150	LRRLUDDU	DS1A DS1V	DS2,DS3	94.08

Table 3 AVSI performance with the proposed method.

Light Direction	Training data for variance values	Optimized Values			Test data	SI rate (%)
		σ_1	σ_2	σ_3		
LRRLUDDU	DS2	53.82	85.56	82.34	DS3	97.48
LRRLUDDU	DS3	56.24	92.05	85.65	DS2	98.26
Average	-	55.03	88.80	84.00	-	97.87

shown in Table 2, the previous SCM achieved better performance compared with the single modality-based SI systems. Table 3 shows the AVSI performances with the proposed VSR measure with the same experimental environments of Table 2. In the experiments, we set the audio reliability value as 1, as we did not add extra noises to the speech database even though the speech signal has a high level of distortion since the audio signal was collected in an office environment. The average AVSI rate of the proposed method is 97.87%, which is 3.79% higher than that of the score-based AVSI system. By adopting the VSR measure, the relative reduction of AVSI error rate is 64.02% while the score-based AVSI system is taken as baseline.

5. Conclusion

In this study, we proposed a VSR measure that can handle video distortion due to illumination change. With the AVSI experimental results, we confirmed that the proposed VSR measure, for estimating the influences of illumination change, is a promising solution in multimodal biometrics. In the future, we will develop a VSR to measure the morphological correctness of detected lips, and we will also study a combining method of the proposed VSR and the morphological reliability for AVSI in real environment.

Acknowledgments

This work is supported by the MKE, Korea, under the ITRC support program supervised by the IITA (IITA-2009-(C1090-0903-0008)) and the NRF of Korea Grant funded by the Korean Government (2009-0077345).

References

[1] T. Wark and S. Sridharan, "Adaptive fusion of speech and lip information for robust speaker identification," Digital Signal Process., vol.11, pp.169–186, 2001.
[2] A.F. Niall, Audio and Video Based Person Identification, Ph.D Thesis of University College Dublin, 2005.
[3] E. Erzin, Y. Yemez, and A.M. Tekalp, "Multimodal speaker identification using adaptive classifier cascaded based on modality reliability," IEEE Trans. Multimedia, vol.7, no.5, pp.840–852, 2005.
[4] M. Heckmann, F. Berthommier, and K. Kristian, "Noise adaptive stream weighting in audio-visual speech recognition," EURASIP J. Applied Signal Process., vol.11, pp.1260–1273, 2002.
[5] U.V. Chaudhari, G.N. Ramaswamy, G. Potamianos, and C. Neit, "Audio-visual speaker recognition using time-varying stream reliability," Proc. IEEE ICASSP'03, vol.5, pp.712–715, 2003.
[6] C. Sanderson, Biometric Person Recognition: Face, Speech and Fusion, VDM-Verlag, 2008.
[7] D.A. Reynolds and R.C. Ross, "Robust text-independent speaker identification using Gaussian mixture speaker models," IEEE Trans. Speech Audio Process., vol.3, no.1, pp.72–82, 1995.
[8] M. Woelfel and J. McDonough, Digital Speech Recognition, Wiley, 2009.
[9] N. Kambhatla and T.K. Leen, "Dimension reduction by local PCA," Neural Comput., vol.9, pp.1493–1503, 1997.
[10] R. Eberhart and J. Kennedy, "A new optimizer using particle swarm theory," Proc. 6th Int. Symposium on Micro Machine and Human Science, pp.39–43, 1995.
[11] C. Sanderson and K.K. Paliwal, "Fast features for face authetication under illumination direction changes," Pattern Recognit. Lett., vol.24, pp.2409–2419, 2003.
[12] C. Kotropoulos, A. Tefas, and I. Pitas, "Morphological elastic graph matching applied to frontal face authentication under well-controlled and real conditions," Pattern Recognit., vol.30, no.12, pp.1935–1947, 2000.