

**PROCEEDINGS OF
THE 2007 INTERNATIONAL CONFERENCE ON
ARTIFICIAL INTELLIGENCE**

ICAI²⁰⁰⁷

Volume II

Editors

**Hamid R. Arabnia
Mary Qu Yang, Jack Y. Yang**

Associate Editors

**Nikolaos G. Bourbakis, Halil Ceylan, Yuehui Chen
Georgios A. Demetriou, David Dodds, Laura L. Elnitski
Okan K. Ersoy, David de la Fuente, Joydeep Ghosh
Peter M. LaMonica, Guo-Zheng Li, Raymond A. Liuzzi
Jun Ni, Andrzej Niemierko, Jose A. Olivas, Bill Rand
Gene Simmons, Ashu M. G. Solo, Andrew Sung, Sule Yildirim
Yanqing Zhang, Matthias Dehmer, Frank Emmert-Streib**



WORLD COMP'07

June 25-28, 2007

Las Vegas Nevada, USA

www.world-academy-of-science.org

©CSREA Press

TIB/UB Hannover

89

130 463 280



This set of volumes contain papers presented at The 2007 International Conference on Artificial Intelligence (ICAI'07). Their inclusion in this publication does not necessarily constitute endorsements by editors or by the publisher.

Copyright and Reprint Permission

Copying without a fee is permitted provided that the copies are not made or distributed for direct commercial advantage, and credit to source is given. Abstracting is permitted with credit to the source. Please contact the publisher for other copying, reprint, or republication permission.

Copyright © 2007 CSREA Press
ISBN: 1-60132-023-X, 1-60132-024-8 (1-60132-025-6)
Printed in the United States of America

CSREA Press
U. S. A.

Speaker Identification and Verification for Intelligent Service Robots

Keun-Chang Kwak, Hae-Jin Kim, Kyung-Suk Bae, Ho-Sub Yoon

Intelligent Robot Research Division,
Electronics and Telecommunications Research Institute (ETRI)
161 Gajeong-dong, Yuseong-gu, Daejeon, 305-700, Korea
Tel: +82-42-860-5752, Fax: +82-42-860-5752
kwak@etri.re.kr

Abstract— This paper is concerned with a text-independent speaker identification/verification and speech recognition for intelligent service robots with humanlike auditory system. The objective of this paper is to develop and integrate audio-based Human-Robot Interaction (HRI) components that can naturally interact between human and robot through audio information obtained from microphones in the noisy environments. Furthermore, we can naturally communicate with intelligent service robot through a Korean-based spontaneous speech recognition and text-independent speaker recognition to provide a customized service such as daily life schedule and selection of preferable TV channel for the identified speaker. The robot platform used in this study is WEVER, which is a network-based intelligent service robot developed at Intelligent Robot Research Division in Electronics and Telecommunication Research Institute (ETRI). This robot refers to an Ubiquitous Robotic Companion (URC) that provides necessary services anytime and anywhere. The experimental results obtained for ETRI speaker database reveal that the approach presented in this paper yields a good identification (94.5%) and verification (FAR: 0.014%, FRR: 0.067%) performance within 3 meter.

I. INTRODUCTION

OVER the past few decades, a considerable number of studies have been conducted on speaker identification/verification [7][8]. In general, the speaker recognition is divided into two specific tasks: speaker identification and speaker verification [1][2]. In the case of speaker identification, there is no a priori identity claim and system determines who the speaker is. On the other hand, speaker verification is that the system must decide if the input voice came from the claimed speaker or not. Although studies have made on speaker recognition in the area of biometrics, this marker is still in its infancy. Furthermore, very few attempts in the previous literature have been made on speaker recognition in contrast to speech recognition in the area of intelligent robots [3]. However, there has recently been a renewal of interest in Human-Robot Interaction (HRI) for

intelligent service robots. Among various HRI components, specifically the concern with audio-based HRI such as speech recognition, sound source separation, sound localization, and speaker recognition has been growing.

In this paper, we present text-independent speaker identification/verification and speech recognition to develop an audio-based HRI components for Ubiquitous Robotic Companion (URC)-based intelligent service robots, which exploit strong Information Technology (IT) infrastructure such as high-speed internet [4]. Here the URC means that it will provide the necessary services at any time and place to meet the user's requirements. Thus, it combines the network function with the current concept of a robot in order to enhance mobility and human interface. For this purpose, Mel-Frequency Cepstral Coefficients-Gaussian Mixture Model (MFCC-GMM) and likelihood ratio sets are developed for speaker identification and verification, respectively. Finally we perform a Korean-based spontaneous speech recognition and text-independent speaker recognition to provide a customized service under wireless network environments. The experimental results obtained for Electronics and Telecommunication Research Institute (ETRI) speaker database reveal that the presented approach yields a good identification and verification performance at a short or long-distance (3m-5m). The robot platform used in this study is WEVER, which is a URC-based intelligent service robot developed at Intelligent Robot Research Division in ETRI.

This paper is organized in the following manner. Section 2 describes the feature extraction and the generation of speaker model. In Section 3, we present text-independent speaker identification/verification for intelligent service robots. Section 4 reports on comprehensive experiment results for speaker identification/verification under robot environments. Furthermore, we show the service scenario for combining text-independent speaker identification with speech recognition. Finally concluding comments are

concluded in Section 5.

II. GENERATION OF SPEAKER MODEL

In order to analyze speech signal obtained from speaker, firstly the endpoint detection algorithm is performed. Fig. 1 shows the signal detected by log energy and zero crossing rates. Fig. 2 visualizes spectrogram using a Short-Time Fourier Transform (STFT) from the detected signal for time-dependent frequency analysis. The spectrogram divides the signal into overlapping segments, windows each segment.

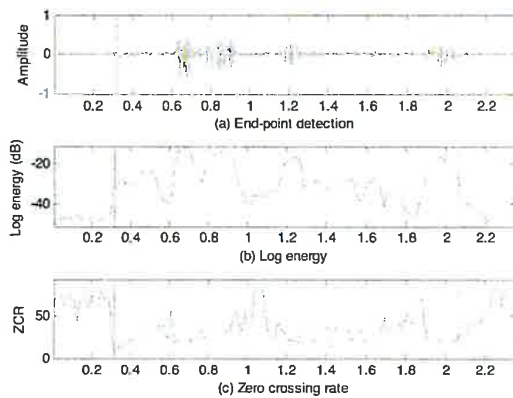


Fig. 1 End-point detection

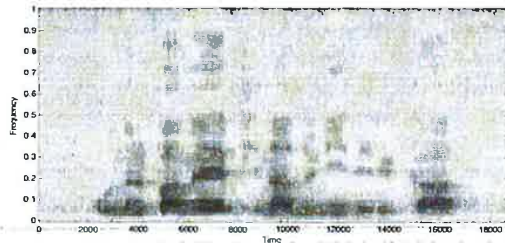


Fig. 2 Spectrogram of Fig. 1(a)

After detecting signal, we perform the feature extraction as shown in Fig. 3. This method is comprised of six stages: pre-emphasis, frame blocking, hamming widow to lessen distortion, Fast Fourier Transform (FFT), triangular bandpass filter (see Fig. 4), and cosine transform to get MFCC. For simplicity, we use 11 MFCC parameters except for the first order.

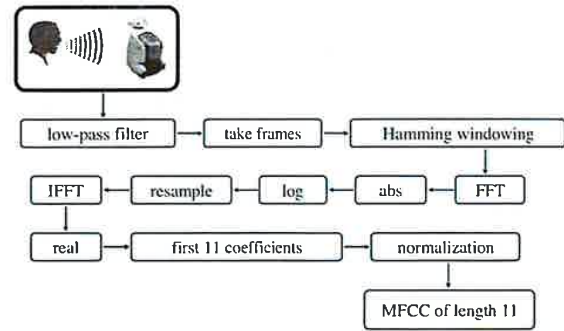


Fig. 3 block diagram for feature extraction

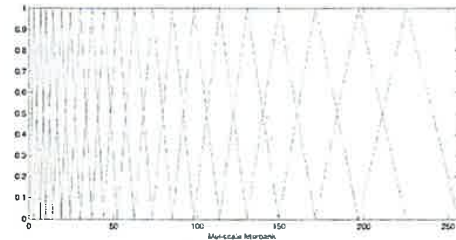


Fig. 4 triangular bandpass filter

In what follows, we describe the well-known GMM frequently used in conjunction with text-independent speaker identification/verification to represent speaker's model under robot environments. Here the distribution of feature vectors extracted from individual speech is performed by a Gaussian mixture density. For a D-dimensional feature vector denoted as $\mathbf{x}$, the mixture density for speaker is defined as

$$P(\bar{\mathbf{x}} | \lambda_s) = \sum_{i=1}^M w_i b_i(\bar{\mathbf{x}}) \quad (1)$$

where w_i is mixture weights and b_i is Gaussian mixture. The density is a weighted linear combination of M Gaussian mixture b_i parameterized by a mean vector μ_i and covariance matrix Σ_i . The Gaussian mixture is defined as the following equation

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

The mixture weights w_i satisfy the constraint $\sum_{i=1}^M w_i = 1$.

Thus, the parameters of speaker's model are denoted as $\lambda_s = \{w_i, \mu_i, \Sigma_i\}$ $i = 1, \dots, M$. For simplicity, diagonal covariance matrices are used to construct GMM, because diagonal matrix are more computationally efficient than full covariance matrix for training [5][6]. Given training speech from a speaker, the speaker model is trained by estimating the parameters of the GMM. The well-known method is maximum likelihood estimation. For a sequence of T training vectors $X = \{\bar{x}_1, \bar{x}_2, \dots, \bar{x}_T\}$, the GMM likelihood can be expressed as

$$p(X | \lambda_s) = \prod_{t=1}^T p(\bar{x}_t | \lambda_s) \quad (3)$$

The maximum likelihood parameter estimation is obtained by using the Expectation-Maximization (EM) algorithm as follows

$$\hat{w}_i = \frac{1}{T} \sum_{t=1}^T p(i | x_t, \lambda_s) \quad (4)$$

$$\hat{\mu}_i = \frac{\sum_{t=1}^T p(i | x_t, \lambda_s) x_t}{\sum_{t=1}^T p(i | x_t, \lambda_s)} \quad (5)$$

$$\hat{\Sigma}_i = \frac{\sum_{t=1}^T p(i | x_t, \lambda_s) x_t^2}{\sum_{t=1}^T p(i | x_t, \lambda_s)} - \hat{\mu}_i^2 \quad (6)$$

where $\hat{w}_i, \hat{\mu}_i, \hat{\Sigma}_i$ are the updated mixture weight, mean, and variance for i 'th Gaussian mixture, respectively. The advantages of GMM as the likelihood function are that it is computationally inexpensive.

III. SPEAKER IDENTIFICATION/VERIFICATION FOR HUNAN-ROBOT INTERACTION

For speaker identification, a group of speakers $S = \{1, 2, \dots, S\}$ 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 [5].

$$\hat{S} = \arg \max_{1 \leq k \leq S} p(\lambda_k | X) = \arg \max_{1 \leq k \leq S} \frac{p(X | \lambda_k) p(\lambda_k)}{p(X)} \quad (7)$$

where the second equation is based on Bayes' rule.

Assuming equally likely speakers, $p(X)$ is the same for all speaker models, the identification simplifies as follows

$$\hat{S} = \arg \max_{1 \leq k \leq S} p(X | \lambda_k) \quad (8)$$

Using logarithms and the independence between observations, the SI is computed as

$$\hat{S} = \arg \max_{1 \leq k \leq S} \sum_{t=1}^T \log p(\bar{x}_t | \lambda_k) \quad (9)$$

On the other hand, given speech feature vectors X , the speaker verification should be decided if X was spoken by the client or not. Based on the similarity domain, this can be obtained as a statistical test between $H_0 : S$ and $H_1 : S'$ where H_0 is the null hypothesis that the claimant is the client S' , while H_1 is the alternative hypothesis that the claimant is an impostor S' [6][7]. The decision according to the Bayesian rule for minimum risk is given by

$$L(X) = \frac{p(X | S)}{p(X | S')} \begin{cases} > \theta : X \in H_0 \\ \leq \theta : X \in H_1 \end{cases} \quad (10)$$

where θ is a prescribed threshold. Taking the logarithm, the likelihood ratio is the following equation

$$\log L(X) = \log p(X | S) - \log p(X | S') \begin{cases} > \log \theta : X \in H_0 \\ \leq \log \theta : X \in H_1 \end{cases} \quad (11)$$

where $\log L(x)$ is the normalized log-likelihood score. The normalized log-likelihood value of X can be computed as

$$\log p(X | S) = \frac{1}{T} \sum_{t=1}^T \log p(\bar{x}_t | S) \quad (12)$$

The well-known geometric mean and the maximum can be used to calculate the normalized log-likelihood score given not the client model. For a set of background speaker models of size B : $S' = \{S_1, S_2, \dots, S_B\}$, the geometric mean method is calculated as

$$\log p(X | S') = \frac{1}{B} \sum_{b=1}^B \log p(X | S_b) \quad (13)$$

The maximum method is computed as

$$\log p(X|S') = \max_{S_b \in S} \log p(X|S_b) \quad (14)$$

In previous literature, this approach has been called likelihood ratio sets [6], cohorts, and background speakers. The main advantage of this approach is that a single speaker-independent model can be trained once for a particular task and then used for all hypothesized speakers in that task [5].

IV. EXPERIMENTAL RESULTS

In this section, we use ETRI speaker database to evaluate the identification and verification performance of the presented speaker recognition system. Fig 5 shows WEVER with eight microphones and sound board, which is URC-based intelligent service robot developed by ETRI. Here sound board with eight channels called Multimodal Interface Module (MIM) sound board has already developed in Intelligent Robot Research Division of 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) $\times$ 8 channels) and test data sets (10 sentences $\times$ 20 people $\times$ 5 distances (1,2,3,4,5m) $\times$ 8 channels). For simplicity, we use ETRI speaker database consisting 2200 sentences for only single microphone. The recording was done in an office environment. The audio is stored as a mono, 16bit, 16kHz WAV file. Firstly we performed text-independent speaker identification under robot environments. Fig. 6 showed the identification performance of GMM obtained from the variation of several mixture size and distances. As shown in Fig. 6, the experimental results within 3 meter showed a good identification performance (94.5%). Furthermore, the results obtained from the variation of GMM size showed a similar performance in this experiment.

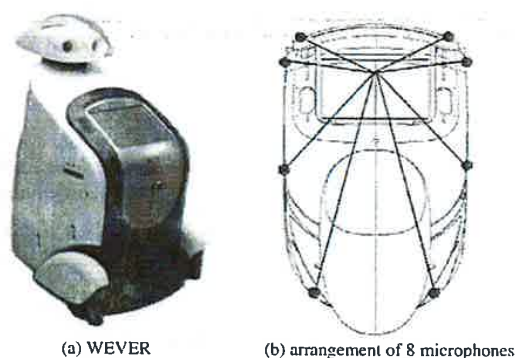


Fig. 5 Robot platform: WEVER

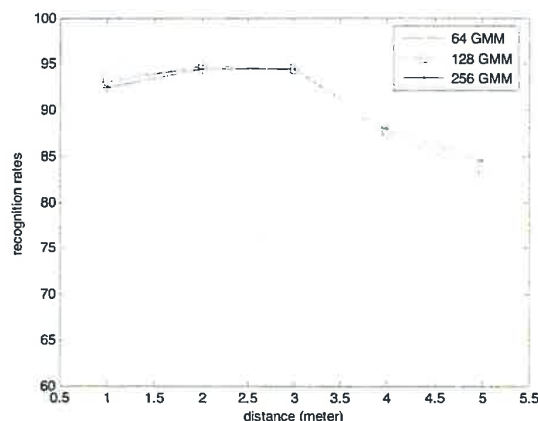


Fig. 6 Identification performance obtained by MFCC-GMM

On the other hand, we performed text-independent speaker verification for 3 clients and 17 impostors under robot environments. As explained before, we used likelihood ratio sets [6] for speaker verification. Fig. 7 showed the histogram of impostor versus the users corresponding to score. Finally we obtained verification performance (FAR: 0.014%, FRR: 0.067%) within 3 meter, when threshold is 0.1 (EER point). These results lead us to the conclusion that the text-independent speaker identification/verification based on GMM yielded a good identification and verification performance under robot environments.

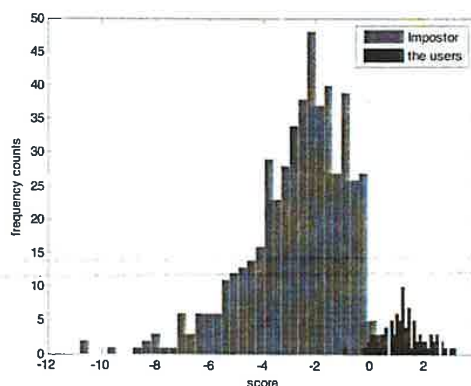


Fig. 7 Verification performance by cohort model

In what follows, we shall show a scenario for combining the text-independent speaker identification presented in this paper and spontaneous speech recognition. On the basis of these methods, the robot can simultaneously recognize what the speaker said as well as identification of the speaker. Firstly the speaker registration is performed through wireless network in real-time. The speaker model is generated by several sentences as GUI shown in Fig. 8. Fig. 9 shows a scenario for providing user-customized service such as speaker's daily life schedule and selection of preferable TV channel for two identified speaker. Fig. 10 shows the

procedure for speech and speaker recognition under network-based robot environments.

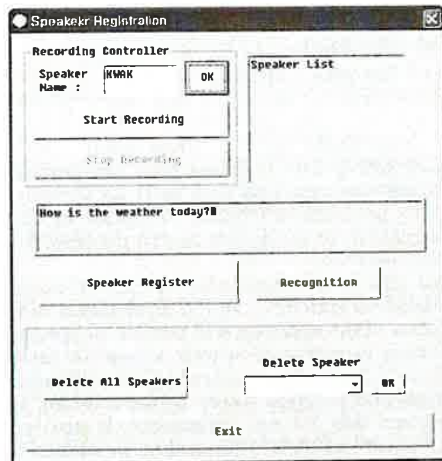


Fig. 8 GUI for speaker registration

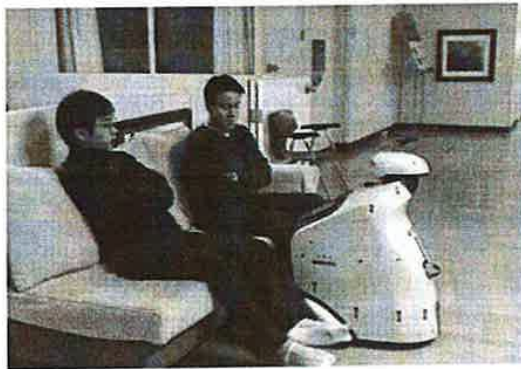


Fig.9 Scenario through combination text-independent SI and spontaneous speech recognition

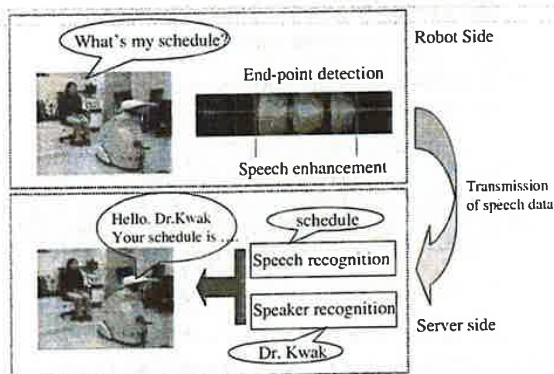


Fig. 10 Procedure for speech and speaker recognition

V. CONCLUSION

We have developed the text-independent speaker identification/verification as audio-based HRI components for URC-based intelligent service robots, which exploit strong IT infrastructure. It is concluded that the approach presented in this paper showed usefulness and effectiveness for speech/speaker recognition under robot environments. A further direction of this study will be to fuse feature information obtained through several microphones. Furthermore, we shall perform further research on multi-modal user identification and verification using audio and vision information.

ACKNOWLEDGMENT

This work is a result of URC project sponsored by MIC of Korean government.

REFERENCES

- [1] A. E. Rosenberg, "Automatic speaker verification: a review," *Proceedings of IEEE*, vol. 64, no. 4, pp.475-487, 1976.
- [2] J. P. Campbell, "Speaker recognition: a tutorial," *Proceedings of IEEE*, vol. 85, no. 9, pp. 1437-1462, 1997.
- [3] H. Lee and H. Ko, "Competing models-based text-prompted speaker independent verification algorithm," *Speech Communication*, vol. 48, pp. 28-44, 2006.
- [4] 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.
- [5] 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.
- [6] 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.
- [7] C. S. Liu, H. C. Wang, and C. H. Lee, "Speaker verification using normalized log-likelihood score," *IEEE Trans. on Speech and Audio Processing*, vol. 4, no. 1, pp.56-60, 1996.
- [8] T. Pham and M. Wagner, "Similarity normalization for speaker verification by fuzzy fusion," *Pattern Recognition*, vol. 33, pp. 309-315, 2000.