# Text-Independent Speaker Recognition for Ubiquitous Robot Companion

Sungtak Kim*, Mikyong Ji*, Hoirin Kim*, Keun-Chang Kwak**, and Su-Young Chi**
*Information and Communications University, Korea
**Electronics and Telecommunications Research Institute, Korea
{stkim,lindaji,hrkim}@icu.ac.kr, {kwak,chisy}@etri.re.kr

***Abstract*** - This paper describes text-independent speaker recognition system which is basic and essential for interaction between users and the robot in Ubiquitous Robot Companion environment. For comfortable interaction between users and the robot, we implement an online speaker enrollment module which trains a GMM for each speaker model and text-independent speaker recognition module which consists of text-independent speaker identification and verification. The task of this system is to recognize one of the family members consisting of four to five persons. A user-specific service can be provided from the results of speaker recognition. Finally, the speaker recognition system is evaluated on speech database collected in URC environment with various distances, and we discuss the results.

***Keywords*** - Online Speaker Enrollment, Text-Independent Speaker Recognition, GMM, SCR

## 1. Introduction

The vision and hearing capabilities of robot play an important role in compensating each other in communications with real world. The visual information provides geometric information to robot. However, the process of this information needs lots of computation power and visual information is much sensitive to surrounding lighting. On the other hand, the hearing capability of robot has a merit of perceiving sound in no-light place or distant place. In addition, the hearing capability of robot plays an important role of protecting robot and human from dangerous situation because a dangerous signal is roaring sound in many cases. Speaker recognition is the first step of interaction between users and robot and the essential technique for response of robot to user's calling.

Speaker recognition, which can be classified into identification or verification, is the process of automatically recognizing who is speaking on the basis of individual information included in speech waveform. Speaker recognition methods can be also divided into text-dependent and text-independent methods. The former requires the speaker to say key words or sentences given to the same text for both training and recognition trials, and the latter does not rely on a specific text being spoken. However, text-dependent or text-independent speaker recognition system has the drawback that it would be defeated by playback of recoded voice of the genuine speaker. To overcome this problem, a text-prompted technique is now commonly adopted [1].

In this paper, we introduce a speaker enrollment module which is for training speaker model and then text-independent speaker recognition method which is used in user interface for robot in URC environment. The task of this speaker recognition system is to recognize one of family members within 3 meters distance, and to provide user-specific service by using the results of speaker recognition.

The remaining part of this paper is organized as follows. Section 2 describes online speaker enrollment and speaker recognition system in URC environment, and experimental results are given and discussed in Section 3. Finally, conclusions and further works are presented in Section 4.

## 2. Speaker Recognition System in URC environment

### 2.1 Online Speaker Enrollment

The purpose of online speaker enrollment is to train the user model and to register this model. Here GMM is used for speaker model. The goal of speaker model training is to estimate the parameters of GMM, which in some sense best matches the distribution of the training feature vectors. Although there are several techniques available for estimating the parameters of a GMM, the most popular and well-established method is Maximum Likelihood (ML) estimation.

The aim of ML estimation is to find the model parameters which maximize the likelihood of the GMM, given the training data. For the sequence of $T$ training vectors $X = \{x_1, x_2, ..., x_T\}$, the GMM likelihood can be represented by

$$P(X \mid \lambda_S) = \prod_{t=1}^{T} p(x_t \mid \lambda_S) \quad (1)$$

This expression is a nonlinear function of the parameters $\lambda_S$ and direct maximization is not possible. However, ML parameter estimates can be obtained iteratively using special case of the Expectation-Maximization (EM) algorithm. The EM algorithm is well explained in [2]. On line speaker enrollment that we implemented also trains

speaker models by using EM algorithm and registers speaker models for speaker recognition.
Fig. 1 shows an example of user interface for online speaker registration in robot.

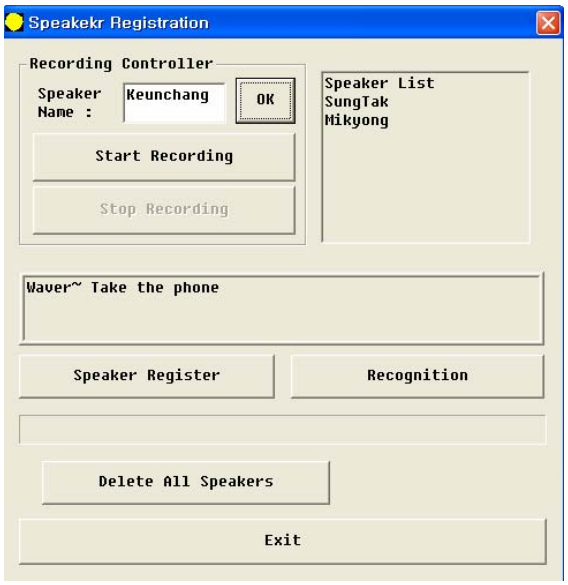


Fig. 1. An example of user interface for online speaker enrollment in robot.

### 2.2 Speaker Recognition System

Fig. 2 shows speaker recognition process in URC environment. Robot (named Waver) catches user's input speech through simple speech interface, and transmits the input speech to a server. In server, speaker recognition process which needs much computation power is performed and then the result is re-transmitted from server to Waver. Finally, Waver informs a message corresponding to the result to the user by using speech synthesis technique.

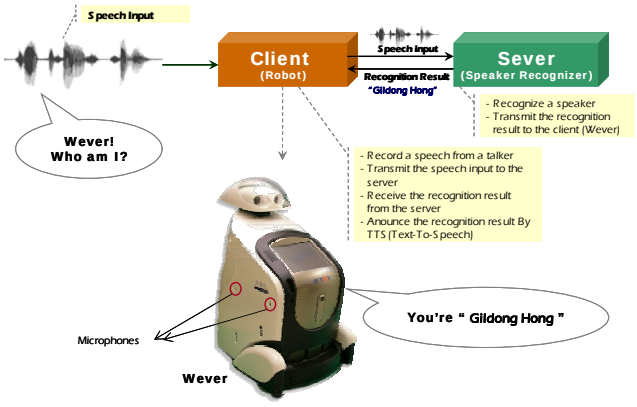


Fig. 2. The process of speaker recognition in Waver.

### 2.3 Speaker Recognition Based on Statistical Models

*A. Speaker Identification Using GMMs*

A Gaussian mixture density is represented by a weighted sum of $M$ component densities using the following equations [3][4].

$$p(x \mid \lambda) = \sum_{i=1}^{M} w_i b_i(x) \tag{2}$$

$$b_i(x) = \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)\right\} \tag{3}$$

where $x$ is a d-dimensional random vector, $b_i(x)$ is the PDF with mean vector $\mu_i$ and covariance matrix $\Sigma_i$ of $i$ component, and $w_i$ is the mixture weight of $i$ component that satisfies the constraint that $\sum_{i=1}^{M} w_i = 1$.

The complete Gaussian mixture density is parameterized by mixture weights, mean vectors, and covariance matrices from all component densities. These parameters are represented as equation (4)

$$\lambda = \{w_i, \mu_i, \Sigma_i\} \quad i = 1,...,M \tag{4}$$

For speaker identification, each speaker is represented by a GMM and is referred to as his/her model $\lambda$. A group of $S$ speakers, $S = \{1,2,...,K\}$ is presented by GMM's $\lambda_1, \lambda_2, ..., \lambda_K$. The objective is to find the speaker model which has the maximum a posteriori probability for a given observation sequence $(x_1, x_2, ..., x_T)$.

$$\hat{S} = \arg\max_{k} \sum_{t=1}^{T} \log(x_t \mid \lambda_k) \tag{5}$$

*B. Speaker Verification Using GMM-UBM(Universal Background Model)*

Given a segment of speech, *Y*, and hypothesized speaker, *S*, the task of speaker verification is to determine if *Y* was spoken by *S*. Speaker verification task can be re stated as a basic hypothesis test between

$H_0$: Y is from the hypothesized speaker S
and
$H_1$: Y is not from the hypothesized speaker S

A likelihood ratio test for the decision between two hypotheses is given by

$$\frac{p(Y \mid H_0)}{P(Y \mid H_1)} \begin{cases} \geq \theta & \text{accept } H_0 \\ < \theta & \text{reject } H_0 \end{cases} \tag{6}$$

where $p(Y \mid H_i)$, $i = 0 \text{ or } 1$, is the probability density function of the hypothesis $H_i$ evaluated for observed speech segment *Y*, also referred to as the likelihood of the hypothesis $H_i$ given speech segment. $\theta$ is the pre-defined threshold value for accepting or rejecting $H_0$. Often, the logarithm of equation (6) is used as like equation (7) [5].

$$\Lambda(Y) = \log p(Y \mid H_0) - \log p(Y \mid H_1) \tag{7}$$

Fig. 3 shows the basic components of speaker verification system.

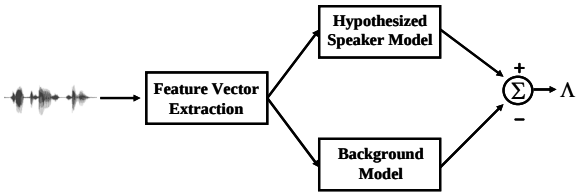


Fig. 3. Likelihood ratio test-based speaker verification system.

### *C. SCR (Speaker Confusion Rate) based Speaker Verification*

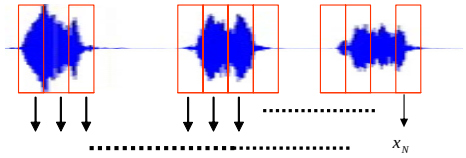


Fig. 4. The feature vectors of input utterence

We discard the silence frames by usig VAD, and using the other frames, feature vectors, $x_1, x_2, ..., x_N$ , are computed as shown in Fig. 4. When considering only $i^{th}$ frame, we can define speaker model $\lambda_{\hat{S}_i}^{\max}$ having maximum similarity as equation (8).

$$\lambda_{\hat{S}_i}^{\max} = \arg\max_k p(x_i \mid \lambda_k) \tag{8}$$

$$\lambda_{\hat{S}}^{C} = \arg\max_k p(x_1, x_2, ..., x_N \mid \lambda_k) \tag{9}$$

The *SCR* of the speaker model, $\lambda_{\hat{S}}^{C}$ which has maximum similarity in terms of total input vectors is described as equation (10) by using equation (9) and equation (11).

$$SCR(\lambda_{\hat{S}}^{C}) = \frac{1}{N}\sum_{n=1}^{N} d(\lambda_{\hat{S}_n}^{\max}, \lambda_{\hat{S}}^{C}) \tag{10}$$

$$d(\lambda_{\hat{S}_n}^{\max}, \lambda_{\hat{S}}^{C}) = \begin{cases} 0, & \text{if } \lambda_{\hat{S}_n}^{\max} = \lambda_{\hat{S}}^{C} \\ 1, & \text{otherwise} \end{cases} \tag{11}$$

$x_1$ $x_2$

The value of SCR is compared with pre-defined threshold value $\theta$ to decide on accepting $(SCR(\lambda_{\hat{S}}^{C}) < \theta)$ or rejecting $(SCR(\lambda_{\hat{S}}^{C}) \geq \theta)$ the hypothesized speaker [6].

## 3. Experimental Results

For performance evaluation, we use speech database which is collected by using sound A/D board of robot with 8 channels. The database is a collection of conversational speech from 23 male and 7 female speakers. Each speaker utters 20 sentences twice. When collecting speech database, the distances between robot and speaker vary with 1m, 2m, and 3m at the front side or in angle of 45 degrees in a quiet room. The speech signals are recorded at 16 kHz and 16 bits per sample. For speaker recognition, speaker models and UBM are GMMs with 160 mixtures. For UBM training, we use 300 sentences from 30 persons.

Table 1 shows the performance of speaker identification of 30 registered speakers according to different channels in terms of distance and degree, and the performance in the case of using the Likelihood Sum (LS) method over 8 channels.

Table 1. The performance of speaker identification (%).

| Spkr. / Mic. | 1m | | 2m | | 3m | |
|---|---|---|---|---|---|---|
| | 0° | 45° | 0° | 45° | 0° | 45° |
| Ch.0 | 92.67 | 92.00 | 91.00 | 85.00 | 62.00 | 84.33 |
| Ch.1 | 90.67 | 92.00 | 81.00 | 81.00 | 58.00 | 68.67 |
| Ch.2 | 95.33 | 93.67 | 89.67 | 87.67 | 64.00 | 88.00 |
| Ch.3 | 94.00 | 92.00 | 85.00 | 83.67 | 63.00 | 75.00 |
| Ch.4 | 91.00 | 86.67 | 78.00 | 77.67 | 59.33 | 72.33 |
| Ch.5 | 91.67 | 91.67 | 81.67 | 82.33 | 70.33 | 80.67 |
| Ch.6 | 94.33 | 92.33 | 79.00 | 81.33 | 60.67 | 79.67 |
| Ch.7 | 91.67 | 90.33 | 86.33 | 83.00 | 52.67 | 78.67 |
| LS | 94.67 | 95.00 | 89.00 | 88.00 | 68.67 | 85.00 |

From the results of speaker identification, the speaker identifier developed is acceptable to our task for identifying one of family members.

For the evaluation of speaker verification, we use only speech data which are collected on the front side. Registered speakers are 5 speakers (male: 3, female: 2). The performance according to distances is shown in Fig. 5 to 7.

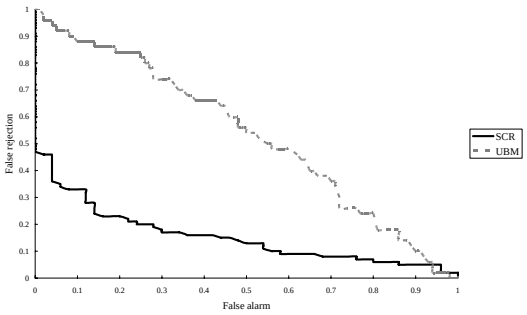


Fig. 5. ROC curves for speaker verification at 1m distance.

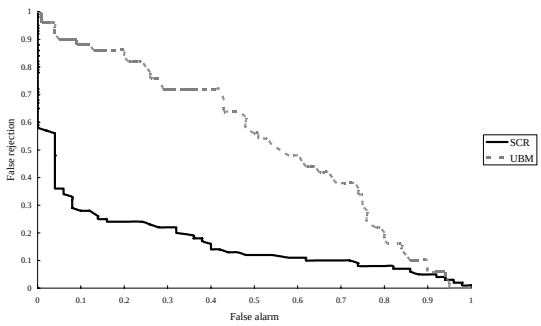


Fig. 6. ROC curves for speaker verification at 2m distance.

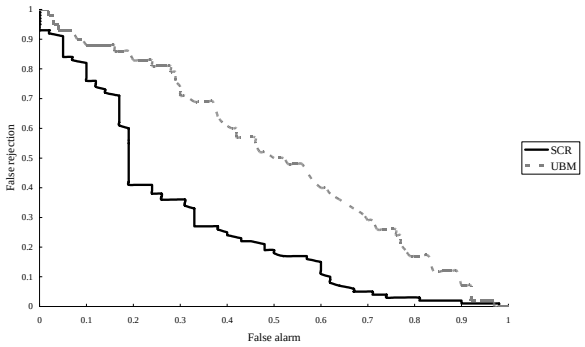


Fig. 7. ROC curves for speaker verification at 3m distance.

From the results of speaker verification, GMM-UBM based speaker verification does not perform well as expected, but SCR based speaker verification shows the acceptable performance compared with GMM-UBM method. The reason of poor performance of GMM-UBM based speaker verification is that the number of speakers for UMB is not enough and the speech data collected are too much corrupted by channel noise of the sound board of robot. Because of the reasons, un-reliable UBM was generated and GMM-UBM based speaker verification could not have a good performance.

## 4. Conclusions and Further Works

In this paper, we implemented a speaker recognition system which consists of text-independent speaker identification/verification and an online speaker registration which trains GMM as a speaker model in URC environment. The methods for speaker recognition we used are GMM based speaker identification and GMM-UBM based speaker verification which show recently successful performance, although GMM-UBM based speaker verification did not show satisfied performance in URC environment. In addition to these systems, we also apply SCR based speaker verification technique which has noise robustness. We evaluate the performance of these techniques in URC environment. The results of speaker identification show satisfied performance for the task of robot in URC environment, and SCR based speaker verification technique shows better performance than GMM-UBM based technique.

For further works, we will try to generate UBM using noise-canceled speech data and increase the number of speakers for UBM. And also we will apply the speaker adaptation technique for more convenient online speaker registration. In result, it is possible to make a speaker model with only a few sentences.

## Acknowledgements

This work was supported in part by MIC & IITA through IT Leading R&D Support Project.

## References

[1] P. Josep, and Jr. Compbell, “Speaker Recognition: A Tutorial,” Proc. of the IEEE, Vol. 85, No. 9, pp. 1437-1462, Sept. 1997.

[2] A. Dempster, N. Laird, and D. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” J. Royal Stat. Soc., Vol. 39, pp. 1- 38, 1997.

[3] D. Reynolds and R.C. Rose, “Robust Text Independent Speaker Identification Using Gaussian Mixture Speaker Models,” Proc. IEEE Transactions on Speech and Audio Processing, Vol.3, No.1, pp.72-83, 1995.

[4] B. Narayanaswamy and R. Ganggadharaiah, “Extracting Additional Information from Gaussian Mixture Model Probabilities for Improved Text-Independent Speaker Identification,” ICASSP, Vol. 1, pp. 621-624, Mar. 2005.

[5] D. Reynolds, T. Quatieri, and R. Dunn, “Speaker Verification Using Adapted Gaussian Mixture Models,” Digital Signal Processing, Vol. 10, pp.19-41, 2000.

[6] Kyuhong Kim, Hoirin Kim, and Minsoo Hahn, “Utterance Verification Using Search Confusion Rate and Its N-Best Approach,” ETRI Journal, Vol. 27, pp. 461-464, Aug. 2005.