SPEAKER'S GENDER IDENTIFICATION FOR HUMAN-ROBOT INTERACTION

Kyung-Sook Bae, Keun-Chang Kwak, Soo-Young Chi
Intelligent Robot Research Division, Electronics and Telecommunications Research Institute(ETRI), 161 Gajeong-Dong, Yuseong-Gu, Daejeon, Korea

Keywords: Gender Identification, GMM, URC.

Abstract: This paper is concerned with a text-independent Speaker's gender Identification (GI) for Human-Robot Interaction (HRI). For this purpose, we perform speaker's gender recognition based on Gaussian Mixture Model (GMM) and use robot platform called WEVER, which is a Ubiquitous Robotic Companion (URC) intelligent service robot developed at Intelligent Robot Research Division in Electronics and Telecommunication Research Institute (ETRI). Furthermore, we communicate with intelligent service robots through a Korean-based spontaneous speech recognition and text-independent speaker's gender identification to provide a suitable service such as selection of preferable TV channel or music for the identified speaker's gender. The experimental results obtained for ETRI speaker database reveal that the approach presented in this paper yields a good identification (94.9%) performance within 3 meter.

1 INTRODUCTION

Speaker Identification has many applications and is a topic of great interest in the speech research community. Speaker's Gender Identification (GI) can be thought of as a subset of speaker identification and also can be contributed to increase performance of Speaker Identification as a preprocessing. In the past, GI has been investigated for clean speech by Wu and Childers (Wu, 1991). Parris and Carey studied GI for different languages using telephone speech data. In their system Paris and Carey trained an GI system using speakers of British English and tested their system using speakers of British English, US English, and 10 other languages. Slomka and Sridharan proposed text-independent GI systems capable of being optimized for multiple adverse conditions, including various coders, and reverberation levels (Slomka, 1997).

Recently, there has been a renewal of interest in Human-Robot Interaction (HRI) for intelligent robots. Among HRI components, specifically the concern with speech-based HRI such as speech recognition, sound source separation, and speaker recognition has been growing. In this paper, we present text-independent GI to develop HRI components for Ubiquitous Robotic Companion (URC) intelligent service robots, which exploit strong Information Technology (IT) infrastructure such as high-speed internet. 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, we perform gender recognition based on Gaussian Mixture Model (GMM) (Reynolds, 1995) through a microphone equipped with WEVER robot developed by ETRI. Furthermore, we communicate with intelligent service robots through spontaneous speech recognition and text-independent speaker's gender recognition to provide a suitable service for the identified speaker's gender. The experimental results obtained for ETRI speaker database reveal that the presented approach yields a good identification performance at a short or long-distance (3m-5m). Proposed GI system is shown in Figure 1.

Bae K., Kwak K. and Chi S. (2006).
SPEAKER'S GENDER IDENTIFICATION FOR HUMAN-ROBOT INTERACTION.
In *Proceedings of the International Conference on Signal Processing and Multimedia Applications*, pages 339-342
DOI: 10.5220/0001569103390342

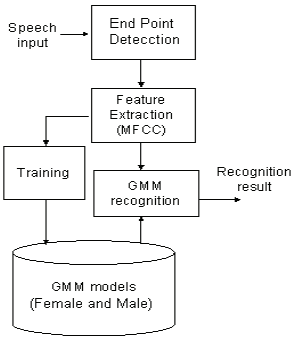

Figure 1: Gender Identification System.

This paper is organized in the following manner. Section 2 introduces the background and concept of URC intelligent service robots. Section 3 describes GMM method for text-independent GI. The well-known approaches for identification are explained in Section 4. Section 5 reports on comprehensive experiment results for GI under robot environments. Furthermore, we shall a scenario that the robot offers a service according to the speaker's sex by text-independent GI. Finally concluding comments are concluded in Section 6.

2 URC INTELLIGENT SERIVCE ROBOTS

In this Section, we briefly introduce URC intelligent service robots. The network-based intelligent service robot refers to a URC that provides necessary services anytime and anywhere. Recently Ministry of Information and Communication (MIC) in Korea has launched IT-based service robot project based on the concept of URC. Thus, the URC will be developed with a priority placed on human-focused functions. Consumers will be able to enjoy various services of the robot at lower costs since the URC will operate by adding network functions to the existing robots. The combination of a robot and IT infrastructure will result in human-oriented interfaces and technologies that enhance our living standards. Figure 2 shows a schematic diagram of distributed control middleware for URC to communicate with URC servers in order to use various home services and software. These robots are very appealing in that they could quite possibly enable Korea to secure competitive superiority over other countries in the robot development area.

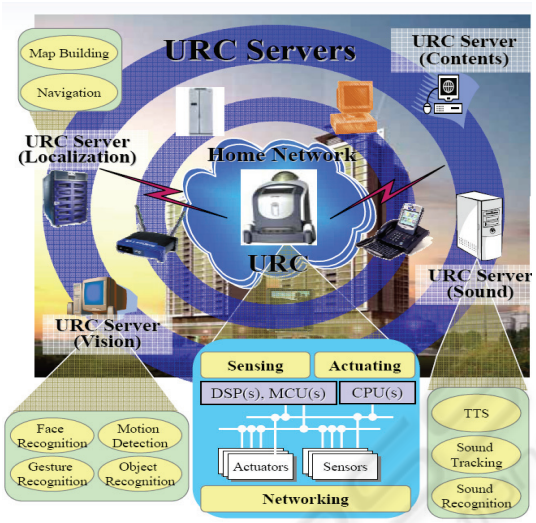

Figure 2: Schematic diagram for a middleware of URC intelligent service robots.

3 GAUSSAIN MIXTURE MODEL

In what follows, we describe the well-known GMM frequently used for performing text-independent GI to represent gender 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 **x**, the mixture density for gender is defined as

$$P(\vec{x} \mid \lambda_s) = \sum_{i=1}^{M} w_i b_i(\vec{x}) \quad (1)$$

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 $\sum_i$. The Gaussian mixture is defined as the following equation

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

The mixture weights w_i satisfy the

constraint $\sum_{i=1}^{M} w_i = 1$. Thus, the parameters of gender model are denoted as $\lambda_s = \{w_i, \mu_i, \Sigma_i\}$, $i = 1, \ldots, M$. For simplicity, diagonal covariance matrices are used to construct GMM, because diagonal matrix are more computationally efficient than full covariance matrix for training (Reynolds, 1995). Given training speech from a speaker, the speaker's gender model is trained by estimating the parameters of the GMM. The well-known method is maximum likelihood estimation. For a sequence of T training vector $X = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_T\}$, the GMM likelihood can be expressed as

$$p(X \mid \lambda_s) = \prod_{t=1}^{T} p(\vec{x}_t \mid \lambda_s) \tag{3}$$

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

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

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

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

where $\hat{w}_i, \hat{\mu}_i, \hat{\Sigma}_i$ are the updated parameters (mixture weight, mean, and variance) for i'th Gaussian mixture. The advantages of GMM as the likelihood function are that it is computationally inexpensive. It is also based on a well-understood statistical model. For text-independent tasks, GMM is insensitive the temporal aspects of the speech by modeling only the underlying distribution of acoustic observations from a speaker [Reynolds, 1995].

4 GENDER RECOGNITION

After detecting signal, we perform the feature extraction as shown in Figure 3. This method is comprised of six stages: pre-emphasis, frame blocking, hamming widow to lessen distortion, Fast Fourier Transform (FFT), triangular bandpass filter, and cosine transform to get MFCC. The size of these feature vectors is 12×82, 24×82, and 36×82 , respectively. For simplicity, we use 11 order Mel-scale cepstrum parameters except for the first order.

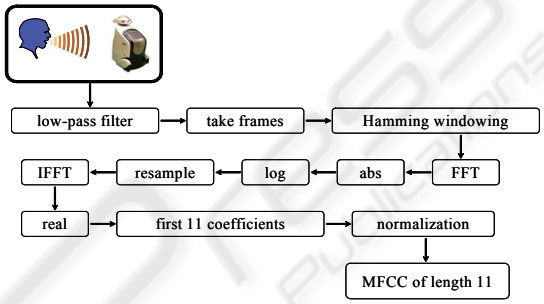

Figure 3: Block diagram for feature extraction.

For GI, a group of speaker's sexuality $S = \{Male, Female\}$ is represented by GMM's parameters. The main goal is to find the speaker's gender model which has the maximum a posteriori probability as the following equation (Reynolds, 1995).

$$\hat{S} = \arg \max_{1 \le k \le S} p(\lambda_k \mid X) = \arg \max_{1 \le k \le S} \frac{p(X \mid \lambda_k) p(\lambda_k)}{p(X)} \tag{7}$$

where the second equation is based on Bayes' rule. Assuming equally likely genders, p(X) is the same for male model and female model, the identification simplifies as follows

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

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

$$\hat{S} = \arg \max_{1 \le k \le S} \sum_{t=1}^{T} \log p(\vec{x}_t \mid \lambda_k) \tag{9}$$

5 EXPERIMENTAL RESULTS

In this section, we use ETRI speaker database to evaluate the identification performance of the presented GI system. Figure 4 shows WEVER, which is URC intelligent service robot developed by ETRI. The database is constructed by audio recording of 80 speakers (20 females and 60 males). The data set consists of 30 sentences for each speaker. We divide the speech data into training (30 sentences × 20 people (10 females and 10 males) × 3 distances (1,2,3m)) and test data sets (30 sentences × 60 people(10 females and 50 males) × 5 distances (1,2,3,4,5m)). The recording was done in an office environment. The audio is stored as a mono, 16bit, 16KHz, Wav file. We performed text-independent GI under robot environments. Figure 5 shows the GI performance of GMM obtained from the variation of several mixture size and distances.

Figure 4: Robot platform WEVER.

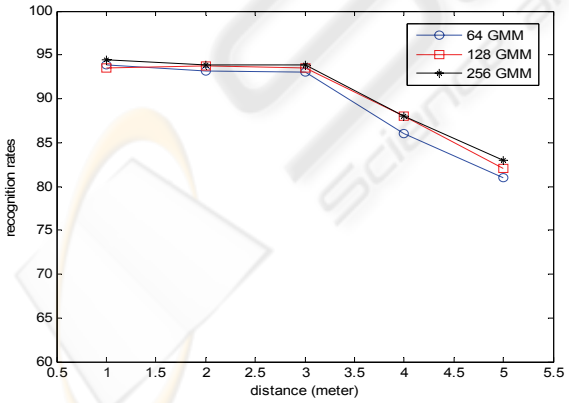

Figure 5: Performance obtained by GMM.

As shown in Figure 6, the experimental results within 3 meter showed a good identification performance (94.9%). Furthermore, the results obtained from the variation of GMM size showed a similar performance in this experiment.

In what follows, we shall show a scenario for combining the text-independent GI presented in this paper and spontaneous speech recognition. Fig. 10 shows a scenario for providing a suitable service such as selection of preferable TV channel.

Figure 6: Scenario through combination text-independent GI and spontaneous speech recognition.

6 CONCLUSION

We have developed the text-independent GI as the speech-based HRI components for URC intelligent service robots, which exploit strong IT infrastructure. It is concluded that the approach presented in this paper showed usefulness and effectiveness for GI 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.

REFERENCES

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.

Slomka, S and Sridharan, S, “Automatic gender identification optimised for language independence,” TENCON ‘97, vol. 1, pp. 145–148, 1997.

J. P. Campbell, “Speaker recognition: a tutorial,” *Proceedings of IEEE,* vol. 85, no. 9, pp. 1437-1462, 1997.

K. Wu and D.G. Childers, “Gender Recognition from Speech. Part I: Coarse Analysis”, J. Acoust. Soc. Am., Vol. 90, No. 4, pp1828-1840, 1991.