Performance Comparison of Gender and Age Group Recognition for Human-Robot Interaction

Myung-Won Lee
Dept. of Control and Instrumentation Engineering,
Chosun University, 375 Seosuk-dong
Gwangju, Korea

Keun-Chang Kwak*
Dept. of Control, Instrumentation, and Robot Engineering,
Chosun University, 375 Seosuk-dong
Gwangju, Korea

***Abstract*—In this paper, we focus on performance comparison of gender and age group recognition to perform robot's application services for Human-Robot Interaction (HRI). HRI is a core technology that can naturally interact between human and robot. Among various HRI components, we concentrate audio-based techniques such as gender and age group recognition from multichannel microphones and sound board equipped with robots. For comparative purposes, we perform the performance comparison of Mel-Frequency Cepstral Coefficients (MFCC) and Linear Prediction Coding Coefficients (LPCC) in the feature extraction step, Support Vector Machine (SVM) and C4.5 Decision Tree (DT) in the classification step. Finally, we deal with the usefulness of gender and age group recognition for human-robot interaction in home service robot environments.**

Keywords-gender recognition; age group recognition; human-robot interaction.

I. INTRODUCTION

Conventional industrial robots perform jobs and simple tasks by following pre-programmed instructions for humans in factories. Meanwhile, the main goal of the intelligent service robot is to adapt to the necessities of life as accessibility to human life increases. While industrial robots have been widely used in many manufacturing industries, intelligent service robots are still in elementary standard. Although the intelligent robots have been brought to public attention, the development of intelligent service robots remains as a matter to be researched further.

Recently, there has been a renewal of interest in Human-Robot Interaction (HRI) for intelligent service robots [1-2]. This is different from HCI (Human-Computer Interaction) in that robots have an autonomous movement, a bidirectional feature of interaction, and diversity of control level. Among various HRI components, we especially focus on audio-based HRI. Audio-based HRI technology includes speech recognition, speaker recognition [3][4], sound source localization [5], sound source separation, speech emotional recognition, speech enhancement, gender and age group recognition. Among various audio-based HRI components, we focus on gender and age group recognition. The robot platform used in this paper is WEVER, which is a network-based intelligent home service robot equipped with multi-channel sound board and three low-cost condenser microphones. Finally, we perform the performance comparison in the step of feature extraction (MFCC, LPCC) and classification (SVM, C4.5) for gender and age group classification.

The material of this paper is organized into following fashion. In section 2, we describe and discuss about well-known feature extraction methods Mel-Frequency Cepstral Coefficients (MFCC) and Linear Prediction Coding Coefficients (LPCC). In section 3, we deal with Support Vector Machine (SVM) and C4.5 Decision Tree (DT) for classification. Here we elaborate on gender and age group recognition in home service robots equipped with multiple microphones and multi-channel sound board. In section 4, we perform the experimental setup and performance comparison. Finally the conclusions and comments are given in section 5.

II. FEATURE EXTRACTION METHODS

The speech signals are obtained from the first channel of sound board at a distance of 1 meter in quite office environments. The speech signal is sampled with 16kHz, and each sample is encoded with 16bits. There are 20 sentences in a long speech signal for gender classification data, only one sentence in a speech signal for age classification data. Speech signals are assumed to be time invariant within a time period of 10 to 30 ms. Short-time processing methods are adopted for speech signal processing. A window sequence is used to cut the speech signal into segments, and short-time processing is periodically repeated for the duration of the waveform. The key problem in speech processing is to locate accurately beginning and ending of a speech. Endpoint detection (EPD) enables computation reduction and better recognition performance. We detect beginning and ending points of speech intervals using short-time energy and short-time zero crossings

The short time energy is as follows

$$E_n = \sum_{m=-\infty}^{\infty}\left[x(m)w(n-m)\right]^2 \tag{1}$$

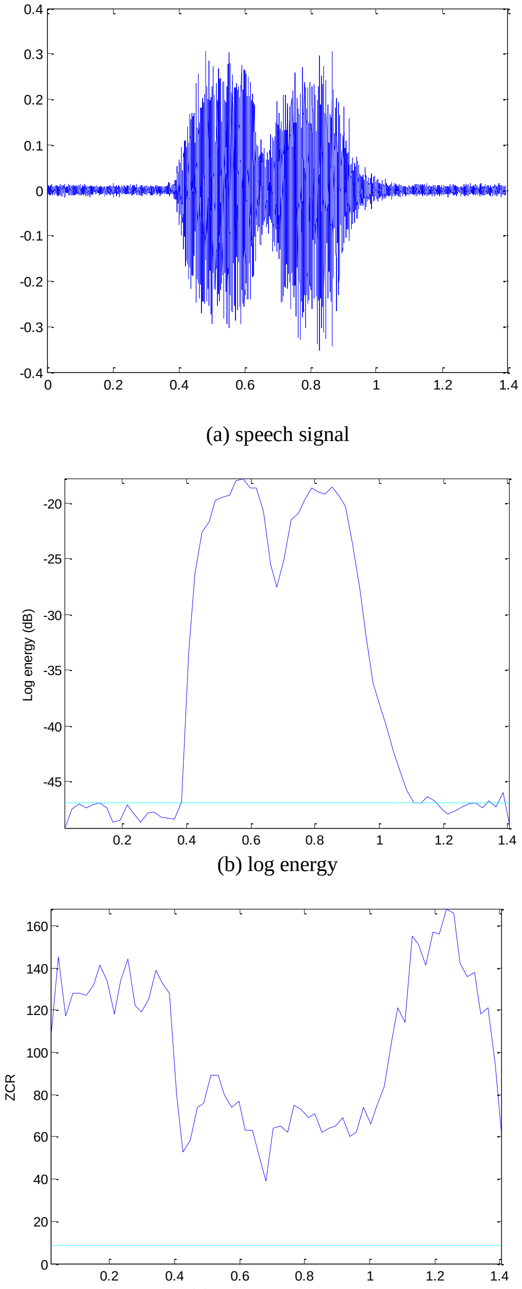
The short time zero crossing rate is as follows

$$Z_n = \sum_{m=-\infty}^{\infty}\left|sgn\left[x(m)\right]-sgn\left[x(m-1)\right]\right|w(n-m) \tag{2}$$

Figure 1 shows the EPD obtained from log energy and zero crossing rate. Rectangular window gives equal weights to all samples in the window. Hamming window gives most weight to middle sample. Rectangular and Hamming windows can be expressed as follows, respectively.

$$w(n)=1,\quad 0\le n\le N \ \ and \ \ 0 \ \ otherwise \tag{3}$$

$$w(n)=\begin{cases}0.54-0.46\cos(2\pi n/(N-1)) & 0\le n\le N-1\\ 0 & otherwise\end{cases} \tag{4}$$



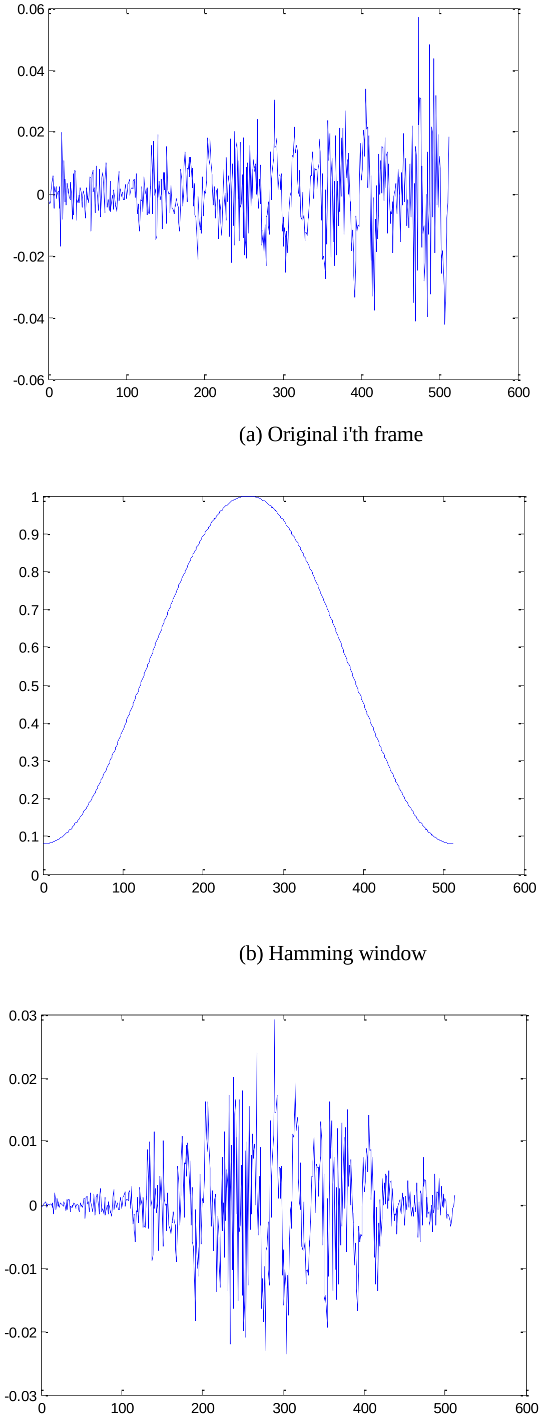
(a) speech signal

(b) log energy

(c) zero crossing rate

Figure 1. Endpoint detection

The window used is hamming window, and the window length is 512 samples with 30% overlap. Figure 2 shows original i'th frame and the transformed i'th frame obtained by hamming window.



(a) Original i'th frame

(b) Hamming window

(c) i'th frame obtained by hamming window

Figure 2. Signal obtained by Hamming window

After detecting endpoints of the speech interval, silence intervals from the speech signal are removed. We cut the speech signal into segments, each segment is 512 sample points in each frame for the signal with 16kHz sample rate. The size of overlapped frame is 171 samples.

The number of the filter bank is 20. The dimension of MFCC is 12. Feature extraction is based on each frame of the speech signals. After detecting signal, the feature extraction step is performed by six stages to obtain MFCC. These stages consist of pre-emphasis, frame blocking, hamming window, FFT (Fast Fourier Transform), triangular bandpass filter, and cosine transform [6]. For simplicity, we use 11 MFCC parameters except for the first order. The construction procedure of MFCC is shown in Figure 3.

The mel scale filter bank is a series of triangular bandpass filters hat have been designed to simulate the bandpass filtering believed to occur in the auditory system. This corresponds to a series of bandpass filters with constant bandwidth and spacing on a mel frequency scale as shown in Figure 4.

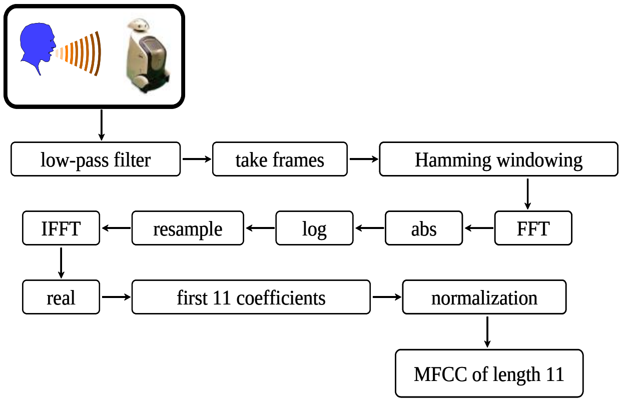


Figure 3. Procedure of MFCC.

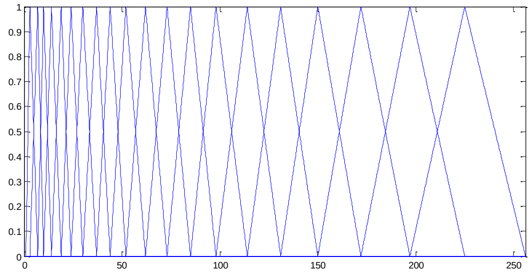


Figure 4. Mel scale filter bank

On the other hand, LPCC (Linear Prediction Coding Coefficients) method provides accurate estimates of speech parameters. The current speech sample can be closely approximated as a linear combination of past samples.

$$x(n) \approx a_1 x(n-1) + a_2 x(n-2) + \cdots + a_p x(n-p) \quad (5)$$

Coefficients are determined by minimizing the sum of squared differences between the actual speech samples and the linearly predicted ones.

III. Classification Methods

We firstly consider the use of the support vector machine (SVM) as a nonlinear classifier. This is legitimated by the fact that SVMs come with high generalization capabilities. Among various SVM models, we use the one known as LIBSVM. LIBSVM is composed of C-support vector classification (C-SVM) and ν -support vector classification (ν -SVM). Here we employ the C-SVM in the form proposed by Vapnik [7] for the implementation of multi-class classification. Furthermore we consider polynomial kernel functions frequently used in conjunction with classification tasks

Polynomial: $K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0$ (6)

where γ and r are kernel parameters.

On the other hand, C4.5 is a method used to generate a decision tree developed by Quinlan [8]. C4.5 is an extension of ID3 algorithm. The decision tree generated by C4.5 can be used for classification. For this reason, C4.5 is often referred to as a statistical classifier. C4.5 builds decision tree from a set of training data in the same way as ID3 using the concept of information entropy. The training data is a set S=s_1, s_2, … of already classified samples. Each sample s_i = x_1, x_2, … is a vector where x_1, x_2, … represents attributes or features of the sample. The training data is augmented with a vector C = c_1, c_2, … where c_1, c_2, … represent the class to which each sample belongs. At each node of the tree, C4.5 choose one attribute of the data that most effectively splits its set of sample into subsets enriched in one class or the other. Its criterion is the normalized information gain that results from choosing an attribute for splitting the data.

The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then recourses on the smaller sublists. The algorithm has a few base cases. All the samples in the list belongs to the same class. When this happens, it simply creates a leaf node of the decision tree saying to choose that class. None of the features provide any information gain. In this case, C4.5 creates a decision node higher up the tree using the expected value of the class. Instance of previously-unseen class encountered. Again, C4.5 creates a decision node higher up the tree using the expected value.

IV. Experimental Results

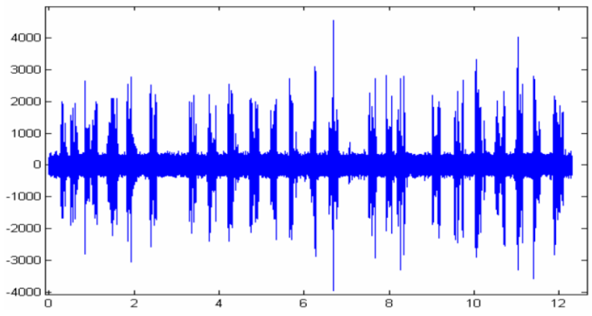
In this section, we describe our comprehensive set of experiments and draw conclusions regarding the classification performance in comparison with well-known methods frequently used in conjunction with the feature extraction and classification. We used hamming window 512 samples to multiply the speech signal to enable short-time speech signal processing. We extract MFCC and LPCC features based on each frame to produce feature data for classification.

The classification data includes 14 long speech signals for gender classification from 7 female and 7 male, 500 speech signals (200 signals from children, 300 signals from adults) for age group classification, respectively. We divide 2/3 of examples for training into the rest for testing. The training and testing data for gender classification is 6960 and 3482, respectively.

The training and testing data for age group classification is 12925 and 6464, respectively. Figure 5 shows 20 sentences in a long speech signal for gender classification data. Figure 6 shows u-robot test bed environments including three rooms and a living room. Figure 7 shows multi-channel sound board. These microphones are low-price condenser. Furthermore, multi-channel sound board was developed for sound localization and speech/speaker recognition in Electronics Communications Research Institutes (ETRI). Figure 8 visualizes MFCC obtained from one sentence.

Table 1 lists the result of performance comparison for gender classification. As listed in Table 1, the experimental results revealed that MFCC-SVM showed good performance (93.16%) in comparison to other presented approaches for testing data set. Table 2 lists the result of performance comparison for age group classification. The experimental results obtained that MFCC-SVM (91.39%) outperformed other methods in like manner.

As a result, the SVM and DT classifiers obtained better performance with MFCC features than LPCC features. Auditory model has been introduced in the MFCC feature, other auditory model embedded features can be extracted for future classification. Other features as pitch period, formants, short-time average energy etc. can be extracted to combine with MFCC or LPCC features for classification.



Speech signals with 20 sentences

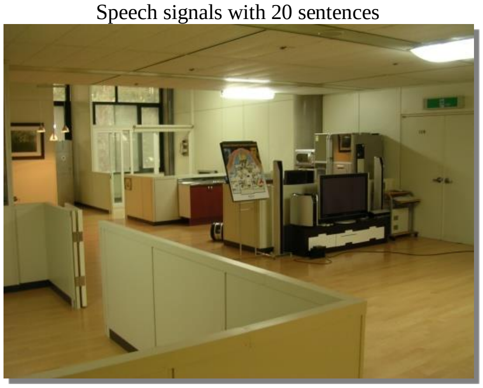


Figure 5. Robot test bed environment

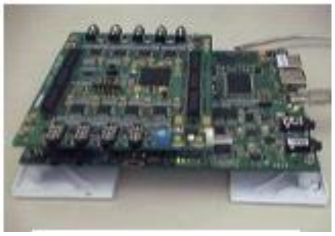


Figure 6. Multi-channel sound board(MIM board)

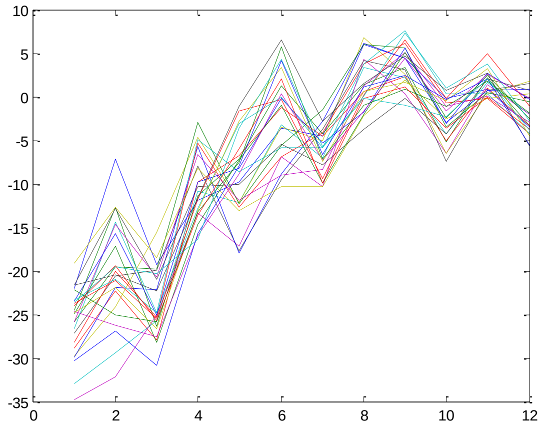


Figure 7. MFCC obtained by one sentence.

TABLE I. PERFORMANCE COMPARISON (GENDER CLASSIFICATION)

	Training data	Testing data
MFCC-SVM	95.89	93.16
MFCC-DT	94.33	91.45
LPCC-SVM	93.28	86.60
LPCC-DT	93.10	83.02

TABLE II. PERFORMANCE COMPARISON (AGE GROUP CLASSIFICATION)

	Testing data
MFCC-SVM	91.39
MFCC-DT	88.37
LPCC-SVM	84.69
LPCC-DT	82.72

V. CONCLUSIONS

We have performed the comparative analysis for gender and age group classification of audio-based HRI components. These components are compared with MFCC-SVM, MFCC-DT, LPCC-SVM, and LPCC-DT. The experimental results revealed that the aggregate of the MFCC-SVM showed better performance in comparison with other methods. We have shown the usefulness and effectiveness of the presented technique through the performance obtained from the constructed databases.

In the future studies, we shall continuously develop other techniques such as sound source separation and fusion of information obtained from multi-microphones for humanlike robot auditory system. Also, we can apply to customized service application based on the integrated robot audition system including gender and age group recognition, speaker and speech recognition, and sound source localization and separation.

The presented technique can be applied to service robots such as home service robots, edutainment robots, and u-health robots as well as various application areas.

REFERENCES

[1] K. C. Kwak, S. S. Kim, "Sound Source Localization With the Aid of Excitation Source Information in Home Robot Environments," IEEE Trans. on Consumer Electronics, Vol. 54, No. 2, 2008, pp. 852-856.

[2] K. C. Kwak, "Face Recognition with the Use of Tensor Representation in Home Robot Environments", IEICE Electronics Express, Vol. 6, No. 4, 2009, pp. 187-192.

[3] M. Ji, S. Kim, and H. Kim, "Text-independent speaker identification using soft channel selection in home robot environments," IEEE Consumer Electronics, vol. 54, no. 1, 2008, pp.140-144.

[4] K. S. R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker recognition," IEEE Signal Processing Letters, vol. 13, no. 1, 2006, pp. 52-55.

[5] V. C. Raykar, B. Yegnanarayana, S. R. M. Prasanna, and R. Duraiswami, "Speaker localization using excitation source information in speech," IEEE Speech and Audio Processing, vol. 13, no. 5, 2005, pp.751-761.

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

[7] V. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.

[8] J. R. Quinlan, C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993

AUTHORS PROFILE

Myung-Won Lee received the B.Sc. and M.Sc. from Chosun University, Gwangju, Korea, in 2010 and 2012, respectively. He is currently pursuing a candidate for the Ph.D. His research interests include human–robot interaction, computational intelligence, and pattern recognition.

Keun-Chang Kwak received the B.Sc., M.Sc., and Ph.D. degrees from Chungbuk National University, Cheongju, Korea, in 1996, 1998, and 2002, respectively. During 2003–2005, he was a Postdoctoral Fellow with the Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada. From 2005 to 2007, he was a Senior Researcher with the Human–Robot Interaction Team, Intelligent Robot Division, Electronics and Telecommunications Research Institute, Daejeon, Korea. He is currently the Assistant Professor with the Department of Control, Instrumentation, and Robot Engineering, Chosun University, Gwangju, Korea. His research interests include human–robot interaction, computational intelligence, biometrics, and pattern recognition. Dr. Kwak is a member of IEEE, IEICE, KFIS, KRS, ICROS, KIPS, and IEEK.