FIFTH INTERNATIONAL CONGRESS ON SOUND AND VIBRATION
DECEMBER 15-18, 1997
ADELAIDE, SOUTH AUSTRALIA

Invited Paper

SPEECH SIGNAL ENHANCEMENT BASED ON A SINUSOIDAL MODEL

M. kazama*, T. Ohnishi**, and M. Tohyama**
風間道子　大西崇浩　東山三樹夫

*Acoustic Consultant　**Kogakuin University

ABSTRACT
This article describes a method for enhancing speech under noisy conditions that is based on a sinusoidal wave model. The significant sinusoidal components necessary for intelligible speech recovery were investigated using the STFT. Since intelligible speech sounds can be synthesized using only a few dominant sinusoidal waves, a spectrum peak-picking method based on the sinusoidal model is proposed for enhancing speech signals. Five major sinusoidal components are extracted for speech signal reconstruction from a noisy speech using the STFT. This proposed method increases the signal-to-noise ratio by about 5 dB, but the tonal quality is not acceptable. The quality is improved by tracking the fundamental frequency and its harmonics of speech signals using the harmonic sinusoidal model.

1. INTRODUCTION

Noise reduction for speech signals has generally been achieved through spectral subtraction[1]. However, under nonstationary noisy conditions, it is difficult to identify speech or silence with this method. Speech transformation based on a sinusoidal model[2] is more promising. We propose a spectrum peak-picking method based on this model to achieve the noise reduction. In this method, speech signals are synthesized using the major sinusoidal components extracted from the noisy signal.

Consonants can be modeled as random noise, while vowels have a formant structure. Fukuda et al[3] analyzed the fundamental frequencies and estimated the harmonics of vowels using the MW-STFT. This MW-STFT method, however, requires a lot of computing time and is not effective for synthesizing consonants. Thus, we tried to improve the intelligibility as well as the quality by using both the spectrum peak-picking method and the MW-STFT method.

The sinusoidal model of speech is described in Section 2. Section 3 discusses the spectrum peak-picking method based on the sinusoidal model as well as MW-STFT. In Section 4, experimental results of speech enhancement and noise reduction are demonstrated.

2. SINUSOIDAL MODEL OF A SPEECH[2]

We assume that the envelope of a speech waveform conveys the important information[4] and thus an intelligible speech signal can be recovered by reconstructing the envelope. The envelope can be expressed by the dominant sinusoidal components that are obtained by spectrum peak-picking. Spectrum peak-picking is performed by picking the sinusoidal components that take the local maximum of the power spectra obtained by the STFT.

The power spectrum of the signal is analyzed by the STFT as shown in Fig. 1. We used a Japanese female speech signal sampled at 16 kHz. The frame length is 512 points (32 msec) and a rectangular window is applied. Each frame is overlapped as shown in Fig. 2. The significant sinusoidal components are found by picking the major power spectrum components from among the largest components.

Figure 3(a) shows the original speech signal and (b) is the waveform reconstructed by the most significant sinusoidal component. We can see that the reconstructed envelope of the waveform is almost the same as the original one. This reconstructed speech signal is quite understandable, but not completely intelligible, particularly for the consonant parts.

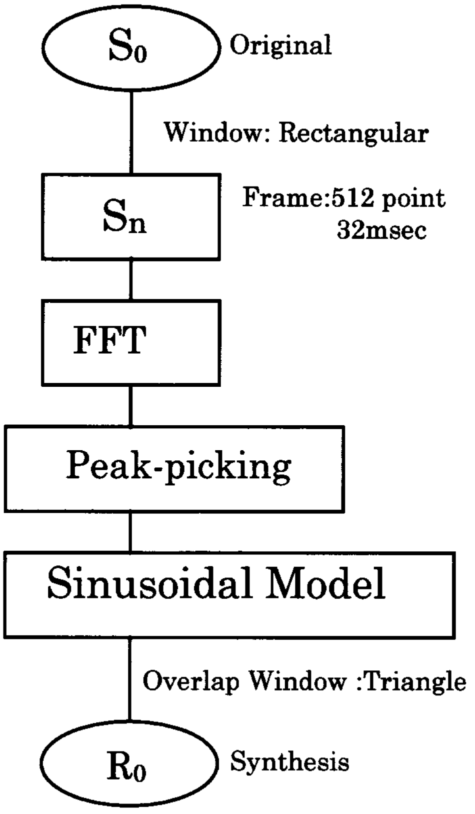


Fig. 1
Process of the peach-picking and STFT

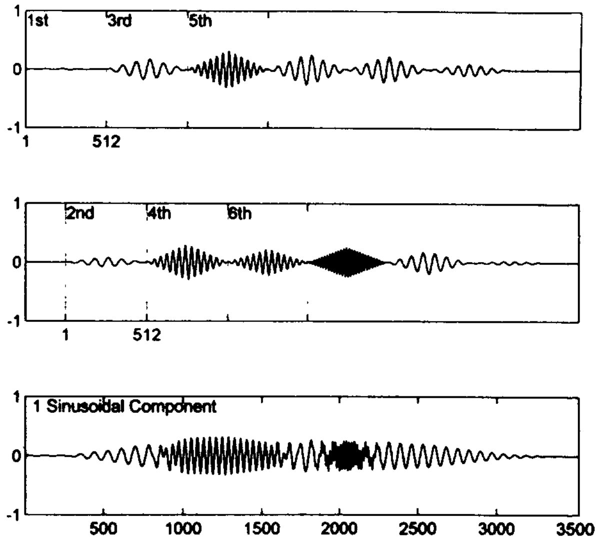
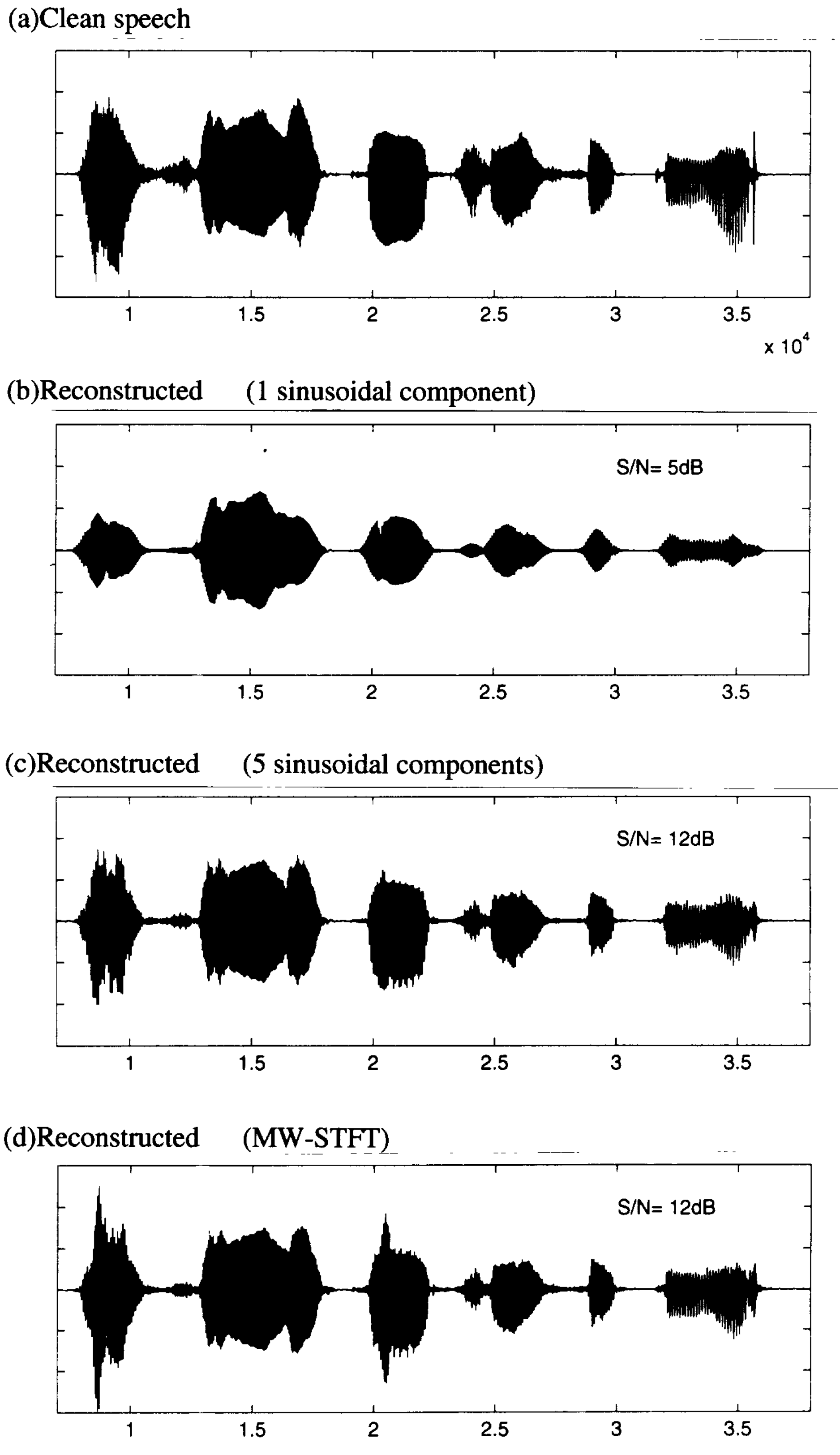


Fig. 2
Frame-based synthesis of a speech waveform using a single sinusoidal component



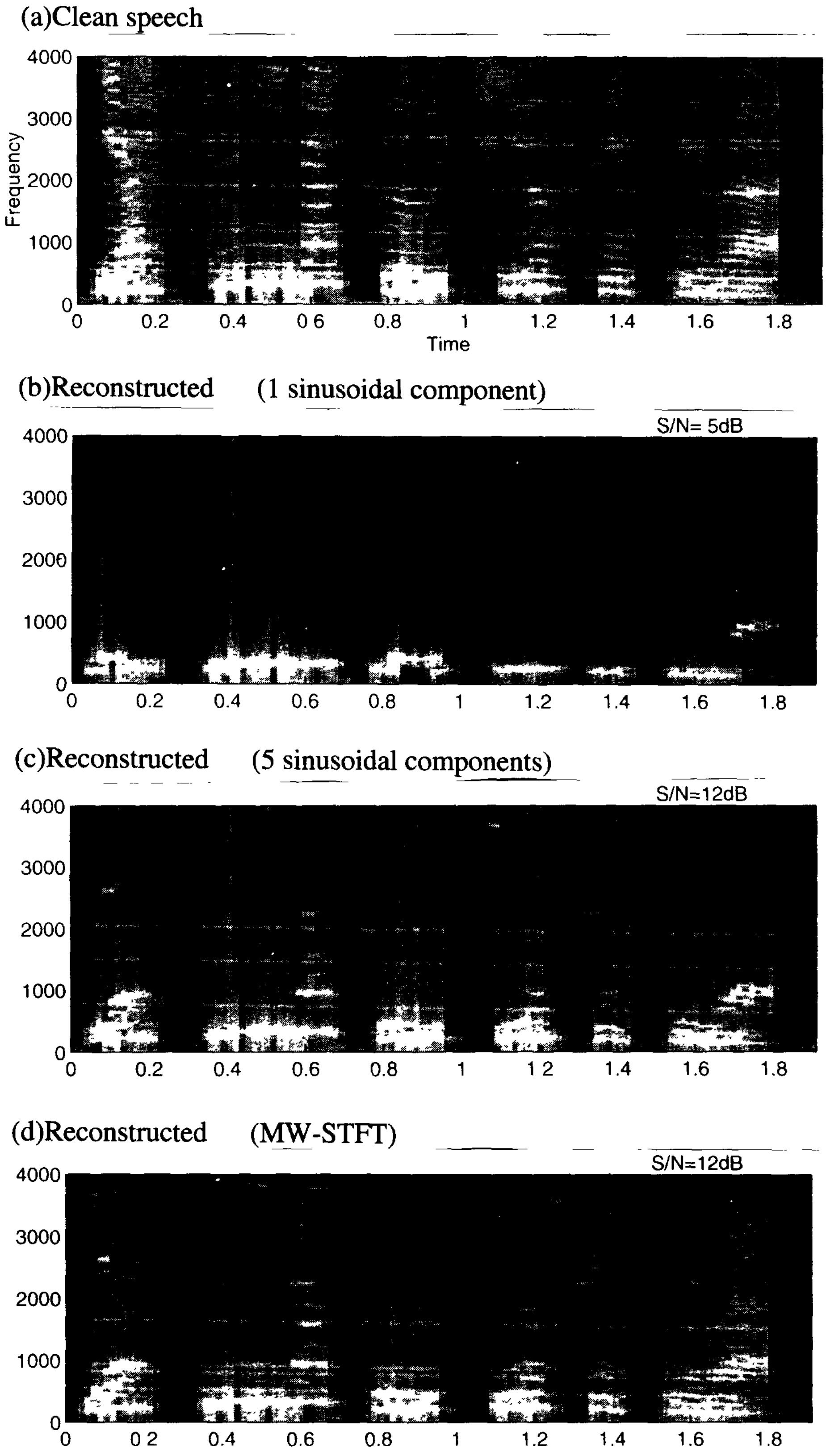
(a)Clean speech

(b)Reconstructed (1 sinusoidal component)

(c)Reconstructed (5 sinusoidal components)

(d)Reconstructed (MW-STFT)

Fig. 3 The time waveforms of clean speech signals



(a)Clean speech

(b)Reconstructed (1 sinusoidal component)

(c)Reconstructed (5 sinusoidal components)

(d)Reconstructed (MW-STFT)

Fig. 4 The spectrograms of clean speech signals

Five major sinusoidal components were found to be necessary for expressing intelligible speech, including the consonants (Fig. 3(c)). Figures 4(a)-(c) show the spectrograms for the waveforms shown in Figs. 3(a)-(c). We can clearly see the difference between Figs. 4(a) and (c), particularly in the middle and high frequency bands, however, both do not yet sound natural.

3. MW-STFT(5)

MW-STFT has been proposed as a method for describing a speech signal by a few sinusoidal components as shown in Fig. 5. Applying the MW-STFT to the speech signal analysis, we can estimate the fundamental frequency of a speech waveform in a super resolution, even when using a short time window(5). High quality speech signals can be reconstructed by using the estimated fundamental frequency and harmonics. We tried to reconstruct the speech waveform using both the spectrum peak-picking and MW-STFT methods to get high quality vowel and consonant reconstructions.

We estimated a fundamental frequency from the largest set of five sinusoidal components in the different window lengths by MW-STFT. We synthesized the speech signal using the five major sinusoidal components and higher harmonics estimated from the fundamental frequency under 8 kHz. Figure 3(d) shows a reconstructed waveform and Fig. 4(d) shows its spectrogram. We can see that the middle and high frequency components are reconstructed well by comparing Figs. 4(c) and (d). We were also able to hear the improvement in speech quality.

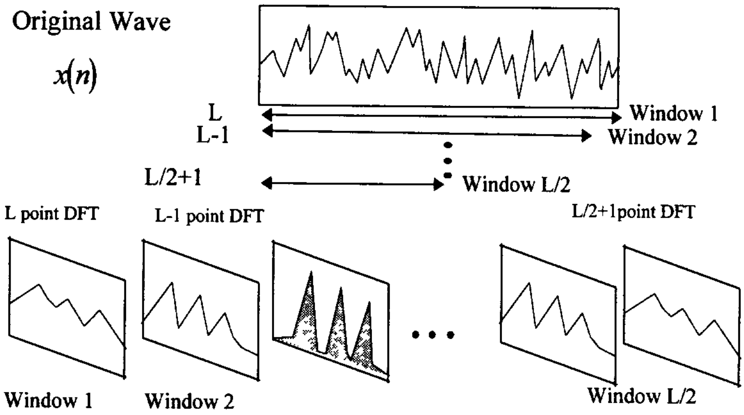
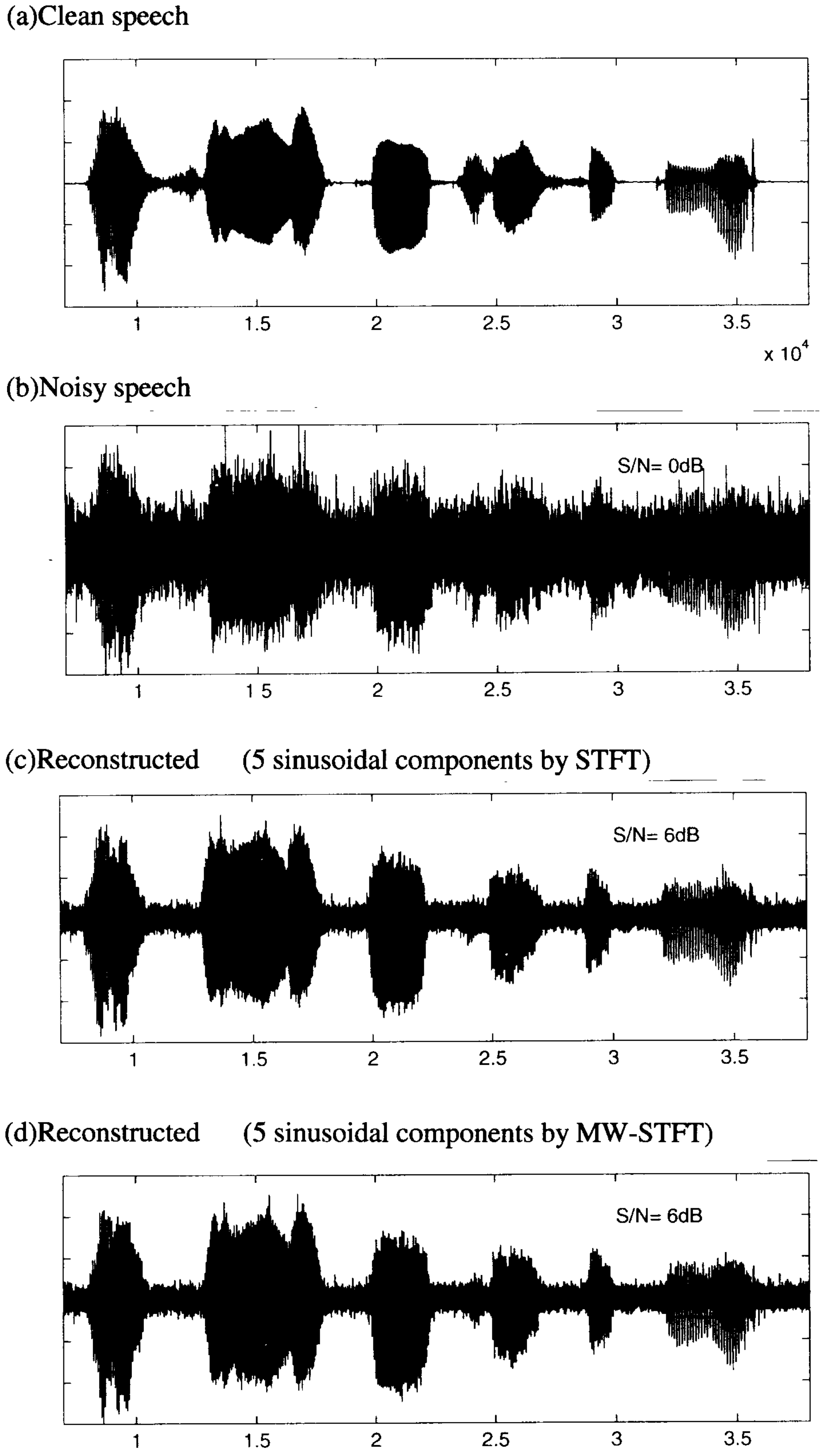


Fig. 5 STFT using multi-rectangular windows(MW-STFT)

4. Speech enhancement from a noisy signal

If we can express a speech waveform by a few sinusoidal components, noise can be reduced by extracting a few major sinusoidal components in noisy speech. Figure 6(a) shows a clean speech waveform while Fig. 6(b) illustrates the original noisy speech. Figure 6(c) demonstrates a noise reduction result by the spectrum peak-picking. The noise added to



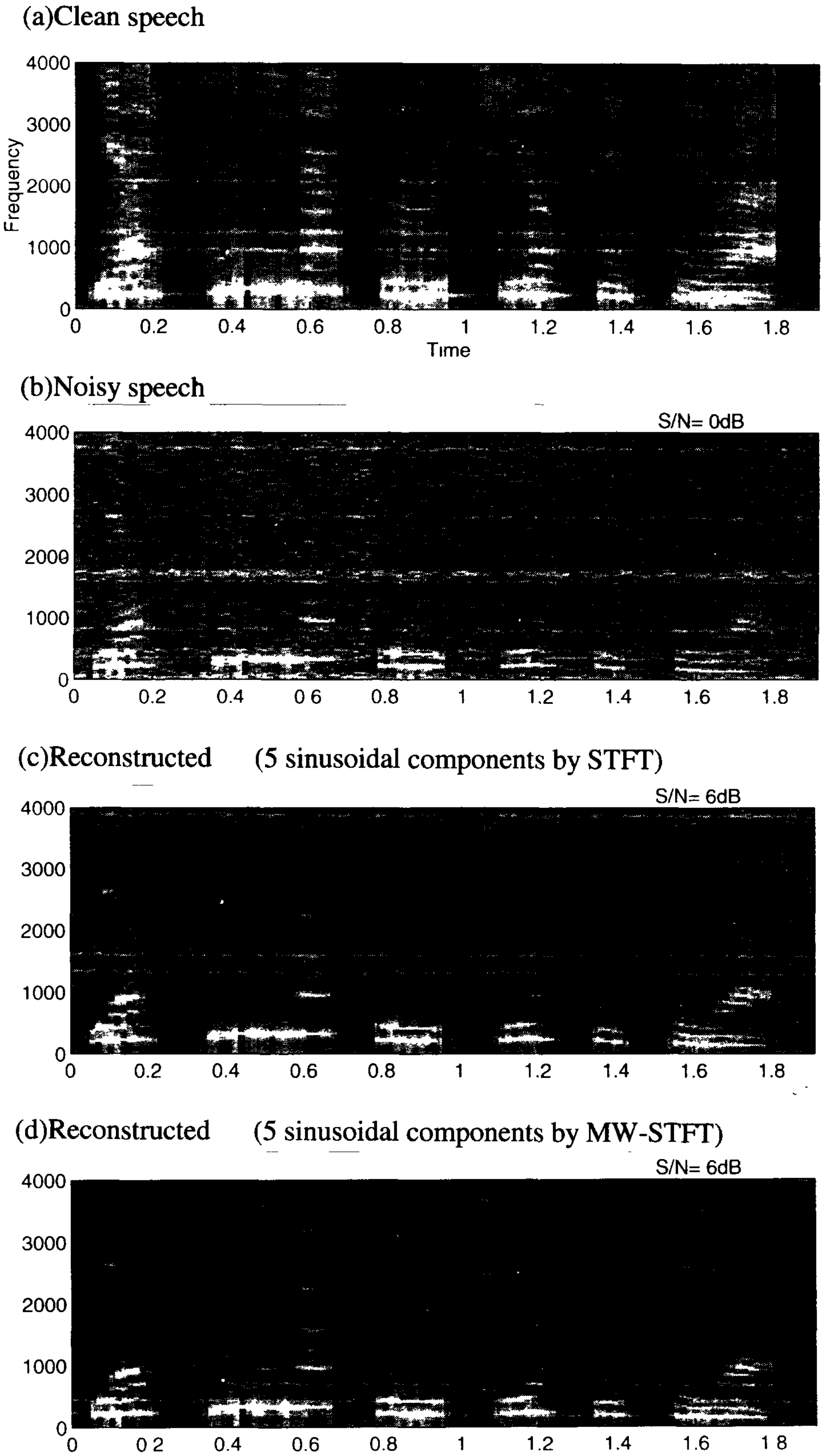
(a)Clean speech

(b)Noisy speech

(c)Reconstructed (5 sinusoidal components by STFT)

(d)Reconstructed (5 sinusoidal components by MW-STFT)

Fig.6 The time waveforms of clean and noisy speech signals



(a)Clean speech

(b)Noisy speech

(c)Reconstructed (5 sinusoidal components by STFT)

(d)Reconstructed (5 sinusoidal components by MW-STFT)

Fig.7 The spectrograms of clean and noisy speech signals

the clean speech is white noise and the S/N ratio is 0 dB. Figures 7(a)-(d) show spectrograms obtained from the waveforms shown in Figs. 6(a)-(d).

Figure 8 shows the noise reduction effect in the S/N ratios performed by changing the number of extracted sinusoidal components such that

$$X = 10 \times \log_{10} \sum_k N(k)^2 - 10 \times \log_{10} \sum_k \left(R(k) - So(k)\right)^2 \quad \text{(dB)}$$

Here, So is the original speech, N is the added noise, and R is the synthesized speech. We can see that noise reduction effects can be obtained when five to ten major sinusoidal components are extracted. Figure 6(d) shows the result of using both the spectrum peak-picking and MW-STFT methods when only five sinusoidal components are extracted without using the harmonics. The noise reduction effect was about 5 dB. The synthesis of the speech waveform including the harmonics does not improve the noise reduction effect.

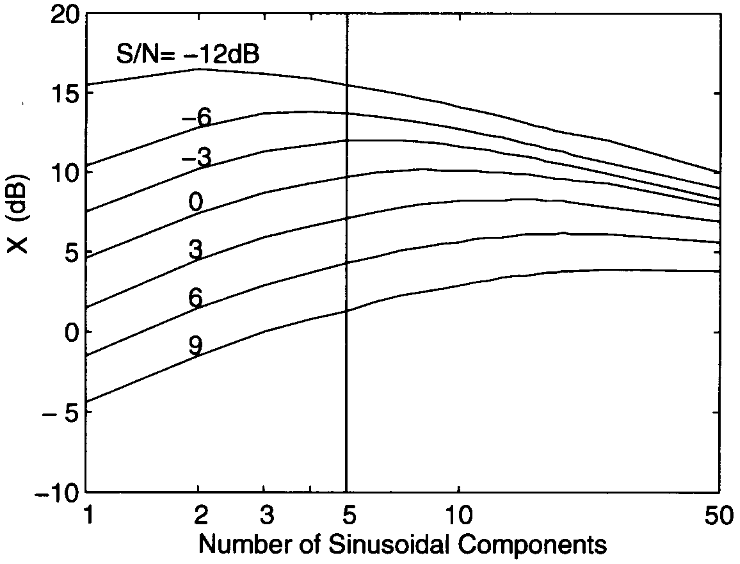


Fig. 8 Noise reduction effect

5. Conclusion

We have proposed a method for reducing noise using power spectrum peak-picking based on the sinusoidal model. Experimental results confirmed that the proposed method works well in reducing noise. The use of five major sinusoidal components were required for reconstructed a understandable speech signals. Noise can be reduced by about 5 dB by extracting these major components. In future work on noise reduction, it will be necessary to estimate or synthesize higher harmonics to improve the speech quality.

6. REFERENCES

(1)K. Ito et al., Technical Report of IEICE, Japan, EA95-59 (in Japanese)(1995)
(2)T.F. Quatieri et al., ASSP 34(6) pp.1449-1464 (1986)
(3)Fukuda et al., IASTED Int. Conf. MSO'96 Paper Code: 242-147 (1996)
(4)T. Houtgast, Private Communication (1996)
(5)T. Ohnishi , et al., ICSV'5,Paper No. 203857 (1997)