Analysis of Speech Signals Depending on the Microphone and Microphone Distance

*Jong-Mok Son, *Young-Ho Son, *Hong-Seok Kwon, *Chi-Su Kim and *Keun-Sung Bae

*This work was supported by ETRI(Electronics and Telecommunications Research Institute) in Korea.

Abstract

Microphone is the first link in the speech recognition system. Depending on its type and mounting position, the microphone can significantly distort the spectrum and affect the performance of the speech recognition system. In this paper, characteristics of the speech signal for different microphones and microphone distances are investigated both in time and frequency domains. In the time domain analysis, the average signal-to-noise ratio is measured for the database we collected depending on the microphones and microphone distances. Mel-frequency spectral coefficients and mel-frequency cepstrum are computed to examine the spectral characteristics. Analysis results are discussed with our findings, and the result of recognition experiments is given.

I. Introduction

There are many sources of acoustical distortion in the speech signal such as additive background noise, room reverberation, and so on. Aside from the additive contamination due to noiselike signals, the speech signal inevitably undergoes a series of spectral distortions before being recorded and processed for speech recognition. Since a microphone is the first link in the speech recognition system, it can significantly distort the speech signal depending on its type and mounting position. Therefore, when the microphone used in testing is different from the one used during training of the reference patterns, the mismatch in the spectrum between them may become one of the major problems. As the severe performance degradations incurred by microphone variations become apparent, researchers have begun to pay attention to the issue of microphone robustness.[1-3] Most common methods are preprocessing techniques that apply signal processing algorithms to the recorded speech in order to compensate for microphone variations before input to the speech recognition system.[2,3] Other methods are usually trying to make up for the mismatched training and testing conditions as part of the speech recognition process.[4]

For practical use of speech recognition system, a system that is robust to the environmental changes including the variations of microphones is really needed. Thus as a basic research for it, in this paper, we investigated the influence of microphone and microphone distance on the characteristics of speech signal. We arbitrarily picked four PC systems having different microphones. Two kinds of notebook PC having internal microphones and a desktop computer with two kinds of external microphones are used for data collection. For the database we collected, then, the average signal-to-noise ratio and mel-frequency spectral characteristics were analyzed with respect to each system and microphone distances. In the next section, we discuss the speech database collected for analysis and recognition experiments. Experimental results are presented with discussions in section III. We conclude by summarizing our observations in section IV.

* Kyungpook National University

Manuscript Received : November 4, 1998.

II. Data Collection and Analysis

2.1 Data Collection

We collected isolated-word speech data from three male and two female speakers. Data collection was done for four PC systems having different microphones, and microphone distance was varied from 10 cm to 50 cm increasing 10 cm at each recording. All the speakers were not familiar with these type of data-entry operation. The speech signal was sampled at 16 kHz with 16 bits quantization per sample. Four PC systems having different microphones are, hereafter, referred to system A, B, C, and D, respectively, for our convenience. The condenser microphone in system C is the most expensive one, and the dynamic microphone in system D is the cheapest one among them.

System A : Samsung notebook PC with an internal microphone (System Model : SPC5900RT)

System B : Sambo notebook PC with an internal microphone (System Model : TG 220DB)

System C : Desktop PC with a Sony condenser microphone (Mic. Model : ECM221)

System D : Desktop PC with a rod-type dynamic microphone

Each speaker uttered previously selected 10 Korean words three times at each recording. Since we have four systems and five cases of microphone distance, it comes to 3,000 utterances from 5 speakers. We then divided them into 3 groups according to the order of three times of utterances for the same word. Due to logistical problems, the recordings for four systems were made not simultaneously but separately at different times. This may introduce a certain amount of variability in our data. We also adjusted the input gain, if needed, to obtain an appropriate signal level at each recording. All the data collection was done in the quiet laboratory environment, and it was maintained during the recording.

2.2 Data Analysis

To begin with, we estimated the average signal-to-noise ratio (SNR) for different systems and microphone distances. The SNR is a common measure to compare the quality of speech signals. The definition of SNR used in this study is given in eq.(1). For all the utterances, speech and silence regions were labeled manually by displaying and listening the waveform segments. Then the average signal power on the sample basis was computed for all the speech segments. Similarly, the average noise power was computed for all the silence regions on the sample basis.

$$\sigma_s^2 = \frac{1}{N} \sum_{i=1}^{N} x^2(i) \tag{1.1}$$

$$\sigma_n^2 = \frac{1}{M} \sum_{i=1}^{M} e^2(i) \tag{1.2}$$

$$SNR_{dB} = 10 \log_{10} \frac{\sigma_s^2}{\sigma_n^2} \tag{1.3}$$

where $x(i)$: speech samples in the voiced/unvoiced region

$e(i)$: speech samples in the silence region

N : total number of samples belonging to voiced/unvoiced region from all the utterances

M : total number of samples belonging to silence region from all the utterances

Mel-cepstrum is a very popular feature parameter used for speech recognition. Taking it into consideration, we examined the mel-frequency spectral coefficient (MFSC) and mel-frequency cepstral coefficient (MFCC) to observe the spectral characteristics of the speech signal depending on the different microphones and microphone distances. To enhance the higher frequency components and reduce the spectral dynamic range, the signal was preemphasized with factor of 0.95 before analysis. In order to obtain the MFSC representation for frequency range from 0 to 8 kHz, 24-channel auditory filter bank was applied to the spectrum of the speech signal on the frame basis. The frame size was set to 20 ms with Hamming window, and the frame moving size was set to 10 ms. Then 2048-point FFT was computed to get the spectrum. Table 1 shows the center frequency of 24-channel auditory filter bank, and spectral characteristic of the filter bank is plotted in Figure 1.[5] A triangular shape filter was overlapped to reduce the effect of discontinuity at the boundary of each band, and filter gain was normalized to have the same gain-bandwidth product.

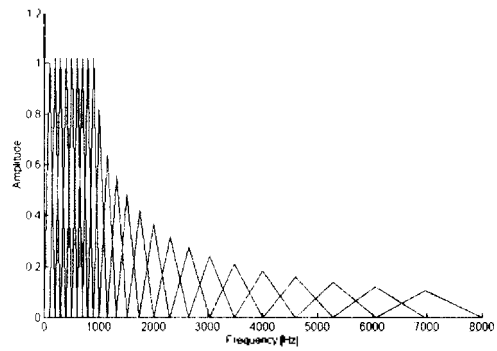


Figure 1. Frequency characteristics for the MFSC filter bank.

Table 1. Center frequency of 24-channel auditory filter bank.

Channel	Center frequency	Channel	Center frequency
1	100 Hz	13	1516 Hz
2	200 Hz	14	1741 Hz
3	300 Hz	15	2000 Hz
4	400 Hz	16	2297 Hz
5	500 Hz	17	2639 Hz
6	600 Hz	18	3031 Hz
7	700 Hz	19	3482 Hz
8	800 Hz	20	4000 Hz
9	900 Hz	21	4595 Hz
10	1000 Hz	22	5278 Hz
11	1149 Hz	23	6063 Hz
12	1320 Hz	24	6964 Hz

From the MFSC, the MFCC is obtained using the eq. (2).[6,7] In our experiment, N was set to 24 and M was set to 14.

$$Y[i] = \sum_{j=1}^{N} X[j] \cos[i(j-\frac{1}{2}) \frac{\pi}{N}], \quad 1 \le i \le M \tag{2}$$

where $X[j]$: MFSC at jth channel
$Y[i]$: ith MFCC coefficient
N : number of MFSC coefficients

III. Experimental Results and Discussion

Table 2 shows the average signal power, noise power, and signal-to-noise ratio measured from all the utterances in the first group of our database. It is shown that system B has the highest SNR and system D has the poorest one. Since systems A and B are notebook PCs with condenser microphone, they may have low fan noise compared to systems C and D that are desktop PCs with condenser microphone and dynamic microphone, respectively. From the results of the system C and D, it is shown that the condenser microphone has background noise suppression capability much better than the rod-type dynamic microphone. Figure 2 shows the SNR for each system depending on the microphone distances. As the microphone distance increases, the SNR decreases gradually but the reduction of SNR shows much difference depending on the systems. As expected, rod-type dynamic microphone showed the largest reduction of SNR as microphone distance increases.

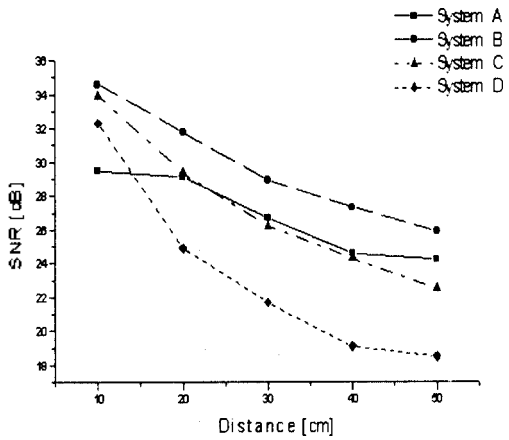
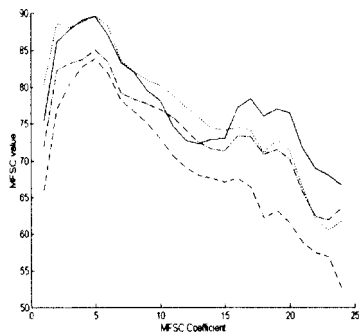


Figure 2. Average SNR depending on the microphone distances.

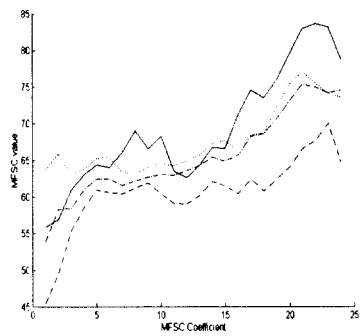
Table 2. Average signal power, noise power, and signal-to-noise ratio.

	signal power [dB]	Noise power [dB]	SNR [dB]
System A	69.5	42.1	27.4
System B	64.0	33.4	30.6
System C	66.0	37.3	28.7
System D	70.2	48.4	21.8

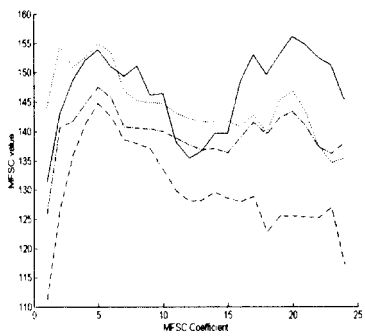
To examine the spectral characteristics, we computed the average MFSC for each system depending on the voiced sound, unvoiced sound, average of voiced sound and unvoiced sound, and silence region. Labeling and segmentation of voiced/unvoiced/silence region were done manually by displaying the waveform, and obscure region for segmentation such as transition region from voiced to unvoiced sounds or vice versa was excluded from the analysis. Figure 3 shows the MFSC characteristics. In Figure 3, we can see that the overall spectral characteristics for each microphone show similar patterns, however, system A shows severe spectral variation compared to other systems. In Figure 3(d), it is seen that system A shows very severe spectral variation in the silence region, especially in some frequency ranges. On the contrary, system C shows very flat spectral characteristics. It is



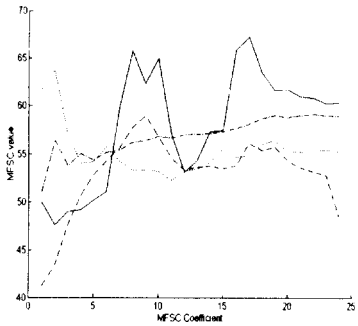
(a) MFSC for voiced sounds



(b) MFSC for unvoiced sounds



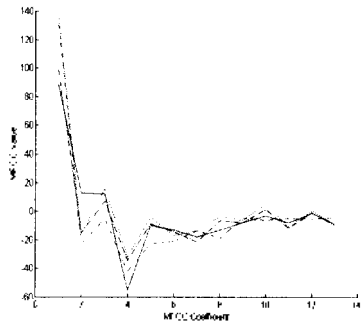
(c) MFSC for voiced/unvoiced sounds



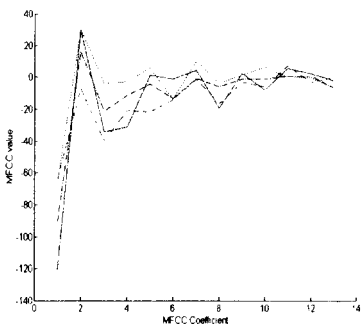
(d) MFSC for silence region

System A : ——— System B : ----------
System C : - - - - - - - - - System D : .....

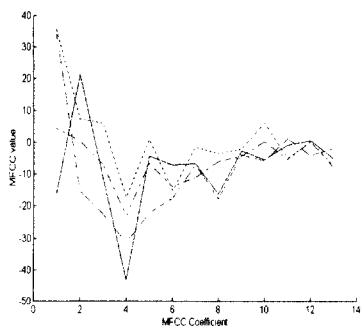
Figure 3. MFSC characteristics depending on the systems.



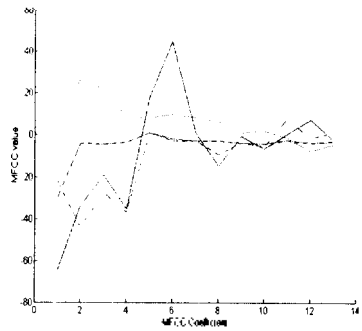
(a) MFSC for voiced sounds



(b) MFSC for unvoiced sounds



(c) MFSC for voiced/unvoiced sounds



(d) MFSC for silence region

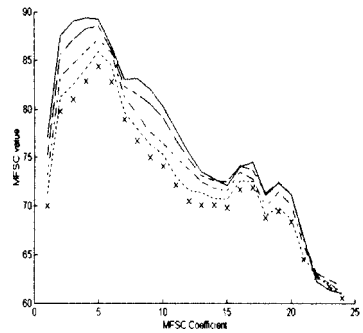
System A : ——— System B : ----------
System C : - - - - - - - - - System D : .....

Figure 4. MFCC characteristics depending on the systems.

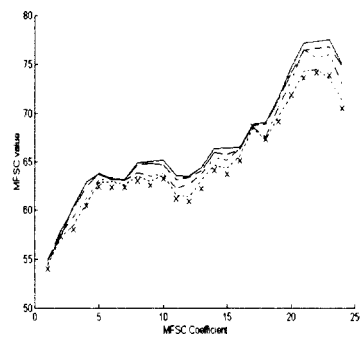
expected that the latter case will influence less than the former in the performance of the speech recognition system. The severe spectral variation in system A has proven to be caused by the system's fan noise. Figure 4 shows the MFCC characteristics. Since the MFCC is computed from the MFSC using eq.(2), it reflects the characteristics of the MFSC. In Figure 4, it is also shown that system A has the largest deviation compared to other systems as shown in Figure 3.

Figures 5 and 6 show the MFSC and MFCC characteristics for different microphone distances. The MFSC showed similar patterns except the mean value, and the MFCC showed almost the same shape for different microphone distances. In other words, microphone distance did not affect much the spectral characteristics of the signal.

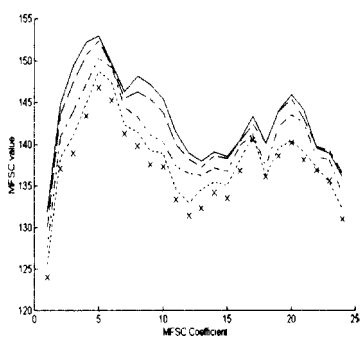
We did the recognition experiments with our database. The word recognition system was constructed by cascading phoneme-like unit based HMM modules properly, and recognition engine was offered from the ETRI(Electronics and Telecommunications Research Institute, Korea). Though our database is made up of 10 words selected from the vocabulary list of the system, the system was not trained by our database. Table 3 shows the test results for each system depending on the microphone distances. Since the recognition experiment was done for all the data we collected, total number of test utterance comes to 3,000. From the Table 3, it is shown that system C that is a desktop PC with a condenser microphone has the highest recognition rate and system A that is a notebook PC with an internal microphone has the poorest result among them. Considering the spectral characteristics shown in Figures 3 and 4, this result is consistent with the previous analysis results as expected. As the microphone distance increases the recognition rate decreases, but not consistently. In



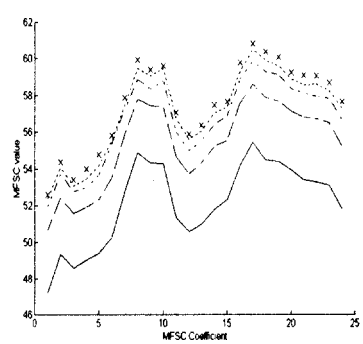
(a) MFSC for voiced sounds



(b) MFSC for unvoiced sounds



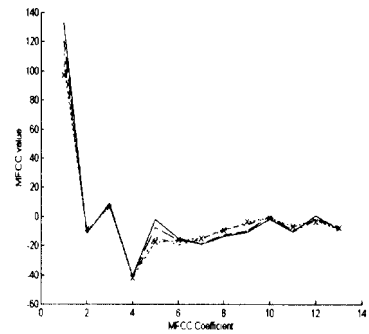
(c) MFSC for voiced/unvoiced sounds



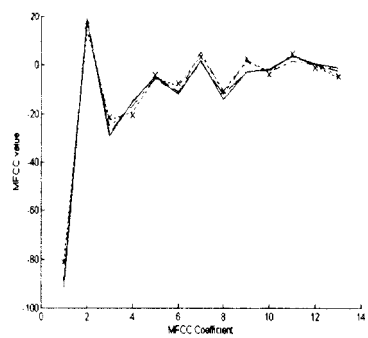
(d) MFSC for silence region

10 cm : ——— 20 cm : ---------- 30 cm : — · — · — · —
40 cm : ····· 50 cm : xxxxxxxx

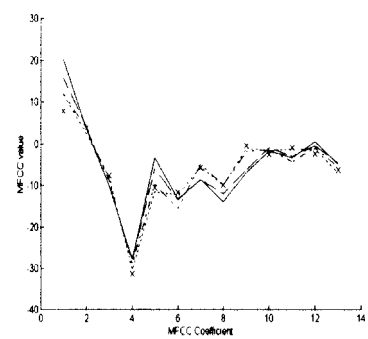
Figure 5. MFSC characteristics depending on the microphone distances.



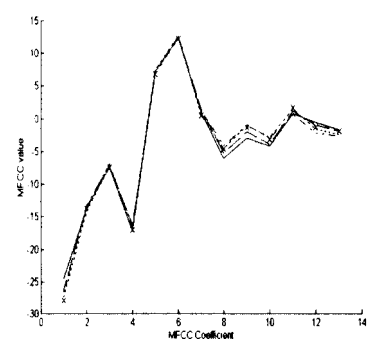
(a) MFSC for voiced sounds



(b) MFSC for unvoiced sounds



(c) MFSC for voiced/unvoiced sounds



(d) MFSC for silence region

10 cm : ——— 20 cm : ---------- 30 cm : — · — · — · —
40 cm : ····· 50 cm : xxxxxxxx

Figure 6. MFCC characteristics depending on the microphone distances.

Table 3. Recognition results depending on the microphone distances.

	System A	System B	System C	System D	Average
10 cm	88.0 %	94.0 %	96.7 %	96.7 %	93.8 %
20 cm	93.3 %	94.0 %	97.3 %	93.3 %	94.5 %
30 cm	88.7 %	91.3 %	92.7 %	91.3 %	91.0 %
40 cm	85.3 %	94.0 %	93.3 %	86.0 %	89.7 %
50 cm	85.3 %	80.0 %	87.3 %	87.3 %	85.0 %
Average	88.1 %	90.7 %	93.5 %	90.9 %	90.8 %

addition, for the microphone distance of 50 cm, lots of recognition errors were caused by the failure of endpoint detection. Therefore, robust endpoint detection algorithm is thought to be an important factor to make up for the variation of microphone distances.

IV. Conclusion

We investigated the influence of microphone and microphone distance on the characteristics of the speech signal, which are related with the performance of the recognition system. The average signal-to-noise ratio is measured for the database we collected depending on the microphones and microphone distances. Mel-frequency spectral coefficients and mel-frequency cepstrum are computed to examine the spectral characteristics. From the analysis, it was shown that a microphone showing much deviation in the mel-frequency spectral coefficients resulted in the poorest recognition rate. Microphone distance did not affect much to the variation of mel-frequency spectral coefficients, but as microphone distance increased recognition errors also increased much and many of them were caused by the failure of endpoint detection.

Acknowledgements

The authors would like to thank Division of Human Interface in ETRI for offering a HMM-based speech recognition system for recognition experiments.

References

1. Subrata Das, Authur N das, David Nahamoo, Michael Picheny, "Influence of Background Noise and Microphone on the Performance of the IBM TANGORA Speech Recognition System," *IEEE International Conf. on Acoustics, Speech, and Signal Processing.*, Vol. II, pp. 71-74, April, 1993.
2. Anastasios Anastasakos, Francis Kubala, John Makhoul, Richard Schwartz, "Adaptation to New Microphones using Tied-Mixture Normalization," *IEEE International Conf. on Acoustics, Speech, and Signal Processing.*, Vol. 1, pp. 433-436, April, 1994.
3. Subrata Das, Arthur N das, David Nahamoo, Michael Picheny, "Adaptation Techniques for Ambience and Microphone Compensation in the IBM TANGORO Speech Recognition System," *IEEE International Conf. on Acoustics, Speech, and Signal Processing.*, Vol. 1, pp. 21-24, April, 1994.
4. R. Lippmann, E. Martin and D. Paul, "Multi-Style Training for Robust Isolated-Word Speech Recognition," *Proc. IEEE International Conf. on Acoustics, Speech, and Signal Processing*, pp. 705-708, 1987.
5. Jane W. Chang, "Speech Recognition System Robustness to Microphone Variations," M.S. Science of M.I.T.
6. Joseph W. Picone, "Signal Modeling Techniques in Speech Recognition," In *Proc. IEEE*, Vol. 81, No. 9, pp. 1215-1247, Semtember 1993.
7. Steven B. Davis, "Comparison of Parametric Representatios for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Trans. on Acoustics, Speech, and Signal Proc.*, Vol. ASSP-28, No. 4, August 1980.

▲Jong-Mok Son

Jong-Mok Son was born in Taegu, Korea, in 1974. He received the B.S. degree in electronic engineering from Kyungpook National University, Taegu, Korea, in 1997. He is currently in the course of M.S. degree in electronic engineering at Kyungpook National University. His current research interests include speech analysis, and recognition with HMM.

▲Young-Ho Son

Young-Ho Son was born in Taegu, Korea, in 1970. He received the B.S. degree in electronic engineering from Kyungpook National University, Taegu, Korea, in 1997. He is currently in the course of M.S. degree in electronic engineering at Kyungpook National University. His current research interests are speech analysis, application of wavelet transform to speech signals, and digital signal processing.

▲Hong-Seok Kwon

Hong-Seok Kwon was born in Taegu, Korea, in 1971. He received the B.S. degree in electronic engineering from Kyungpook National University, Taegu, Korea, in 1997. He is currently in the course of M.S. degree in electronic engineering at Kyungpook National University. His current research interests are speech analysis, speech conversion, and digital signal processing.

▲Chi-Su Kim

Chi-Su Kim was born in Taegu, Korea, in 1975. He received the B.S. degree in electronic engineering from Kyungpook National University, Taegu, Korea, in 1997. He is currently in the course of M.S. degree in electronic engineering at Kyungpook National University. His current research interests include digital signal processing, speech analysis and speech/nonspeech classification.

▲Keun-Sung Bae

Keun-Sung Bae was born in Kyungpook,, Korea, in 1953. He received the B.S. degree in electronic engineering from Seoul National University, Seoul, Korea, in 1977 and M.S. degree in electrical engineering from Korea Advanced Institute of Science and Technology, Seoul, Korea in 1979. From August 1984 to May 1989, he joined the Mind-Machine Interaction Research Center at the University of Florida, Gainesville, Florida, U.S.A. He received the Ph.D. degree in electrical engineering from the University of Florida in 1989. Since 1979, he has been with the Department of Electronic Engineering, Kyungpook National University, Taegu, Korea. He is now working there as a professor. His current research interests include wavelet analysis, speech coding, speech recognition, digital signal processing, and digital communication. He is interested in all aspects of signal processing.