THE INFLUENCE OF NOISE ON THE SPEAKER RECOGNITION PERFORMANCE USING THE HIGHER FREQUENCY BAND

Shoji HAYAKAWA and Fumitada ITAKURA

School of Engeering, Nagoya University
Furocho 1, Chikusa-ku, Nagoya, 464-01, JAPAN

ABSTRACT

In our previous studies, we have shown the effectiveness of using the information in the higher frequency band for speaker recognition. However, the energy spectrum of speech in the higher frequency band is weak, except for some fricative sounds. Therefore, it is important to investigate the speaker individual information in that region under noisy conditions. In this study, we examine the influence of additive noises on the performance of speaker recognition using the higher frequency band. Experimental results show that high performance is obtained in the wideband case under many typical noisy conditions. It is also shown that the higher frequency band is more stable against noises than the lower one. For that reason, the higher frequency band gives good performance even if the SNR of the higher frequency region is worse than the lower one.

1. INTRODUCTION

It has been shown that the speaker individual information does exist in the higher frequency range above the telephone band [1]. On the other hand, the energy spectrum in the higher frequency band is relatively weaker than that in the lower frequency band. Therefore it is important to study the effectiveness of using the higher frequency range under noisy conditions. Considering the use of the system under practical conditions, it is desirable to use actual noises to evaluate the influence of noise on the system.

In this study, we investigate the influence of noise on the performance of speaker recognition using the higher frequency band. The following section explains the experimental conditions. Section 3 shows the speaker recognition performance for several kinds of noise and variable frequency bandwidth. Section 4 shows the case of changing the SNR for each frequency band. Finally, Section 5 concludes the whole paper.

2. EXPERIMENTAL CONDITIONS

2.1. Database

The token database consists of 5 Japanese words (with duration varying from about 0.3 to 1.0 second). Each talker of 15 male speakers uttered each word 5 times a session. The recording sessions were made every 3 months over 1 year, in a acoustically isolated room. The words uttered by 18 male impostors were also recorded in the same environment for speaker verification experiments. The recorded utterances were digitized at the sampling rate of 32 kHz.

Four types of environmental noise are used as additive noise. The long-term averaged spectra of these kinds of noise, and a Japanese word "/bakuoN/" uttered by a Japanese male speaker are shown in Figure 1. These spectra are normalized by the total power of the whole band from 0 to 16kHz. These noises were recorded by the same devices (microphone, DAT recorder) which recored the speech sound of the database. Gaussian white noise was also used for comparison.

Table 1. Main sources and noise level of each noise.

Noise	Main sources	dBA
Busy office room	dot impact printer human chattering	62
Hospital lobby	sound of coin human chattering	62
Heavy traffic road	sound of passing cars	75
Computer room	9 work stations 11 hard disks, etc.	56

They were added to the speech signal using the following formula:

$$SNR = 10 \log \left(\frac{\sum_{n=1}^{M} S(n)^2}{\sum_{n=1}^{M} N(n)^2} \right) [dB],$$

where M is the number of samples in speech and noise data, $S(n)$ and $N(n)$ are sampled speech and noise data, respectively.

Table 2. LPC analysis order of each frequency band

Band[kHz]	order	Band[kHz]	order
0-3	12	1-10	30
0-4	14	2-10	22
0-5	16	3-10	20
0-6	22	4-10	14
0-8	26	5-10	12
0-10	32	6-10	8
0-12	36	0-16	48

0-7803-2431-5/95 $4.00 © 1995 IEEE

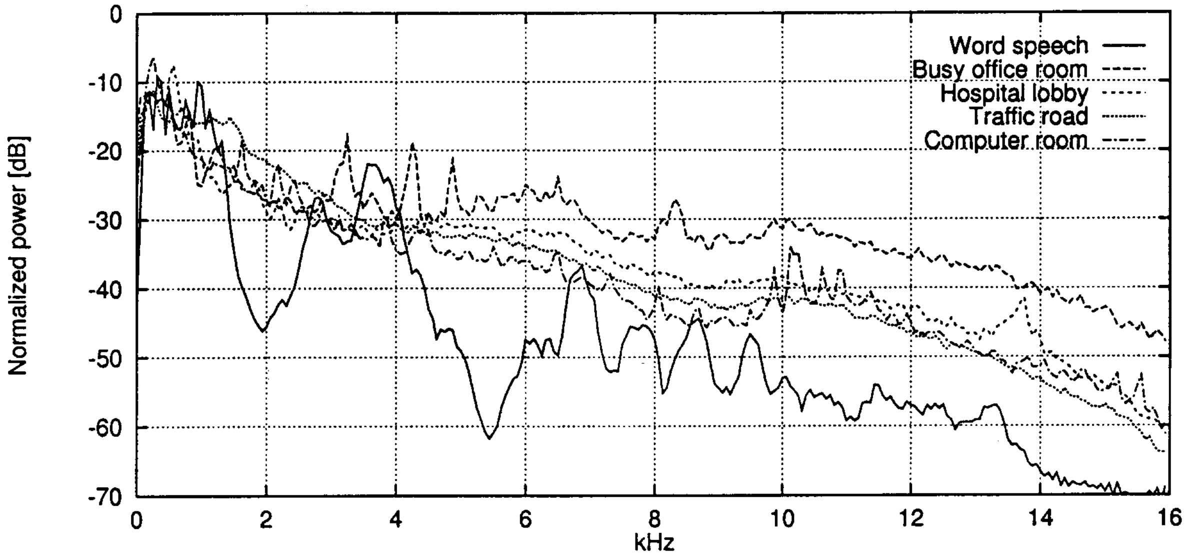

Figure 1. Long-term average spectra of a Japanese word "/bakuoN/" uttered by a Japanese male speaker and five types of noise.

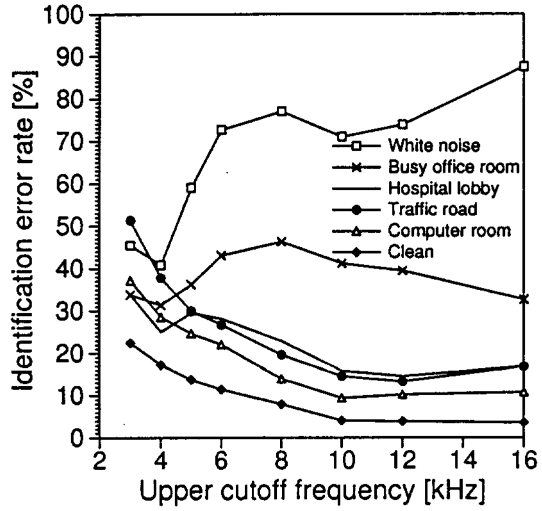

Figure 2. Identification error rates vs. upper cutoff frequency in low-pass processing.

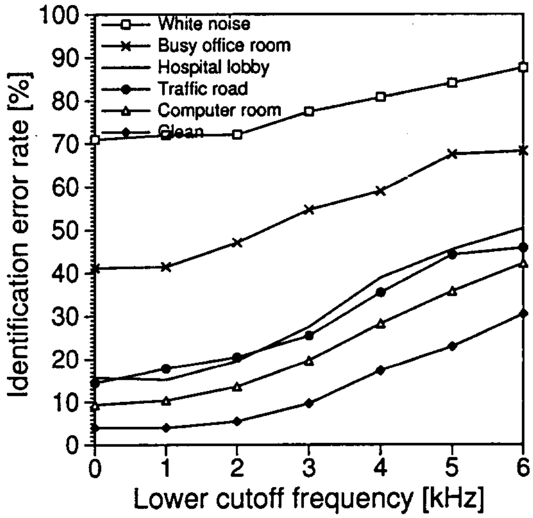

Figure 3. Identification error rates vs. lower cutoff frequency in high-pass processing.

2.2. Speaker recognition system

In both speaker identification (SI) and speaker verification (SV) experiments, an unconstrained-end-point dynamic time warping with modified adjustment window was used. All of speaker recognition experiments are performed in text-dependent mode. The utterances recorded in the first session were used as the reference data, and those of the other 4 sessions were used as the test data. Consequently, a total of 1500 test data were examined. In the speaker verification experiment, the threshold of distance was determined for each speaker to give the same customer rejection rate and impostor acceptance rate.

2.3. Feature extraction

In order to select the frequency range, selective linear prediction is used in all experiments [2]. Cepstral coefficients are calculated by the selective LP analysis using a Hamming window, with a frame shift of 16ms, and a frame length of 32ms. The LPC analysis order of each frequency band was determined by a preliminary experiment and is shown in Table 2. The cepstral coefficient order is twice as many as the LPC analysis order.

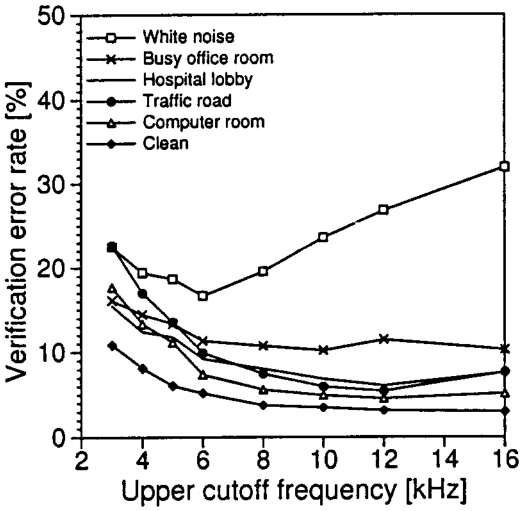

Figure 4. Verification error rates vs. upper cutoff frequency in low-pass processing.

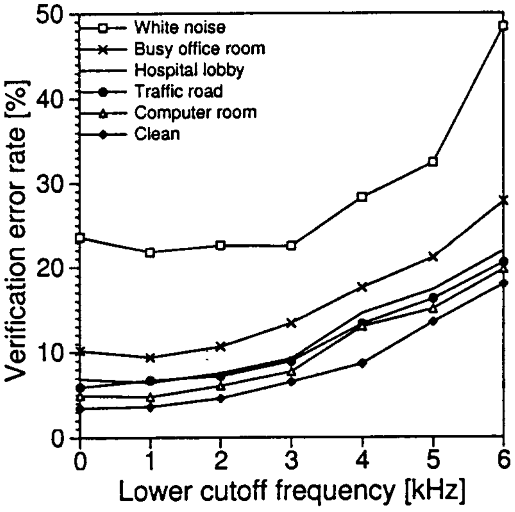

Figure 5. Verification error rates vs. lower cutoff frequency in high-pass processing.

3. SPEAKER RECOGNITION PERFORMANCE FOR SEVERAL KINDS OF NOISES

In the first experiment, the influence of noise on several kinds of frequency band is investigated using the low-pass bands from 0 to 3,4,5,6,8,10,12 and 16 kHz; and the high-pass bands from 10 to 6,5,4,3,2,and 1 kHz to 10kHz. In this experiment we fixed the overall SNR to 20dB, since we consider the situation where the speaker is relatively close to the microphone, and the noise is added as background noise.

Identification and verification error rates in the case of low-pass processing for adding several kinds of additive noise, are shown in Figure 2 and Figure 4, respectively. From these figures, the following performance ranking is found in the case of the frequency band from 0 to 3kHz:

Hospital lobby > Busy office room >
Computer room > White noise >
Traffic road

However, we find in the case of the frequency band from 0 to 10kHz:

Computer room > Traffic road >
Hospital lobby > Busy office room >
White noise

These orders depend on the energy distribution of the different of types of noise. It is also seen that the wider the frequency band is, the higher recognition rate is obtained under soft noise conditions. On the other hand, the effectiveness of using higher frequency band decreases under harsh noise conditions, such as the busy office room and white noise; which have high energy spectrum in the higher frequency region.

Identification and verification error rates in the case of high-pass processing are shown in Figure 3 and Figure 5, respectively. It is seen that the error rate of the frequency band from 0 to 4kHz is the same as that of the frequency band from 4-10kHz in the case of computer room noise. In the case of traffic road noise in which lower frequency spectrum is dominant, the recognition rate of the higher frequency band of 4-10kHz is better than that of its lower counterpart of 0-4kHz.

Table 3 shows the average SNR and standard deviation of both the lower and the higher frequency region for five types of noise. It is observed that harsh noises, such as busy office room and white noise, give relatively low SNR in the higher frequency region. In the case of computer room and traffic road noise, it is interesting to note that the higher frequency band gives the same or better recognition performance than the lower one, while the SNR of the higher frequency region is worse than the lower one

Table 3. SNR of both the lower and the higher frequency region for five types of noise.(dB)

	0-4kHz		4-10kHz	
NOISE	mean	SD	mean	SD
Computer room	20.222	0.107	13.984	4.333
Traffic road	20 233	0.080	11.975	4.316
Hospital lobby	20.295	0.067	9.874	4.304
Office room	21.039	0.204	4.031	4.402
White noise	26.154	0.088	0.629	4.328

Identification error rates of each speaker are shown in Figure 6. Some speakers show more increases of error rates for the lower band noise (speakers #1,3,4,5,9,11,12,14 and 15). Remaining speakers (#2,6,7,8,10 and 13) are more susceptible to the higher band noise. Under noisy conditions, most of the analysed speakers give the best performance for wide band cases.

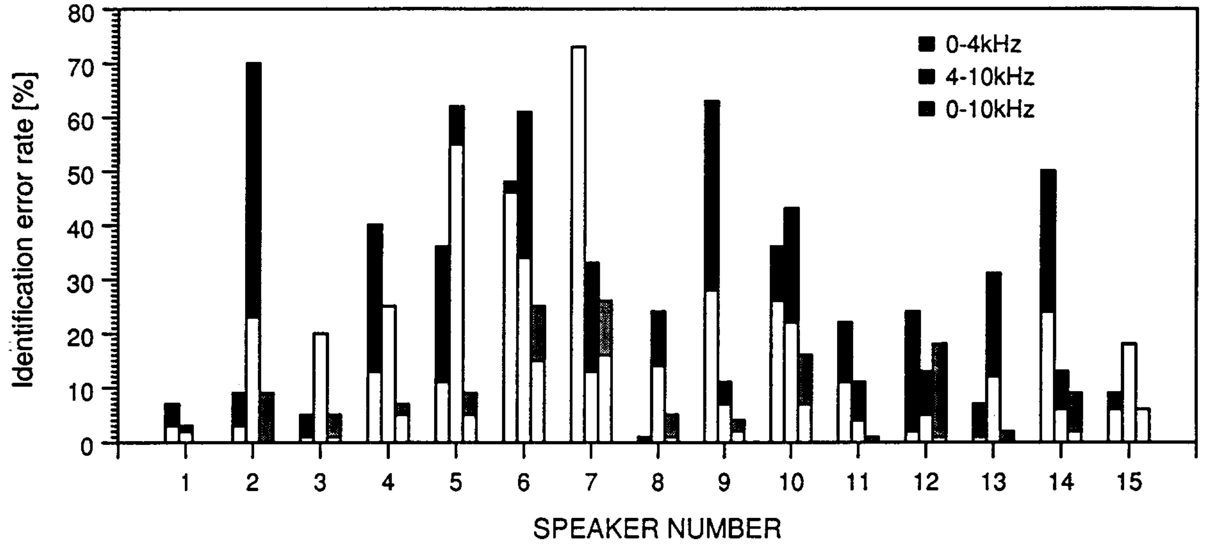

Figure 6. Identification error rates of each speaker for the case when the additive noise is computer room noise. (White boxes show error rates under clean conditions.)

4. THE CASE OF CHANGING SNR FOR EACH FREQUENCY BAND.

In previous experiments, the SNR was calculated for the whole band, namely from 0 to 16kHz, and the SNR of each band depended on the spectral property of noise. In order to investigate the robustness in the lower and the higher frequency bands, white noise is added to give a specified SNR in each frequency band.

Figure 7 shows both identification and verification error rates when the SNR is changed from 0dB to infinity in each frequency band. Error rates (verification and identification) increase rapidly when the SNR is lowered in the case of 0-4kHz whereas the error rates of 4-10kHz increase relatively slowly. Therefore, it can be said that the higher frequency band is more robust to noise than the lower one under the same noisy conditions.

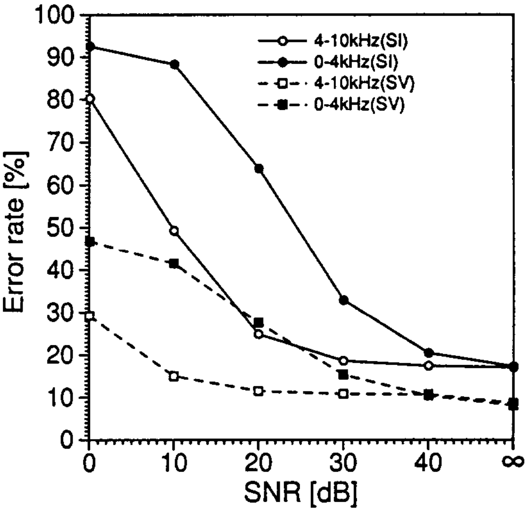

Figure 7. Error rates of SI and SV for the case of changing SNR for each frequency band of both 0-4kHz and 4-10kHz.

5. CONCLUSIONS

The influence of noise on the speaker recognition performance using the higher frequency band was investigated. It was found that the lower the noise energy spectrum in the higher frequency region is, the higher the performance obtained in the case of the wideband signal is. Because the fact that the higher frequency band is more stable than the lower one, the wider frequency band gives better performance even if the SNR in the higher frequency band is worse than in the lower one. We conclude that it is useful to use a 10 kHz bandwidth for speaker recognition under typical noisy conditions when the energy spectrum of noise concentrates mainly in the lower frequency region.

REFERENCES

[1] S.Hayakawa and F.Itakura: "Text-dependent speaker recognition using the information in the higher frequency band", in *Proc. of ICASSP*, pp. 137–140 (1994).

[2] J.Makhoul: "Spectral linear prediction:properties and applications", *IEEE Trans.Acoust.,Speech, & Signal Process.* **ASSP-23**, 3, pp. 283–296 (1975).