마이크로폰어레이를 이용한 사용자 정보추출

Personal Information Extraction Using A Microphone Array

김 혜 진[1], 윤 호 섭[1]

Hye-Jin Kim[1], Ho-Sub Yoon[1]

Abstract This paper proposes a method to extract the personal information using a microphone array. Useful personal information, particularly customers, is age and gender. On the basis of this information, service applications for robots can satisfy users by offering services adaptive to the special needs of specific user groups that may include adults and children as well as females and males. We applied Gaussian Mixture Model (GMM) as a classifier and Mel Frequency Cepstral coefficients (MFCCs) as a voice feature. The major aim of this paper is to discover the voice source parameters of age and gender and to classify these two characteristics simultaneously. For the ubiquitous environment, voices obtained by the selected channels in a microphone array are useful to reduce background noise.

Keywords : Microphone array, GMM, MFCC, Personal information

1. Introduction

A better understanding of personal characteristics on voice sources will help to make it easy for a variety of applications including cell phones, computers and robots to be applied.

While it is generally believed that human listeners are able to judge a speaker's age within ±10 years, few robots exist that can accomplish this task. There are acoustic correlations with age in every phonetic dimension, and their relative importance in age perception has yet to be fully explored (Hollien 1987 [1], Jacques & Rastatter 1990 [2], Linville 1987 [3], Ptacek & Sander 1996 [4], Schöts 2003 [5])

Other attempts to predict age include Minematsu et al. 2003 [6], who tested age prediction with 30 listeners for approximately 400 male speakers and achieved 90% accuracy. In addition, they found that it was easy to classify older speakers (82.5% for female and 94.4% for male speakers) but that it was more difficult to judge younger speakers (0.97% for female and 52.2% for male speakers). Another effort to recognize elderly users (Müller C. et al., 2003[7]) was able to classify non-elderly users (85.6%) from elderly users (98.3%) and categorized female and male groups at 74.0% and 86.1%, respectively.

[1] 한국전자통신연구원

While previous techniques have primarily been applied to a computer environment, this paper presents a method adaptive to a ubiquitous robot environment providing useful services for robot users. User diversity occurs in various situations with a wide range of needs and capabilities. Elderly people often find their usages of robots difficult due to cognitive disabilities associated with age-degenerative processes. Therefore, it is necessary that a robot acts in a more specific manner as they propose suggestions to make elderly users feel more comfortable. On the other hand, children prefer an active attitude and often want faster services.

Under the ubiquitous environment, robustness to noise has been an important issue. There are lots of papers to solve the problem using multiple microphone array[8-9].

The focus of this paper is an analysis of voice source characteristics related to age and gender, of which dependencies are evaluated by a microphone array.

2. Voice Feature and Age/Gender Dependences

There are many speech features related to speech such as pitch, energy, Teager energy, jitter, shimmer, rate of

speech, formant, HNR, PLCC and MFCC.

Müller C [7], previously mentioned above, introduced jitter and shimmer as appropriate features to determine the age and gender of a speaker. Dave et al. [10] proposed the harmonics-to-noise ratio (HNR), a measure of the amount of noise in a speech signal, as acoustic features that could be used to identify a person's gender and age. In addition, Iseli [11] suggested fundamental frequency (F_0), the open quotient ($H_1^* - H_2^*$) and the concept of spectral tilt ($H_1^* - A_3^*$) as age- and gender-dependent parameters.

These voice features are sufficiently useful to reveal the characteristics of gender and age, but some of the features are limited with vowels and some are insufficient for environmental noise. In this paper, MFCC, a well-known feature in speech recognition and speaker recognition, is proposed to address the best attributes of age and gender.

2.1 Mel Frequency Cepstral Coefficients (MFCCs)

MFCC is a well-known feature extraction method used to recognize aspects of speakers, speech, and emotion. The advantages of MFCCs are that it is a more robust method in terms of noise and spectral estimation errors compared to other factors. Voice sources include an impulse train and a noise source. The impulse train is generated by the vibration of a human organ. The glottal slit and resonance in the vocal tract create distinct speech sounds. On time domain, each source contributes to generate voice features. Considering x[n] as speech data, v[n] a vocal tract, g[n] a glottal flow, and p[n] speech data can be written by the following equation in the time domain:

$$x[n] = v[n] * g[n] * p[n] \quad (1)$$

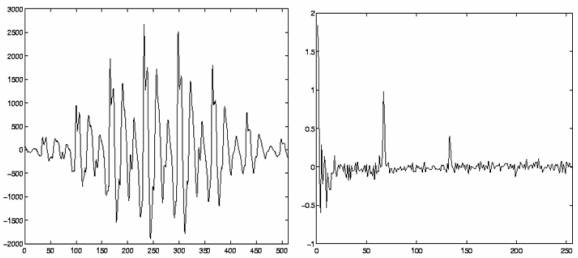
This speech data is filtered by preemphasis, segmented by a hamming widow, and transformed by FFT. In a frequency domain, the equation becomes

$$X(\omega) = V(\omega)G(\omega)P(\omega) \quad (2)$$

To obtain MFCCs, the speech data passes a Mel-scaled filter bank. The Mel-scaled filter bank models the cochlea and has a non-uniform bandwidth.

$$X'(\omega) = V'(\omega)G'(\omega)P'(\omega) \quad (3)$$

Following this, the logarithm of the data is used to separate each part of the source.



(a) Original speech (b) Speech after MFCC

Fig. 1. Comparison of original speech to transformed speech by MFCC.

$$\log X'(\omega) = \log V'(\omega) + \log G'(\omega) + \log P'(\omega) \quad (4)$$

Finally, the Discrete Cosine Transform (DCT) is used to reduce noisy information and abstract speech information. As shown in Fig.1 b), most information is focused on the front parts, while data are distributed all around in Fig.1(a).

3. Classifier of Age and Gender: GMM

The Gaussian Mixture Model is employed as a classifier of age and gender. The Gaussian Mixture model (GMM) represents a single state model of the Hidden Markov Model (HMM), which is a popular technique in speech recognition. GMM can create a model for a nonlinearly distributed class by adapting numerous mixtures while previous studies ([1]-[5]) have shown that it is difficult to decipher the characteristics of age. For age and gender classification, three tests are formulated: (1) age test (A = {adult, child}), (2) gender test, (G = {male, female}), and (3) age-gender test (S = {male adult, female adult, and child}). It is assumed that gender characteristics are generally revealed after the adolescent period. These tests are conducted using the maximum posterior probability from GMM model for each group.

4. Speech Data

To reflect the characteristics of the robot environment, it is necessary to collect data in a robotas well. ETRI-VoiceDB2006 was captured in Wever, a ubiquitous robot. As a robot is mobile, circumstances around robots vary according to time and distance changes. For instance, the distances between a robot and a speaker can change nearly continually. To evaluate the distance factor's effect on age prediction, data were collected at several ranges: 1m, 3m and 5m. The relative locations between a robot and a user are also needed to be considered. To simplify this, two positions were set: a frontal view and a diagonal view.

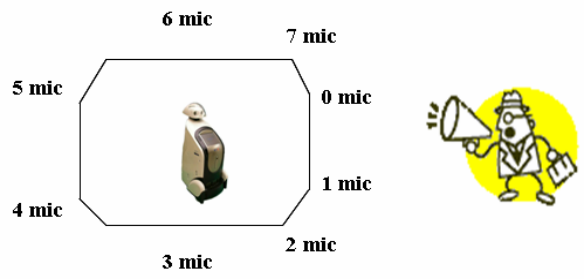


Fig. 2. Multi-array microphones in Wever

In addition, multi-array microphones were used. One advantage of a robot is that it is possible to use multiple microphones. Wever is equipped with eight microphones, as shown in Fig.2.

To short, ETRI-VoiceDB2006 was constructed to reflect robot-specific characteristics such as the gap between a robot and a user, the position of eight numbers of microphones, and the directional effect of voice. ETRI-VoiceDB has nearly the same number of adult speakers and child speakers, at 30 and 28, respectively. The ages of the adults ranged from 22 to 45 years old while the children were aged either ten of eleven; thus, they were pre-adolescents.

5. Experiments

The proposed method was tested using ETRI-VoiceDB06, as previously discussed. Fig.3 presents the entire procedure concerning the classification of age groups. First, the robot, Wever, captured speech data. These data were then transmitted to the main server. The server was used to compute the processes and return the output to the robot. In this way, the end-point of the speech data was detected and age groups were classified in the server system. In addition, a robot is capable of adding numerous noises to the speech data parts such as sensors and motors; therefore, it was necessary to add a noise-elimination algorithm in the form of a Winer filter. From this enhanced data, twelve-dimensional MFCCs as well as energy were extracted as speech features. To classify the data, GMM was applied assuming that the covariance matrices were diagonal in order to improve the computation efficiency. The number of mixtures varied from 1 mixture to 512 mixtures. The effect on the position of microphones in addition to the relative location and interval between the robot and the user were checked.

6. Results

For training data, the set of speech data with a 1m

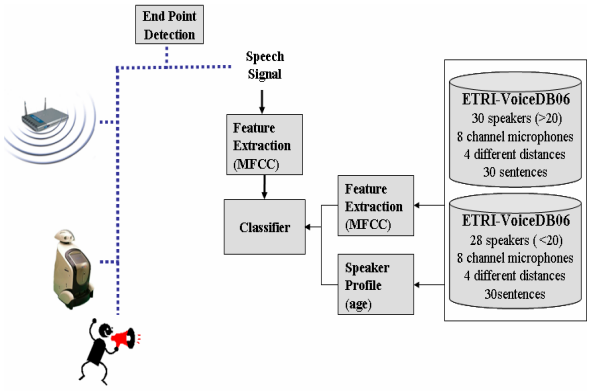


Fig. 3. Age-classification procedure

Table 1. Results: prediction accuracy with various machine-learning techniques and voice features

Feature	Jitter & Shimmer	HNR	MFCC
Classifier	ANN	MLP	GMM
Gender	81.09%	94.4%	94.9%
Age	96.57%	-0.1 year error, 6.86 std	94.6%

interval and a frontal view was chosen. Three types of tests were conducted. Their results are shown in Figs.4-6. For gender classification, the accuracy was 94.9% and from the age test, it was 94.6%. Table 1 shows the results with regard to the predictive accuracy using different features and various machine-learning techniques. This enabled the interpretation of the results comparing to other methods. The previous features and methods presented cling to specific classes: the method using ANN and jitter and shimmer features are suitable for a classification of age but not suitable for gender, while the reverse is true for the method involving HNR features and MLP. However, the proposed method has good accuracy in both cases over 90%.

Here, three group classifications were tested: S = {adult_female, adult_male, child}. Tastes of customers largely depend on these three groups. Given that that MFCC features were confirmed to be effective for classifying both gender and age, S group classification was applied. As seen in Fig. 6, 98.96% was achieved for the child class, 89.89% was achieved for the female class and 77.77% for the male class.

The correlation between the position of the voice data and that of a talker was also tested, as the position of each microphone in a mobile robot plays an important role in the reliability of the classifications. The locations of microphones in a robot can influence the classification capability in many ways. Figs.6 and 7 show the age

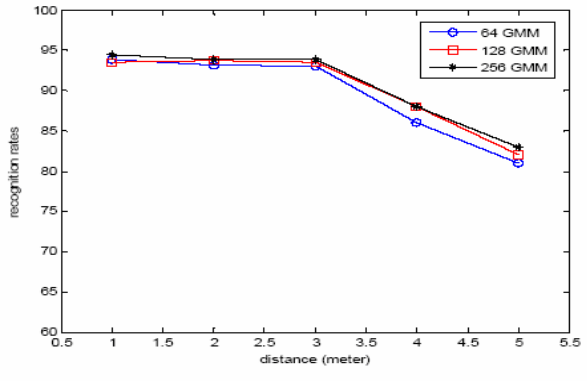


Fig. 4. Gender classification rate using GMM

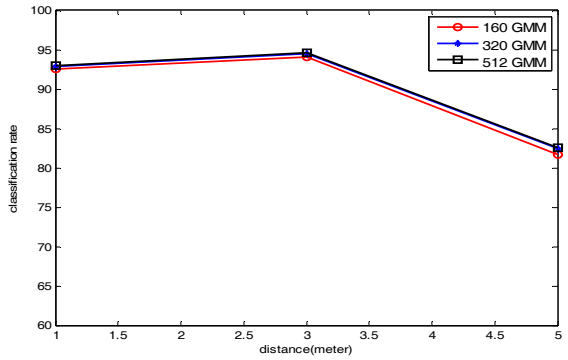


Fig. 5. Age classification rate using GMM

recognition with different positions of the microphones. Fig.6 used microphones in channels 0, 3, 5 and 6, and Fig. 7 used them in channels 0, 1, 2, and 7. The curve of Fig.6 is concave while that that of Fig.7 is convex. As the levels of accuracy using eight microphones are identical, the curve shape in Fig.6 implies that the fewer the number of channels, the more accurate the classification is. As seen in Fig. 2, microphones 0, 3, 5, and 6 cover all directions of the robot, while microphones 0, 1, 2, and 7 are concentrated on the frontal side. The relationships and the number of microphones in these two figures were investigated further. The result implied that fewer microphones would give less reliable output, as a speaker can be located in different places.

The use of multiple microphone systems includes speech classification of age/gender. Our experiment showed that speech data from microphones chosen selectively, achieved higher identification rate that those from all microphones. Fig.8 showed two spectrums of which signals were collected in the same time and different channels. The signal from channel 0 is close to the original source (see Fig.2) while the signal from channel

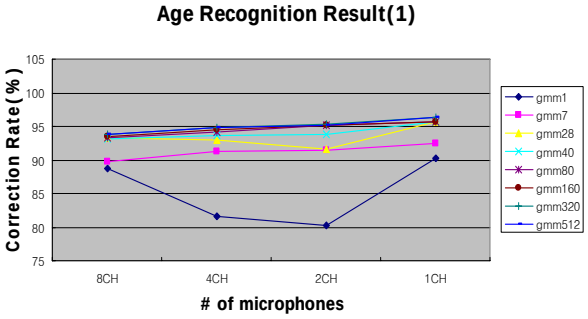


Fig. 6. Age Recognition Result (1) using microphone channels 0, 3, 5, and 6.

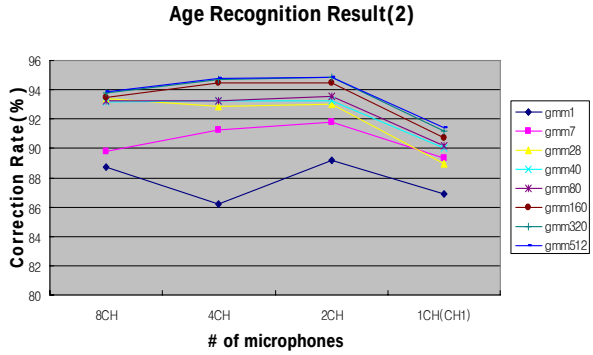
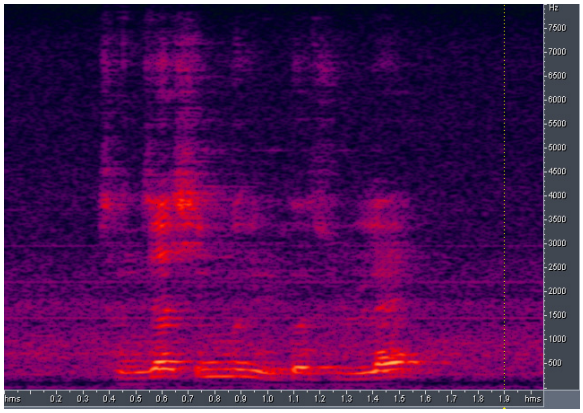


Fig. 7. Age Recognition Result (2) using microphone channels 0, 1, 2, and 7.

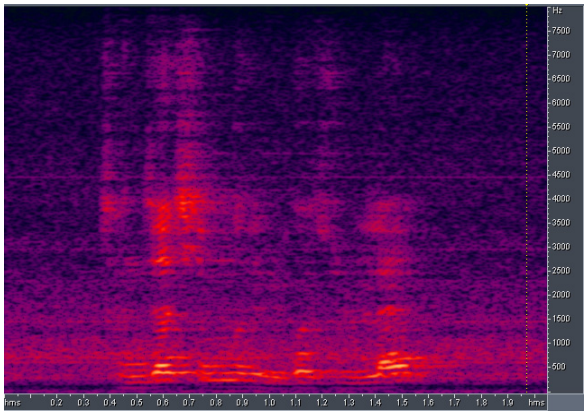
5 is the opposite side of the original one. Although both signals were archived in the same situation, Fig.8 revealed that the CH0 signal is less contaminated than the CH5 signal. Also, Fig.9 implies that speeches of all channels do not give positive effect on identification. It is generally believed that more channels can give more information. In our experiment, eight channels were used. If the common belief was true, the green lines in Fig.9 would always be located in higher position than blue and red lines. Rather, channels selectively chosen showed better identification rate. The total number of identification test samples is 14400 using ETRI-VoiceDB2006 and 1602 samples were failed. The y- axis in Fig.9 shows re-corrected number of samples among misclassified ones. Here, the microphones were picked according to its closeness to the original source.

7. Discussion

In this paper, age and gender classification method is presented for a home service robot. It was found that MFCC features and GMM could classify age groups as



(a) The signal from the channel 0(CH0)



(b) The signal from the channel 5(CH5)

Fig. 8. From the same source, signals from channel 0 (a) and channel 5 (b) were compared. Identification of the (a) signal was correct while the identified speaker using the (b) signal was wrong.

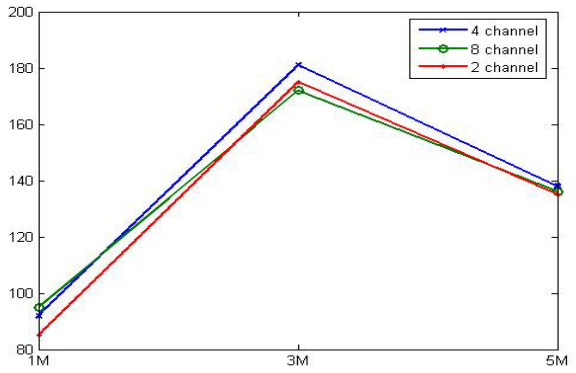


Fig. 9. Comparisons of identification accuracy according to the number of microphones. The original signals is located in the frontal direction. The 4 channels are CH0, CH1, CH2, and CH7 ; The 2 channels are CH0, and CH1.

well as gender groups. Moreover, mixed groups including adult females, adult males and children were successfully classified. Simultaneously, classification of the three groups has advantages such as a reduction of computational cost and time.

In addition, it was determined that the proposed method can be applied to a ubiquitous environment. For instance, a user can give orders to, or communicate with, an intelligent robot, even over a great distance. In addition, the proposed method can also classify age and gender groups correctly although the user may be positioned in diverse viewpoints, such as at the frontal side or diagonal side of a robot.

Finally, the findings imply that signals from selectively chosen channels are more useful in the identification issue. It was confirmed that microphone selection provides better accuracy even when the all signals are used. Moreover, in the case of URC robots, which use the main server and wireless network to overcome the robot's computing power, the microphone selection according to the source position is more powerful because it can reduce the amount of signals that should be transmitted to the server through wireless network. The future work is how to pick microphones channels. We will employ sound source localization method to find proper channels.

참 고 문 헌

[1] H. Hollien, "Old Voices, What Do We Really Know About Them?" Journal of Voice, vol. 1, no. 1, pp. 2-13 1987.

[2] R. D. Jacques, & M. P. Rastatter, "Recognition of Speaker Age from Selected Acoustic Features as Perceived by Normal Young and Older Listeners," *Folia Phoniatrica*, vol. 42, pp. 118-124. 1990.

[3] S. E. Linville, "Acoustic-perceptual studies of aging voice in women," *Journal of Voice*, vol. 1, no. 1, pp. 44-48 1987.

[4] Ptacek, P. H. & Sander, E. K. "Age Recognition from Voice," *Journal of Speech and Hearing Research*, vol. 9, pp. 273-277 1966.

[5] Schöts "A First Step from Analysis to Synthesis," *Proceedings of the XVth ICPhS. Barcelona.* Pp. 2585-2588 2003.

[6] N. Minematsu, M. Sekiguchi, K. Hirose, "Automatic Estimation of Perceptual Age Using Speaker Modeling Techniques," *Proceedings of Eurospeech* Geneva. Switzerland 2003.

[7] C. Müller, F. Wittig and J. Baus "Exploiting Speech for Recognizing Elderly Users to Respond to their Special needs," *Proceeding of Europspeech* Geneva

Switzerland. 2003.
[8] Ilyas Potamitis, Huimin Chen, and George Tremoulis, “Tracking of Multiple Moving Speakers With Multiple Microphne Array,” IEEE Speech & Audio processing Vol. 12 No. 5 pp.520-529 September 2004
[9] P. Barger and Scridha Sridharan, “Robust Speaker Identification Using Mutliple-microphone systems,” IEEE Tencon Speech and Image Technologies for Compuing and Telecommunications pp.261-264 1997
[10] T. Dave, F. David and R. Korin “Acoustic Features for Profiling Mobile Users of Conversational Interfaces,” *MobileHCI* 2004, LNCS 3160 pp. 394-398, 2004.
[11] M. Iseli, Y.-L. Shue and A. Alwan, “Age- and Gender-Dependent Analysis of Voice Source Characteristics,” *Proceedings of ICASSP* May 2006

김 혜 진

2001 포항공과대학교 화학공학과 (공학사)
2003 포항공과대학교 컴퓨터공학과 (공학사)
2004~현재 한국전자통신연구원 연구원

관심분야 : Machine learning, Vision and Audio

윤 호 섭

1989 숭실대학교 공학사
1991 숭실대학교 공학석사
2003 KAIST 공학박사
1991~1998 KIST 시스템공학연구원 선임연구원
1998~현재 한국전자통신연구원 u-로봇연구본부 책임연구원

관심분야 : HRI, 영상처리, 음성처리, 패턴인식, 로봇비전