International Conference on Control, Automation and Systems 2007
Oct. 17-20, 2007 in COEX, Seoul, Korea

An application of Speech/Speaker Recognition System for Human-Robot Interaction

Hyun Jo[1], Gyeongho Kim[2] and Youngjin Park[3]

[1] Department of Mechanical Engineering, Korea Advanced Institute of Science and Technology, Daejeon, Korea
(Tel : +82-42-869-3075; E-mail: e.w.smagel@kaist.ac.kr)

[2] Department of Mechanical Engineering, Korea Advanced Institute of Science and Technology, Daejeon, Korea
(Tel : +82-42-869-8211; E-mail: gyeongho.kim@gmail.com)

[3] Department of Mechanical Engineering, Korea Advanced Institute of Science and Technology, Daejeon, Korea
(Tel : +82-42-869-3036; E-mail: yjpark@kaist.ac.kr)

Abstract: We will introduce a real time, robust speech/speaker recognition system for isolated word recognition using distance microphone. Applying proposed system to a robot platform, robust human-robot interaction can be established for reverberant office environments. For computational effectiveness, dynamic time warping algorithm is used for pattern matching. We select the gamma distribution contrary to the conventional Gaussian distribution to model the probability density function of total accumulated distance. By creating reference speeches at different distances, proposed algorithm shows better speech/speaker recognition performance than the case when creating reference speeches at the same distance. Experimental results show that recognition accuracy is more than 99% by creating five reference speeches at different distances in a reverberant office environment.

Keywords: Isolated word recognition, DTW, pitch, MFCC, gamma distribution, and office environment

1. INTRODUCTION

For the human-robot interaction, speech/speaker recognition algorithms can be used for a robot platform. When specified user calls the name of the robot, for example, the robot should recognize what he/she speaks, and who is speaking. In office environments, however, performance of speech/speaker recognition algorithm gravely decreases due to reflections of sound, Lombard effect, environmental noises, etc. To overcome the reverberation of the sound, dereverberation technique and adaptation scheme were investigated. [1-2] Because those two are based on the inverse filter technique [1] and the adaptive filter design technique [2], previous researches are not appropriate to apply real time, recognition system.

In this paper, proposed method is based on the stochastic model of reverberant reference speeches. With concerning reverberation or Lombard effect, simple and robust implementation of recognition system can be practicable by creating reference speeches varying the distance from a speaker to a microphone.

2. DESCRIPTION FOR TOTAL SYSTEM

There are three steps for the real time, speaker/speech recognition.

2.1 Preprocessing

After sampling acoustic pressure with 22.05kHz, total signal is segmented in a bunch of 30ms signals with 50% overlapping and Hanning windowing. And then, energy level and zero-crossing rate are calculated for each frame to determine the exact location of pure speech signal. [3]

2.2 Feature extraction

After preprocessing, feature extraction process is accomplished. There are two feature vectors to recognize human speech: Pitch for the speaker recognition and Mel Frequency Cepstral Coefficient (MFCC) for the speech recognition.

To extract pitches from speech signal, modified autocorrelation method is applied. [4] To extract MFCCs from speech signal, we used HTK's MFCC which is web-published by Dan Ellis. [5]

2.3 Pattern matching

Dynamic Time Warping (DTW) algorithm is applied between two different feature vectors. We use UE2-1 for DTW algorithm and delta we chose is from 0 to 7. [6] Finally, the total accumulated distance D_T between two feature vectors, $R(n)$ and $T(m)$, is calculated by

$$D_T = \min_{\{w(n)\}} \sum_{n=1}^{N} D\left(R(n), T(w(n))\right). \qquad (1)$$

where $w(n)$ is the optimum path.

3. STOCHASTIC MODEL

5 individual speeches are used to create reference speeches. Using Pitches and MFCCs of reference speeches, we can obtain 10 total accumulated distances. And then we model the total distribution as gamma distribution. The gamma pdf is

$$y = f(x \mid \eta, \lambda) = \frac{\lambda^{\eta}}{\Gamma(\eta)} x^{\eta-1} e^{-\lambda x} \quad \text{for} \quad x > 0. \qquad (2)$$

where Γ is the gamma function, η is the shape parameter, and λ is the inverse scale parameter.

Modeling of total accumulated distances is obtained by using the maximum likelihood estimates (MLEs) for the parameters η and λ of the gamma distribution [7]. When the n samples of statistic variables are exist, MLE of $\hat{\eta}$ and $\hat{\lambda}$ can be represented as:

978-89-950038-6-2-98560/07/$15

$$\hat{\lambda} = \frac{\bar{x}(n-1)}{\sum_{n=1}^{n}(x_i - \bar{x})^2} = \frac{\bar{x}}{s^2}, \quad (3)$$

$$\hat{\eta} = \frac{\bar{x}^2(n-1)}{\sum_{n=1}^{n}(x_i - \bar{x})^2} = \hat{\lambda}\bar{x}. \quad (4)$$

where $\bar{x}$ and s are mean and standard deviation, respectively.

The Figs. 1~2 represent histograms of accumulated distances for pitches and MFCCs and gamma pdf of those. In the case of pitch distances, modeling of distribution is as an exponential function. In the case of MFCC distances, however, we can obtain bell shaped gamma distribution.

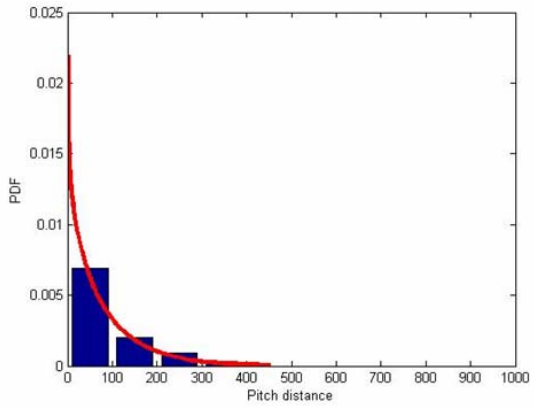

Fig. 1 Histogram of Pitch distances (bar) and a gamma distribution curve (solid line). The reference speech was "iri*h*wa". Reference speech was uttered for five times by 22 years old male at distance 1m.

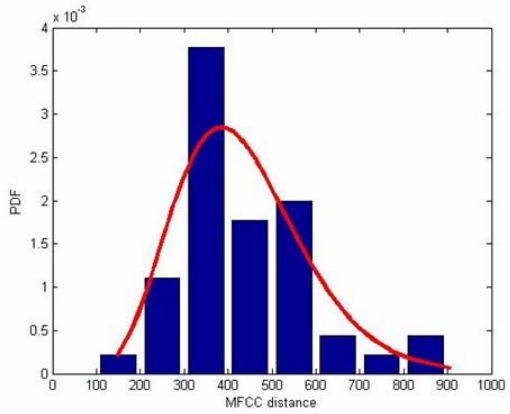

Fig. 2 Histogram of MFCC distances (bar) and a gamma distiribution curve (solid line). The reference speech was "iri*h*wa". Reference speech was uttered for five times by 22 years old male at distance 1m.

When the test speech is uttered at recognition process, we can calculate the probabilities, which are the same number of reference speeches, by using the cdf of estimated gamma distribution. The gamma cdf is

$$p = F(x \mid \eta, \lambda) = \frac{\lambda^{\eta}}{\Gamma(\eta)} \int_0^x t^{\eta-1} e^{-\lambda t} dt \quad \text{for} \quad x > 0. \quad (5)$$

The probability of speaker recognition and speech recognition can be defined as

$$P_{Speaker\_recogntion} = \max\{1 - p_{Pitch}\}, \quad (6)$$

$$P_{Speech\_rocogntion} = \max\{1 - p_{MFCC}\}. \quad (7)$$

Finally, probability of recognition, which means probability of both speaker and speech are recognized, is defined as:

$$P_{recognition} = P_{Speaker\_rocognition} \times P_{Speech\_rocognition}. \quad (8)$$

Decision can be made by setting the threshold of probability of recognition as 0.05.

4. EXPERIMENTS

We made experiments to analyze the performance of proposed system in a office environment. In these experiments, three subjects and five isolated words were used. Three subjects were 22, 26, 27 years old males, respectively. Five isolated words(reference speeches) were "an*j*ʌŋ", "iri*h*wa", "dʒ ʌ riga", "tʃʌŋsohæ", "gman". Averaged reverberation time is 1.07 second when the distance between a sound source and a microphone is 2m. To estimate reverberation time of the room, T_{20} of ISO 3382 was used [8].

Fig. 3 shows particular example when a reference model is made at the distance 1m only and Fig. 4 shows another particular example when a reference model is made by changing the distance between a subject and a microphone – in this case, to make reference model, three reference speeches were measured at the distance 1m and two reference speeches were measured at the distance 2m and 3m, respectively. Figs 3~4 are the case that the subject who made reference model and the subject who uttered test speech are the same. The meaning of y-label in two figures is described in Table 1.

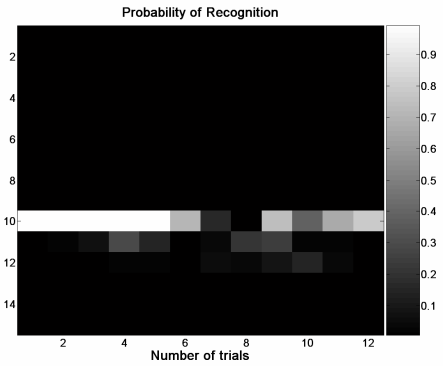

Fig. 3 Probability of recognition when the reference speech is "tʃʌŋsohæ". Reference model is made at distance 1m only.

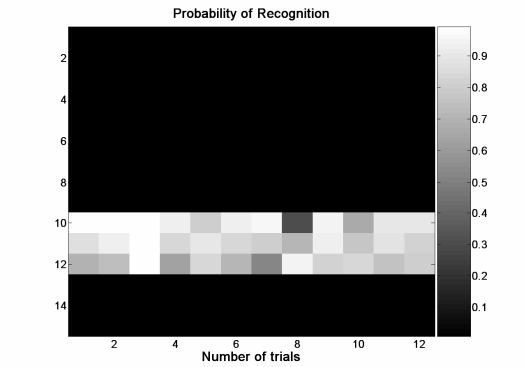

Fig. 4 Probability of recognition when the reference speech is "tʃΛŋsohæ". Reference model is made at distance 1m, 2m, 3m.

Table 1 Description for y-label of Figs. 3~4.

y-label	Distance(m)	Test speech
1	1	an *j*Λŋ
2	2	an *j*Λŋ
3	3	an *j*Λŋ
4	1	iri*h*wa
5	2	iri*h*wa
6	3	iri*h*wa
7	1	dʒ Λ riga
8	2	dʒ Λ riga
9	3	dʒ Λ riga
10	1	tʃΛŋsohæ
11	2	tʃΛŋsohæ
12	3	tʃΛŋsohæ
13	1	gman
14	2	gman
15	3	gman

Making reference model at a specified distance only, (Fig. 3), probability of recognition is relatively higher at the specified distance than the other distances. Making reference model by changing the specified distance (Fig. 4), however, probability of recognition is high at all the distances where reference speech was uttered.

Figs. 5~6 represent the results of averaged recognition accuracy for three subjects and five isolated words. Results obtained in a condition that setting the threshold of probability of recognition as 0.05. Fig. 6 shows better recognition performance at the all distances (1m, 2m, 3m) than Fig. 5. Moreover, this tendency can be observable even when the test speech is uttered at the distance 1m (The first bar of Fig. 6 is higher than the first bar of Fig. 5).

In the case that the subject who made reference model and the subject who uttered test speech are different, probability of recognition is almost zero. Therefore, there are no occasions that wrong speaker was recognized as a proper one.

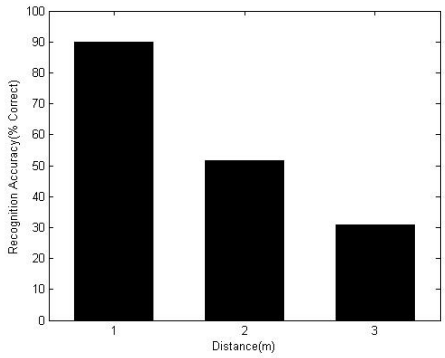

Fig. 5 Averaged recognition accuracy for three subjects and five isolated words. This figure shows the results when reference speeches were made at distance 1m only. In this experimental result, SNR of uttered speech were from 16.5dB to 26.2dB.

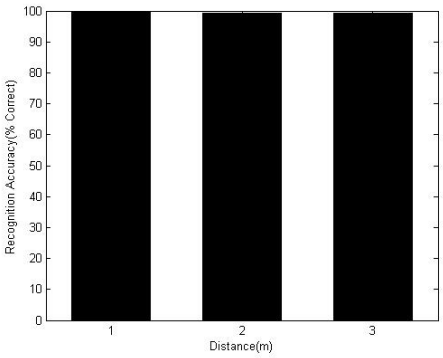

Fig. 6 Averaged recognition accuracy for three subjects and five isolated words. This figure shows the results when reference speeches are made at distance 1m, 2m, 3m. In this experimental result, SNR of uttered speech were from 19.6dB to 25.8dB.

5. SUMMARY & CONCLUSIONS

This paper presents stochastic model for speaker/speech recognition using DTW-based pattern matching method. To create the pdf of the accumulated distances of reference model, we proposed gamma distribution which well estimates distributions of both Pitch and MFCC distances. Five reference speeches is used to create reference model, and high speaker/speech recognition performance can be obtained by letting the threshold of probability as an order of 0.05. Creating reference model at different distances, the recognition accuracy of proposed system is more than 99% at all distances. Therefore, proposed system gives us a simple and robust, isolated word recognition performance in a reverberant room condition.

6. ACNOWLEDGMENT

This work was partly supported by the Korea Science and Engineering Foundation (KOSEF) through the National Research Lab. Program funded by the Ministry of Science and Technology (M10500000112-05J0000 -11210), the Brain Korea 21 Project, and the IT R&D

program of MIC/IITA [2007-S001-01, "Audio-visual Signal Processing SoC for Intelligent Robot"]

REFERENCES

[1] P. Hatziantoniou, I. Potamitis, N.-A. Tatlas, J. Mourjopoulos and N. Fakotakis, "Robust speech recognition in reverberant environments based on complex-smoothed responses," *10th International Conference on Speech and Computer*, 2005.

[2] P. Raghavan, R. J. Renomeron, C. Che, D.-S. Yuk and J. L. Flanagan, "Speech recognition in a reverberant environment using matched filter array processing and linguistic-tree maximum likelihood linear regression adaptation," *The Journal of the Acoustical Society of America*, Vol. 104, Issue. 3, p. 1819, 1998.

[3] L. R. Rabiner and M. R. Sambur, "An algorithm for determining the endpoints of isolated utterances," *The Bell System Technical Journal*, Vol. 54, No. 2, pp. 297-315, 1975.

[4] J. J. Dubnowski, R. W. Schafer, and L. R. Rabiner, "Real-time digital hardware pitch detector," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-24, No. 1, pp. 2-8, 1976.

[5] http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/mfccs.html

[6] L. R. Rabiner, A. E. Rosenberg and S. E. Levinson, "Considerations in dynamic time warping algorithms for discrete word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-26, No. 6, pp. 575-582, 1978.

[7] G. J. Hahn and S. S. Shapiro, *Statistical Models in Engineering*, John Wiley & Sons, Inc., NY·London·Sydney, 1967, pp. 87-88.

[8] *Acoustics – Measurement of the reverberation time of rooms with reference to other acoustical parameters*, STD.ISO 3382-ENGL, 1997.