論文 / 著書情報
Article / Book Information

Title	Audio-visual person authentication using speech and ear images
Author	Koji Iwano, Tomoharu Hirose, Eigo Kamibayashi, Sadaoki Furui
Journal/Book name	Workshop on Multimodal User Authentication (MMUA 2003), Vol. , No. , pp. 85-90
発行日 / Issue date	2003, 12

Audio-Visual Person Authentication Using Speech and Ear Images

Koji Iwano, Tomoharu Hirose, Eigo Kamibayashi, and Sadaoki Furui

Tokyo Institute of Technology, Department of Computer Science
2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552 Japan
{iwano, hirose, kamibaya, furui}@furui.cs.titech.ac.jp

Abstract

This paper proposes a multimodal, biometric person authentication method using speech and ear images to attempt to improve the performance in mobile environments. It is well known that the performance of person authentication using only speech is deteriorated by acoustic noises and feature changes with time. Since the ear shape of each person does not change over time, integrating its image with speech information increases robustness of person authentication. Experiments are conducted using audio-visual database collected from 38 male speakers at five sessions over a half year period. Speech data are contaminated with white noise at various SNR conditions. Experimental results show that the authentication performance is improved by combining the ear image with speech in every SNR condition.

1. Introduction

The necessity of person authentication is spreading in the recent network society. Biometric authentication, which identifies an individual person using physiological and/or behavioral characteristics, such as face, fingerprints, hand geometry, handwriting, iris, retina, vein, and speech, is one of the most attractive and effective methods. These methods are more reliable and capable than knowledge-based (e.g., password) or token-based (e.g., a key) techniques, since biometric features are hardly stolen or forgotten.

Although "speech" is one of the most useful and effective features for person authentication in mobile environments, its performance deteriorates due to additive noise and session-to-session variability of voice quality. Therefore, the combination with other biometric features to improve the performance has attracted a great deal of attention. Along this line, various audio-visual biometric authentication methods have been proposed[1, 2, 3, 4, 5]. Although most of them use "face" information in combination with speech, the face features also change due to make-up, mustache, beard, hair styles and so on, and derives degradation of the performance. Therefore, it is worth investigating other biometric features with high permanence.

From this point of view, this paper proposes an authentication method using "ear" shape information in combination with speech. It is well known that the ear shape hardly changes over time[6, 7]. Although several authentication methods using ear images have already been proposed[7, 8, 9], there is no research on multimodal authentication using both speech and ear images. Since ear images could be captured using a small camera installed in a mobile phone, ear information is expected to be easily used in mobile environments than other biometrics, such as fingerprint, iris, and retinal, that need special equipment.

Our authentication method and audio-visual database are described in Section 2. Section 3 reports experimental results and Section 4 concludes this paper.

2. System structure and experiments

Figure 1 shows our multimodal person authentication system using speech and ear images. Audio and visual data are respectively converted into feature vectors. Each set of features is matched with both a claimed person model and a speaker independent (SI) model. Then, audio and visual scores are integrated with appropriate weighting and a decision is made whether he/she is a true speaker or an impostor. If the score is larger than a threshold value, the speaker is accepted as a claimed speaker.

2.1. Integrated score

A posterior probability is used as the authentication score. The posterior probability of being a claimed speaker S^c after observing a biometric feature set x, is denoted by $p(S^c|x)$. Since x is composed of speech (audio) features x_s and ear (visual) features x_e, $p(S^c|x)$ can be transformed as follows:

$$p(S^c|x) = p(S^c_s|x_s) \cdot p(S^c_e|x_e) \quad (1)$$

where S^c_s and S^c_e represent the claimed speaker's speech and ear models, respectively. Bayes' Rule derives the following

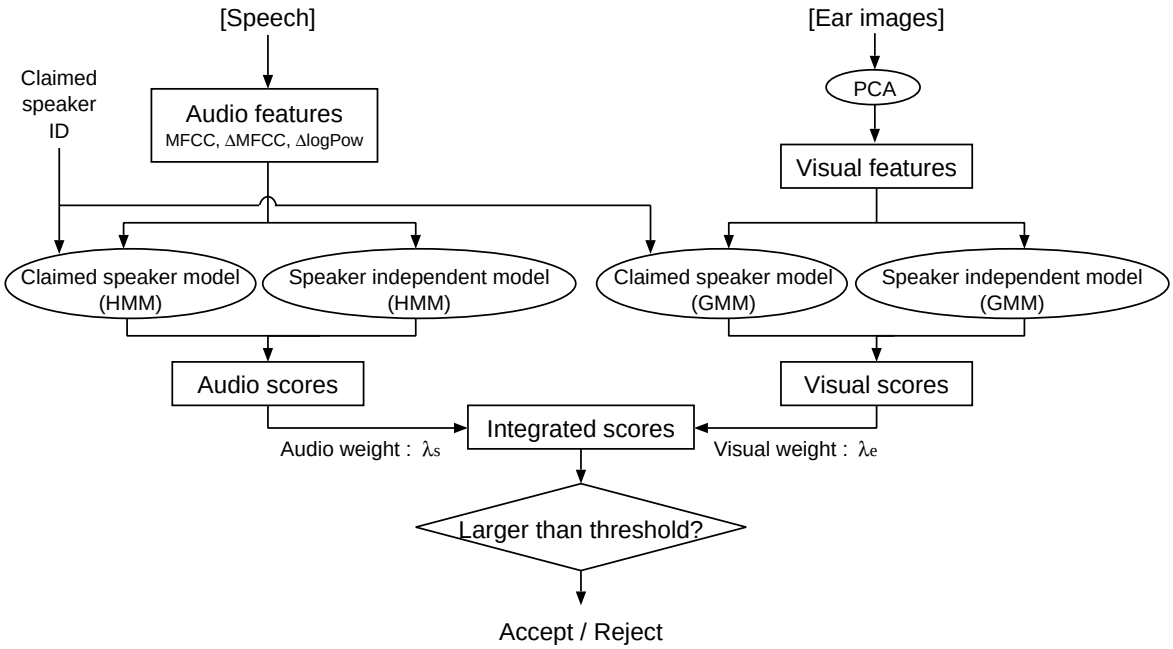

Fig. 1. Multimodal person authentication system using speech and ear images.

equation:

$$p(S^c|x) = \frac{p(x_s|S_s^c)p(S_s^c)}{p(x_s)} \cdot \frac{p(x_e|S_e^c)p(S_e^c)}{p(x_e)} \quad (2)$$

where $p(x_s|S_s^c)$ and $p(x_e|S_e^c)$ are likelihood values with claimed speaker's speech and ear models, respectively. The probabilities in the denominator are approximated by using likelihood values with general speaker's speech model $p(x_s|S_s^g)$ and ear model $p(x_e|S_e^g)$:

$$p(S^c|x) \approx \frac{p(x_s|S_s^c)p(S_s^c)}{p(x_s|S_s^g)p(S_s^g)} \cdot \frac{p(x_e|S_e^c)p(S_e^c)}{p(x_e|S_e^g)p(S_e^g)} \quad (3)$$

$$\propto \frac{p(x_s|S_s^c)}{p(x_s|S_s^g)} \cdot \frac{p(x_e|S_e^c)}{p(x_e|S_e^g)} \quad (4)$$

Equation (4) means that the posterior probability for the claimed speaker's model is calculated by the product of likelihood values normalized using speaker independent (SI) models. By defining authentication scores for speech (p_s) and ear (p_e) as

$$p_m = \log p(x_m|S_m^c) - \log p(x_m|S_m^g) \ \ (m = s, e) \quad (5)$$

an integrated score p_{se} which balances the effectiveness of speech and ear features can be modeled by the following equation.

$$p_{se} = \lambda_s p_s + \lambda_e p_e \ \ (\lambda_s + \lambda_e = 1) \quad (6)$$

where λ_s and λ_e are audio and visual weights, respectively.

2.2. Audio-visual database

2.2.1. Recording conditions

Audio-visual data were recorded at five sessions with intervals of approximately one month. The data were collected from 38 male speakers, and each speaker uttered 50 strings of four connected digits in Japanese at each session. Speech data were sampled at 16kHz with 16bit resolution. One right ear image for each speaker taken by a digital camera with 720×540 pixel resolution was collected at each session. Figure 2 shows the arrangement of a speaker and a camera when recording. An image of the whole ear, with no hair obscuring it, was captured by the camera positioned perpendicular to the ear. The camera was located approximately 20cm away from each speaker's ear. A flash was used to keep constant illumination.

2.2.2. Training and testing data

A set of data recorded at sessions 1∼3 was used for training and that recorded at sessions 4 and 5 was used for testing. The database was separated into two groups in terms of speakers as shown in Figure 3. This figure shows the case that the speaker #01 was used as the claimed speaker. The SI model was trained using the utterances by all the speakers in the speaker group B which did not include the claimed speaker. When one of the speakers in the speaker group B was used as the claimed speaker, the utterances by

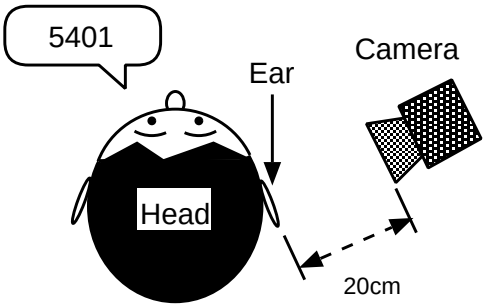

Fig. 2. Location of speaker and camera.

Speaker ID	Session 1,2,3	Session 4,5	
#01 : : #19	Uesd for speaker model	True speaker	Group A
		Impostors	
#20 : : #38	Used for speaker independent model	Impostors	Group B

Fig. 3. Training and testing data for the authentication experiment when the speaker #01 is the claimed speaker.

the speaker group A were used for the SI model training. In this way, the SI model was always trained using the data of a speaker group not including the claimed speaker. All the speakers in both speaker groups A and B, except for the claimed speaker himself, were used as imposters.

White noise was added to the audio data for training at 30dB SNR level to increase the robustness against noisy speech, and testing data were contaminated with white noise at 5, 10, 15, 20, and 30dB SNR conditions.

As image data, we first extracted gray-scaled ear images with 80×80 pixel resolution. An example of the extracted ear image is shown in Figure 4. The ear location and rotation in the image were manually adjusted. In order to increase robustness of visual models, the following variations were given to training data:

(1) Shifting the ear location in vertical and horizontal directions within ±6 pixels at a 2 pixel interval. Consequently, 49 variations were made for each ear image.

(2) Rotating the ear images within ±30 degrees at one degree interval. Accordingly, 61 variations were made for each ear image.

Fig. 4. An example of the extracted ear image.

The both operations made approximately 9,000 (= 3 sessions × 49 × 61) ear images for training each speaker's model. For testing data, we applied only the rotating operation (2).

Both training and testing data were filtered to emphasize the ear feature. The following three conditions were experimentally compared to find the best filtering method:

(a) No filtering (Figure 5(a)).

(b) Laplacian filtering (Figure 5(b)).

(c) Laplacian-Gaussian filtering (Figure 5(c)).

Finally, all ear images were circularly sampled and digitized for reducing hair effects and avoiding the window shape effects caused by rotation of the images.

2.3. Audio and visual features

Audio features were 25-dimensional vectors consisting of 12 MFCCs, 12ΔMFCCs, and Δ log energy. The frame shift was 10ms and the analysis window length was 25ms. For ear images, "eigen-ear" space was built by using Principal Components Analysis (PCA) in the same way as the eigenface approach used in face recognition[10]. The PCA was applied to the ear images recorded at the first session using 19 speakers in one of the two speaker groups that did not include the claimed speaker. The original ear images with no shifting or rotating were used for the analysis. Figure 6 shows examples of the first eight eigen-ear images obtained by the PCA using the Laplacian-Gaussian filtered images. All the ear images were converted into 18-dimensional visual feature vectors using the first 18 eigen-ears.

2.4. Speech and ear models

The audio features were modeled by digit-unit HMMs. Each digit HMM has a standard left-to-right topology with $n \times 3$ states, where n is the number of phonemes in the digit. The

(a) no filtering (b) Laplacian filtering (c) Laplacian-Gaussian filtering

Fig. 5. Examples of the filtered ear images.

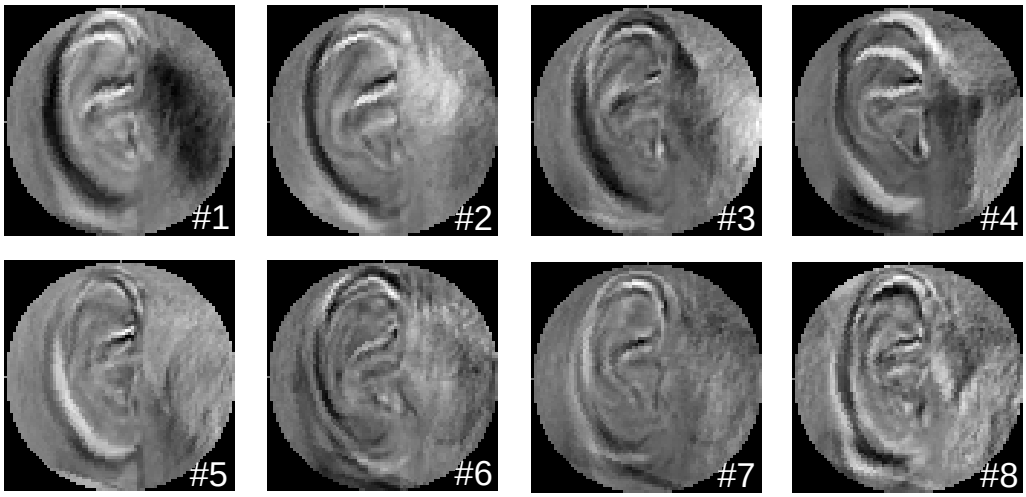

Fig. 6. Examples of the first 8 eigen-ear images.

authentication score for the speech features represented in Equation (4) is calculated as follows:

$$\frac{p(x_s|S_s^c)}{p(x_s|S_s^g)} = \frac{\Sigma_w p(x_s|S_s^c,w)p(w)}{\Sigma_w p(x_s|S_s^g,w)p(w)} \approx \frac{\max_w p(x_s|S_s^c,w)}{\max_w p(x_s|S_s^g,w)} \quad (7)$$

where w is a string of four connected digits.

The visual features were modeled using GMMs. In each testing experiment, 61-feature vectors converted from the rotated images were input to the GMMs. Log likelihood values calculated for the claimed speaker and the SI models were used to obtain the authentication score for each ear image according to the Equation (5).

3. Experimental results

3.1. Results of the authentication using ears

An experiment using only ear images was first conducted for investigating the effects of shifting and filtering the ear images. Table 1 shows equal error rates (EER) for the person authentication at various conditions of filtering and image processing applied to the training data. In the experiment, optimum numbers of mixtures: eight mixtures for speaker GMMs and one mixture for SI GMM, were experimentally chosen.

The results show that both filtering methods are effective for improving the authentication performance. The Laplacian-Gaussian filtering yields better results than the Laplacian filtering. The shifting operation for training data also improves the performance irrespective of filtering methods. This probably means that there are some mismatches of ear location between training and testing data due to the manual image extraction process.

Table 1. Equal error rate (%) in person authentication using ear images with various kinds of filtering and processing in the training stage.

	only rotating	shifting & rotating
no filtering	14.5	14.0
Laplacian filtering	13.6	13.3
Laplacian-Gaussian filtering	13.2	**11.9**

The best result, 11.9% EER, is observed at the condition using the Laplacian-Gaussian filtering and shifting as well as rotating operations. This condition is used in the following visual authentication experiments.

3.2. Results of the multimodal authentication

Multimodal authentication results in various SNR conditions obtained by using optimum audio weights (λ_s) are shown in Figure 7. The optimum weights (λ_s) were determined experimentally to minimize the error rate at each condition. The optimum values are also shown in the figure. Results using only speech ($\lambda_s = 1.0$) and only ear ($\lambda_s = 0.0$) are also shown for the purpose of comparison. The number of mixtures in audio HMMs was optimized based on the experimental results at the 30dB SNR condition; the number of mixtures was set to four for both speaker and SI HMMs.

Although the authentication performance using only speech is highly degraded by the noise effect, it is clearly shown that multimodal authentication is robust. The proposed method is most effective when the SNR is 15dB; the error rate is reduced by 53.0% from the audio only method and 43.9% from the visual only method. The best performance of 0.3% EER is observed at the 30dB SNR condition.

Figure 8 shows EER as a function of the audio weight (λ_s). Improvement using the ear images is observed over a wide range of λ_s. It is also shown that the optimum λ_s values exist in the range of $0.6 \sim 0.8$ at all the noise conditions with the exception of the 5dB SNR condition. This means that the proposed multimodal method is not sensitive to the change of weights and the weight can be easily optimized.

3.3. Comparing ears with faces as biometrics

We previously conducted person authentication experiments using speech and face features[5] in the similar way as that described in this paper. Although the speech and face database has the same number of speakers and recording sessions as the speech and ear database, 38 male speakers and 5 sessions, actual speakers are different between the two databases.

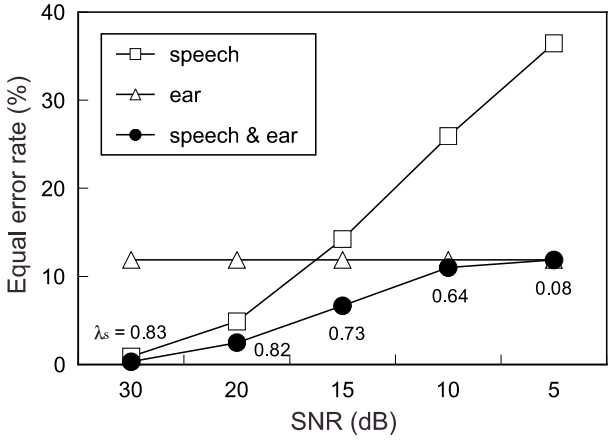

Fig. 7. Person authentication results in various SNR conditions.

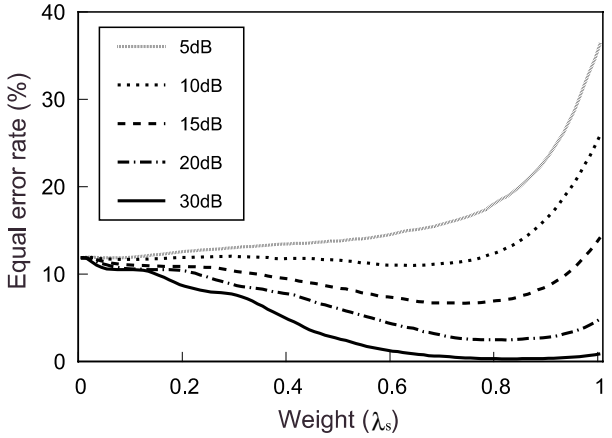

Fig. 8. Equal error rate as a function of the audio weight λ_s.

The previous work showed that the EER using only the face information was 7.0%, which was better than the EER using the ear information, 11.9%.

One of the reasons is that ear images are more changeable than face images by a tilt of the camera, since the ear surface is more irregular than the face surface. However, since the ear itself is not as changeable as the face, the authentication using ear biometrics has a possibility to become a practical method, if the above observation problem can be solved.

4. Conclusions

This paper has proposed a multimodal authentication method using the combination of speech and ear images with the aim of increasing noise robustness in mobile environments. The proposed method has been confirmed to be more robust than the speech only method in various SNR conditions.

Future works include 1) improving the authentication performance using the ear information by increasing the robustness against ear image variation caused by a tilt of a camera, 2) reducing the effects of hair and sideburns, 3) developing an automatic method for ear area detection, and 4) investigating the robustness of ear features against their changes over time.

5. Acknowledgements

This research has been conducted in cooperation with NTT DoCoMo. The authors wish to express thanks for their support.

6. References

[1] R. Brunelli and D. Falavigna, "Personal identification using multiple cues," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol.17, no.10, pp.955–966, Oct. 1995.

[2] B. Duc, E.S. Bigun, J. Bigun, G. Maitre, and S. Fischer, "Fusion of audio and video information for multi modal person authentication," *Pattern Recognition Letters*, vol.18, no.9, pp.835–843, Sept. 1997.

[3] P. Jourlin, J. Luettin, D. Genoud, and H. Wassner, "Acoustic-labial speaker verification," *Pattern Recognition Letters*, vol.18, no.9, pp.853–858, Sept. 1997.

[4] N. Poh, and J. Korczak, "Hybrid biometric person authentication using face and voice features," in *Audio- and Video-Based Biometric Person Authentication, Third International Conference, AVBPA 2001*, J. Bigun and F. Smeraldi, Eds., pp.348–353, Springer, 2001.

[5] T. Hirose, K. Iwano, and S. Furui, "Multi-modal speaker verification using speech and face images," *Proc. ASJ Spring Meeting 2003*, vol.1, pp.107-108, March 2003. (In Japanese)

[6] A. Iannarelli, *Ear Identification*. Forensic Identification series. Paramont Publishing Company, Fremont, California, 1989.

[7] M. Burge and W. Burger, "Ear biometrics," in *Biometrics: Personal Identification in Networked Society*, A. Jain, R. Bolle, and S. Pankanti, Eds., pp.273–285, Kluwer Academic, Boston, MA, 1999.

[8] N. Tashiro, K. Shinohara, M. Abe, and T. Okamura, "Individual identification by outline of components on pinna and superposition," *ITE Tech. Rep.*, MIP2001-54, vol.25, no.22, pp.7–13, March 2001. (in Japanese)

[9] Y. Wang, K. Takeda, K. Sato, and S. Nakayama, "Study on human recognition by ear image with eigenear," *Tech. Rep. of IEICE*, IE2002-95, vol.26, no.76, pp.37–42, Nov. 2002. (in Japanese)

[10] M. Turk and A.P. Pentland, "Face recognition using eigenface," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp.586–591, 1991.