International Journal of Latest Research in Science and Technology **ISSN (Online):2278-5299**
Volume 2, Issue 6: Page No.26-31,November-December 2013
https://www.mnkpublication.com/journal/ijlrst/index.php

AN ENHANCED MULTIMODAL SOUND LOCALIZATION WITH HUMANLIKE AUDITORY SYSTEM FOR INTELLIGENT SERVICE ROBOTS

[1]Keun-Chang Kwak
[1]Department of Control, Instrumentation, and Robotic Engineering, Chosun University, Korea

Abstract*- *This paper is concerned with an enhanced multimodal sound localization with humanlike auditory system for a network-based intelligent service robot, which exploits strong information technology infrastructure. The objective of this paper is to integrate several audiovisual-based Human-Robot Interaction (HRI) components that can naturally interact between human and robot through audiovisual information obtained from robot camera and microphones in the noisy environments or the presence of multiple persons. The proposed approach is comprised of two main stages. The first stage performs speech recognition, sound localization, and speaker recognition to know whether the user calls the robot or not as well as the direction and identification of the caller respectively, when someone calls robot's name. In the second stage, an intelligent robot moves forward to the specific caller based on multiple face detection/recognition with the aid of the information identified by speaker recognition among multiple persons. The robot platform used in this study is WEVER, which is a network-based intelligent service robot developed in Electronics and Telecommunication Research Institute. This robot refers to an Ubiquitous Robotic Companion (URC) that provides necessary services anytime and anywhere. The effectiveness of the proposed approach is compared with other multimodal methods and sound localization itself.

***Keywords* –** Multimodal Sound Localization, Human-Robot Interaction, Ubiquitous Robotic Companion, Speech/Speaker Recognition

I. INTRODUCTION

Nowadays, we have been witnessing a rapid growth in the number and variety of applications of robots, ranging from conventional industrial robots to intelligent service robots. Conventional industrial robots perform simple tasks by following pre-programmed instructions for humans in factories. These robots have been widely used in many manufacturing industries. On the other hand, intelligent service robots consist of toy robots, cleaning robots, humanoid robots and information service robots. In the case of the cleaning robots, the American company iRobot recently sold more than one million vacuum cleaner robots called Roomba. In Korea, the cleaning robot has drawn the attention of the people, and further support and development are concentrated on the information service robot. Especially a network-based intelligent service robot is regarded as a ubiquitous robotic companion that provides necessary services anytime and anywhere for information service [1]. Here this concept exploits strong Information technology infrastructure such as the high-speed internet. In order to develop IT-based service robot project based on this concept as one of the next-generation growth engine industries, it is necessary to develop core technologies such as Human-Robot Interaction (HRI), navigation, operation control, and so on. Among various core technologies, especially HRI components play an important role that can naturally interact between human and robot based on audiovisual information obtained from robot camera, microphones, and various sensors for network-based intelligent service robots. In this paper, we elaborate on an enhanced multimodal sound localization with humanlike auditory system in the noisy environments and the presence of multiple persons. Over the past few years, several studies have been completed on sound localization. Choi [2] has developed an audio-visual integration based on a probabilistic sound localization system by Time Delay of Arrival (TDOA) and face tracking by Open Computer Vision (OpenCV). Especially, when several face images are detected within robot's Field of View (FOV), the disadvantage of these approach is that the robot moves forward to person with large face image among rectangle bounds found by face detection. Hara [3] has proposed a robust speech interface based on audio and video information fusion by Bayesian network for humanoid HRP-2. Here sound localization was performed by the Multiple Signal Classification (MUSIC) method extended to the broadband signal with eigenvalue weighting. Human tracking includes face detection by skin-color model and face tracking by kernel-based face model, respectively. Huang [4] has presented a model-based sound localization system based on a model of the precedence effect of the human auditory system to cope with echoes and reverberations. The features of this system are high time resolution, less computation and correct localization in comparison to beamforming [5], MUSIC [6], and Maximum Likelihood (ML) method [7]. Valin [8] has proposed a robust sound localization method in three-dimensional space using an array of eight microphones. This method is based on a frequency-domain implementation of a steered beamformer along with a probabilistic post-processor. Nakadai [9] has used two pairs of microphone in a humanoid robot called SIG. One pair is installed on both sides of the head, while the other pair is placed inside the

Publication History
Manuscript Received : *25 December 2013*
Manuscript Accepted : *28 December 2013*
Revision Received : *29 December 2013*
Manuscript Published : *31 December 2013*

head to record internal sounds such as motor noise for noise cancellation.

Even though these methods have been revealed a good localization performance, the development of sound localization with humanlike auditory system remains as a matter to be researched further under robot environments. These systems may encounter problems when the sound localization is performed without the aid of speech/speaker recognition as well as multiple face recognition under the noisy environments or the presence of multiple persons. For this purpose, we elaborate on the three main aspects for an enhanced multimodal sound localization. First we use a multimodal approach based on the integration of speech recognition, sound localization, and speaker recognition to know whether the user calls a robot or not, as well as the position and identification of the caller simultaneously, when someone calls robot's name. Here sound localization is based on Generalized Cross Correlation-Phase Transform (GCC-PHAT) method. The speech and speaker recognition are performed by Hidden Markov Model (HMM) and Mel-Frequency Cepstral Coefficient-Gaussian Mixture Model (MFCC-GMM) for the network-based intelligent service robots, respectively. Second, the robot moves forward to the specific caller based on multiple face recognition with the aid of the prior information by speaker recognition among multiple persons after turning around. Here the face detection is performed by Revised and Modified Census Transform (RMCT), Adaboost, and face certainty map in the step of preprocessing, detection, and postprocessing, respectively. Furthermore, multiple face recognition has been developed by Multiple Principal Component Analysis (MPCA) and Support Vector Machine (SVM) method. Finally these two aspects should be performed on the concept of URC that provide various advanced functions and services by adding the network to the existing robot, enhancing mobility and user interface. Thus the sound localization is performed in robot body (client) equipped with three microphones and sound board. On the other hand, the speech/speaker recognition and face detection/recognition are performed on audiovisual information transmitted through wireless network. Over the past few years, we have been developed and implemented these HRI components for network-based intelligent service robot [10][11][12][13][14][17]. The robot platform used in this study is WEVER, which is a URC-based intelligent service robot developed at Intelligent Robot Research Division in Electronics and Telecommunication Research Institute. The experimental results obtained for sound localization database reveal that the proposed approach presented in this paper yield a good localization in comparison to the results obtained by other multimodal approaches and sound localization itself.

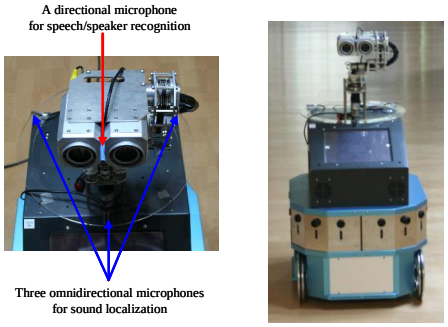
This paper is organized in the following manner. Section 2 describes the developed audio- and video-based HRI components for an enhanced multimodal sound localization with humanlike auditory system. Section 3 presents two main stages of multimodal sound localization under network-based environments. Section 4 covers the experimental results concerning these sound localization and the related HRI components. Finally concluding comments are covered in Section 5.

II. AUDIOVISUAL HUMAN-ROBOT INTERACTION COMPONENTS

In this section, we describe the audiovisual HRI components used in conjunction with multimodal sound localization. As mentioned before, the proposed approach is realized by sound localization, speech/speaker recognition, and multiple face detection/recognition to perform humanlike auditory and visual system. In confronting real-world environments, it is frequently advantageous to use several audiovisual techniques synergistically rather than exclusively, resulting in construction of multimodal sound localization system.

2.1 Sound localization

As shown in Fig. 1, we use low-priced three microphones and sound board equipped with the robot for sound localization. Here sound board with eight channels called multimodal interface module has already developed in ETRI.



(a) Arrangement of microphones (b) Wever robot

Fig.1 Arrangement of microphones and robot platform

We use GCC method in the frequency domain to obtain time delay. The sound localization based on GCC-PHAT in the frequency domain has the merit that this method can estimate relatively correct localization angle in the noisy and reverberant environments. Given that $m_1(n)$ and $m_2(n)$ are sound sources obtained from the first and second microphone, respectively, the time delay between $m_1(n)$ and $m_2(n)$ in the frequency domain is computed as follows

$$R_{m_1 m_2}(n) = \frac{1}{2\pi}\int_{-\infty}^{\infty} W_{PHAT}(w) M_1(w) M_2^*(w) e^{jwn} dw \quad (1)$$

where $M_1(w)$ and $M_2(w)$ are Fourier transform of $m_1(n)$ and $m_2(n)$, respectively. $W_{PHAT}(w)$ is the inverse of $M_1(w)M_2^*(w)$ as the frequency weighting function called PHAT. However, it is usually difficult to obtain the detailed prior knowledge of noise spectra in the reverberant environments. Thus the alternative is calculated as following PHAT weighting function

$$W_{PHAT}(w) = \frac{1}{\left|M_1(w) M_2^*(w)\right|} \quad (2)$$

The resultant time delay based on GCC-PHAT is expressed as follows

$$D = arg\,max\, R_{m_1 m_2}(n) \tag{3}$$

After computing time delay, we decide the section between microphones based on time delay. Suppose that the sound wave at the microphone is a plan wave and the angle between microphones is 120° , respectively. Finally azimuth θ is estimated as follows [10]

$$\theta = cos^{-1}\left(\frac{\Delta t\, v}{L}\right) - 30 \tag{4}$$

where θ is the angle of sound source obtained in the above equation. Δt is time delay between two microphones and v is the velocity of sound source. Moreover, L is the distance between two microphones. The sound localization is performed through the robot's name. This system is effectively combined with speech and speaker recognition to construct humanlike auditory system.

2.2 Speech/Speaker Recognition

In what follows, we present the techniques developed in conjunction with the speech and speaker recognition. On the basis of these methods, the robot can simultaneously recognize whether the user calls a robot or not as well as identification of the caller. We perform speech and speaker recognition with directional microphone from speech signal transmitted through wireless network, as shown in Fig.1. In the case of speech recognition, Korean-based spontaneous recognition is performed by HMM method. As a preprocessing process, we employ endpoint detection algorithm based on log energy and zero crossing rates as well as speech enhancement method based on Winer filter. For further details on speech recognition, see [11]. On the other hand, speaker recognition is comprised of four main stages including on-line speaker registration, feature extraction, generation of speaker model, and text-independent speaker identification. After detecting speech signal by endpoint detection algorithm, we obtain the feature vectors of MFCC. For simplicity, we use 11 order MFCC parameters except for the first order. This method is comprised of six stages to get MFCC: pre-emphasis, frame blocking, hamming widow to lessen distortion, Fast Fourier Transform (FFT), triangular bandpass filter, and cosine transform. In what follows, we briefly describe the well-known GMM frequently used in conjunction with text-independent speaker identification. Here the distribution of feature vectors extracted from individual speeches is obtained by a Gaussian mixture density. For a feature vector denoted as **x**, the mixture density for speaker is defined as

$$P(\vec{x} \mid \lambda_s) = \sum_{i=1}^{K} w_i\, b_i(\vec{x}) \tag{5}$$

where w_i is mixture weights and b_i is i'th Gaussian mixture. The density is a weighted linear combination of K Gaussian mixture parameterized by a mean vector and covariance matrix. The Gaussian mixture is defined as the following equation

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2}\left|\Sigma_i\right|^{1/2}} \times exp\left(-\frac{1}{2}(\vec{x}-\mu_i)^T(\Sigma_i)^{-1}(\vec{x}-\mu_i)\right) \tag{6}$$

The mixture weights w_i satisfy the constraint $\sum_{i=1}^{K} w_i = 1$. Thus the parameters of speaker's model are denoted as $\lambda_s = \{w_i, \mu_i, \Sigma_i\},\ i = 1,\ldots,K$. For simplicity, diagonal covariance matrices are used to construct GMM, because diagonal matrix are more computationally efficient than full covariance matrix for training [15][16]. Given training speeches from a speaker, the speaker model is trained by estimating the parameters of the GMM. The well-known method is ML estimation. For a sequence of T training vectors $X = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_T\}$, the GMM likelihood can be expressed as

$$p(X \mid \lambda_s) = \prod_{t=1}^{T} p(\vec{x}_t \mid \lambda_s) \tag{7}$$

The maximum likelihood parameter estimation is obtained by using the Expectation-Maximization (EM) algorithm. For speaker identification, a group of speakers $S = \{1,2,\ldots,S\}$ is represented by GMM's parameters. The main goal is to find the speaker model which has the maximum a posteriori probability as the following equation

$$\hat{S} = arg \max_{1 \le k \le S} p(\lambda_k \mid X) = arg \max_{1 \le k \le S} \frac{p(X \mid \lambda_k)\, p(\lambda_k)}{p(X)} \tag{8}$$

where the second equation is based on Bayes' rule. Assuming equally likely speakers, p(X) is the same for all speaker models, the identification is simplified as follows

$$\hat{S} = arg \max_{1 \le k \le S} p(X \mid \lambda_k) \tag{9}$$

Using logarithms and the independence between observations, the speaker identification is computed as

$$\hat{S} = arg \max_{1 \le k \le S} \sum_{t=1}^{T} log\, p(\vec{x}_t \mid \lambda_k) \tag{10}$$

On the other hand, on-line speaker registration is performed by Universal Background Model (UBM). Here each speaker is registered by a few sentences based on the adaptation of UBM in real-time [12].

2.3 Face Detection/Recognition

After performing sound localization, we use multiple face detection/recognition to compensate localization error and realize humanlike visual system. In general, face detection/recognition is used in multimodal sound localization. However, it is difficult to find the specific caller among multiple persons, because the robot does not know who the caller is. To solve this problem, we employ the prior information obtained by speaker recognition. Furthermore, multiple face recognition is used to identify multiple persons

shown in the robot's FOV. Here face detection is comprised of three steps including preprocessing, detection, and postprocessing. In the first stage, we revise the modified census transform to compensate the sensitivity to the change of pixel values. The second stage performs Adaboost that constructs the weak classifier which classifies the face and nonface patterns and the strong classifier which is the linear combination of weak classifiers. The last stage performs face certainty map based on facial information such as facial size, location, rotation, and confidence value to reduce False Acceptance Rate (FAR) with constant detection performance.

On the other hand, we use a MPCA and SVM for face recognition[13][14]. The most representative recognition technique frequently used in conjunction with face recognition is PCA. The PCA approach, also known as eigenface method, is a popular unsupervised statistical technique that supports finding useful image representations. It also exhibits optimality when it comes to dimensionality reduction. The use of the MPCA in this setting is motivated by its insensitivity to variation in comparison to PCA itself. This method consists of preprocessing, feature extraction, and identification. In the preprocessing, we select the face region such as eigenface, eigenUpper, and eigenTzone for multiple PCA. Furthermore, a geometric and photometric normalization is used to adjust the location of facial features and improve the quality of the face image, respectively. In the feature extraction, the weight and edge distribution vectors are obtained.

III.AN ENHANCED MULTIMODAL SOUND LOCALICATION

In this section, we present the integration of audiovisual HRI components for an enhanced multimodal sound localization. The proposed approach is comprised of two main stages. The first stage uses the integration of audio-based HRI components including sound localization and speech/speaker recognition to possess humanlike auditory system under network environments. The second stage of the approach concerns the integration of video-based HRI components including multiple face detection/recognition to have humanlike visual system and compensate the localization error. In what follows, we describe the procedure to integrate audio- and video-based HRI components for the enhanced multimodal sound localization as follows

[Step 1] To develop the enhanced multimodal sound localization, firstly the speaker and face registration are performed through wireless network in real-time. The speaker model is generated by a simple few sentences due to the adaptation of UBM. The face images are registered by continuous five frames obtained from robot camera.

[Step 2] When the user calls the robot's name, the speech signal is transmitted to server through a directional microphone used for speech/speaker recognition. The use of the robot's name can effectively interact between human and robot with respect to human auditory system.

[Step 3] The speech and speaker recognition are simultaneously performed through the transmitted signal in the server. Robot can simultaneously recognize whether the user calls the robot or not and who the caller is, in contrast to the conventional intelligent robot performed by only speech recognition. Furthermore, this approach can applied to user-customized service such as daily life schedule and favorite TV channel selection. On the other hand, the sound localization is performed with signals obtained from three microphones in the robot. The server informs the robot of the information concerning the start and end point of the transmitted signal.

[Step 4] The section decision and azimuth are obtained from sound localization based on GCC-PHAT. If the name of robot is recognized by speech recognition, the server brings the result of the localization angle computed in the robot body. Thus the robot can know the information concerning the identification and direction of user.

[Step 5] The server adjusts the rotation angle by the axis of robot wheels, if the circle center by three microphone is different from that of robot wheels. The robot turns around by the estimated angle.

[Step 6] The robot transmits the image shown in the robot's FOV to the server after turning. After that, face detection is performed as mentioned before. If the face images are not detected in the robot's FOV, the robot requests that the user calls the robot's name again.

[Step 7] After detecting the face images, multiple face recognition is performed. The robot moves forward to the specific caller based on face recognition with the aid of the priori information identified by speaker recognition among multiple persons. The proposed approach outperforms the conventional multimodal methods and sound localization itself, in which the intelligent robot can recognize whether a member of family calls myself or not as well as the direction and identification of the specific member

IV. EXPERIMENTAL RESULTS

In the series of experiments, we report on the development and performances of multimodal sound localization and the related HRI components. Firstly we evaluate the performance of the presented speaker recognition system for speaker database constructed by Intelligent Robot Research Division in ETRI. The database is constructed by audio recording of 20 speakers (4 females and 16 males). The data set consists of 30 sentences for each speaker and channel. We divide the speech data into training (20 sentences $\times$ 20 people $\times$ 3 distances (1, 2, 3m)) and test data sets (10 sentences $\times$ 20 people $\times$ 5 distances (1,2,3,4,5m)). The speaker database set consists of 2200 sentences. The recording was done in an office environment. The audio is stored as a mono, 16bit, 16kHz, and WAV file. Based on this database, text-independent speaker identification under robot environments was performed.

The identification performance obtained about 95% (1~3m), 87% (4m), 85%(5m) recognition rates from the variation of several mixture size and distances.

Furthermore, the results obtained from the variation of

GMM size showed a similar performance in this experiment. If the number of family is limited to 10 persons for intelligent home service robot, the identification performance will be further improved. On the other hand, the size of an image 320 x 240 pixel array whose gray levels ranged between 0 and 255. Among various face databases, we specifically considered ETRI face database constructed by Intelligent Robot Research Division for face recognition. This database consists of 2100 samples from 105 individuals. There are 20 images per subject. Here we divided into 5 images for enrollment and the remaining 15 images for test set. Fig. 2 visualizes the results of recognition performance. As shown in Fig. 2, the experimental results revealed that the proposed method yield a better classification performance in comparison to the results produced by the eigenface (L_2 and cosine similarity) and edge distribution.

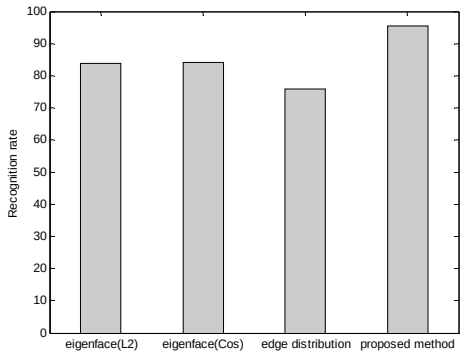


Fig. 2 Comparison of recognition rates

In what follows, sound localization was performed by GCC-PHAT method as mentioned in Section 2. The distance and angle between microphones are 0.32 meter and 120 degree, respectively. The database for sound localization contains 1800 sound sources recorded at every 30 degree and each distance (2, 3, 5m). The sound source used in this paper is the name of robot called WEVER. These speeches were recorded in the office environments. The error of the maximum location is 11.27 degree at 0 location point. Furthermore, the error of the average location is 4.85 degree at 2 meter. However, the location performance became worse due to the attenuation of sound as the distance increases.

The error of average location at 5 meter is about 15 degree. Because these location errors exist within the robot's FOV ($\pm 25^{\circ}$), the localization error can be compensated in the multimodal sound localization with the aid of face detection and recognition. Here the result of face detection showed a good detection rates in comparison to the well-known methods. The number of false detection for the proposed method and Viola-Jones is 3 and 78 on CMU+MIT frontal face test set, respectively. Therefore, if the face images exist within the robot's FOV after turning around, the proposed approach could yield a good performance close to 0 degree. Furthermore, the multimodal sound localization proposed in this paper can be effectively used for network-based intelligent service robots, although the experimental results obtained by only sound localization showed a worse localization performance in the noisy and reverberant environments.

V. CONCLUSIONS

We have discussed the enhanced multimodal sound localization with the aid of speech/speaker recognition, sound localization, and multiple face detection/recognition. It has been experimentally demonstrated that the proposed approach leads to a better localization and humanlike auditory system in comparison to the previous multimodal methods and sound localization itself. The main characteristics of this paper can be summarized as follows: 1) The enhanced multimodal sound localization was developed as the one of HRI components that is realized by URC environments for network-based intelligent service robots. 2) Based on speech/speaker recognition, the intelligent robot can recognize whether the user calls myself or not as well as the identification of the caller simultaneously, when he/she calls robot's name. 3) The robot can move forward to the caller based on multiple face recognition with the aid of the information identified by speaker recognition among multiple persons. The integration of these complementary approaches, together with certain audiovisual HRI components, results in the enhanced multimodal sound localization with humanlike auditory and visual system.

REFERENCES

[1] Y. G. Ha, J. C. Sohn, Y. J. Cho, and H. Yoon, "Towards ubiquitous robotic companion: design and implementation of ubiquitous robotic service framework", *ETRI Journal*, vol. 27, no. 6, pp. 666-676, 2005.

[2] J. S. Choi, M. Kim, and H. D. Kim, "Probabilistic speaker localization in noisy environments by audio-visual integration", *Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 4704-4709, Beijing, Oct., 2006.

[3] I. Hara, F. Asano, Y. Kawai, F. Kanehiro, and K. Yamamoto, "Robust speech interface based on audio and video information fusion for humanoid HRP-2", *Proceedings of the 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp.2404-2410, Sendai, Sep., 2004.

[4] J. Huang, T. Supaongprapa, I. Terakura, F. Wang, N. Ohnishi, and N. Sugie, "A model-based sound localization system and its application to robot navigation", *Robotics and Autonomous Systems*, vol. 27, pp. 199-209, 1999.

[5] D. H. Johnson and D. E. Dudgeon, *Array signal processing: concepts and techniques*, Prentice-Hall, Englewood Cliffs, 1993.

[6] R. O. Schmidt, "Multiple emitter location and signal parameter estimation", *IEEE Trans. on Antennas and Propagation*, vol.34, no.3, pp. 276-280, 1986.

[7] P. Stoica and K.C. Sharman, "Maximum likelihood method for direction-of-arrival estimation", *IEEE Trans. on Acoustics and Speech Signal Processing*, vol. 38, no.7, pp.1131-1143, 1990.

[8] J. M. Valin, F. Michaud, B. Hadjou, and J. Rouat, "Localization of simultaneous moving sound sources for mobile robot using a frequency-domain steered beamformer approach", *Proceeding of the 2004 IEEE International Conference on Robotics and Automation*, pp. 1033-1038, 2004.

[9] K. Nakadi, H. G. Okuno, and H. Kitano, "Real-time sound source localization and separation for robot audition", *In Processing IEEE International Conference on Spoken Language Processing*, pp. 193-196, 2002.

[10] J. Choi, J. Lee, S. Jeong, K. C. Kwak, S. Y. Chi, M. Hahn, "Multimodal sound source localization for intelligent service robot", *The 3rd International Conference on Ubiquitous Robots and Ambient Intelligence (URAI06)*, 2006.

[11] H. S. Lee, "Spontaneous dialogue recognition with large out-of-vocabularies", *IEEE Asia Pacific Conference on circuits and systems*, pp. 247-251, 1996.

[12] S. Kim, M. Ji, H. Kim, K. C. Kwak, and S. Y. Chi, "Text-independent speaker recognition for ubiquitous robot companion", *The 3rd International Conference on Ubiquitous Robots and Ambient Intelligence (URAI06)*, 2006.

[13] D. H. Kim, H. S. Yoon, S. Y. Chi, Y. J. Cho, "Face identification for human robot interaction: intelligent security system for multi-user working environment on PC", *The 15th IEEE International Symposium on Robot and Human Interactive Communication (ROMAN06)*, pp. 617-622, 2006

[14] D. H. Kim, J. Y. Lee, H. S. Yoon, H. J. Kim, Y. J. Cho, and E. Y. Cha, "A vision-based user authentication system in robot environments by using semi-biometrics and tracking", *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2005)*, pp. 246-251, 2005.

[15] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83. 1995.

[16] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19-41, 2000.

[17] K. C. Kwak, K. D. Ban, K. S Bae, H. J. Kim, S. Y. Chi, and Y. J. Cho, "Speech-based Human-Robot Interaction Components for URC intelligent service robots", *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2006)*, Video session, 2006.