Journal of Sound and Vibration (2001) **241**(1), 41–52
doi:10.1006/jsvi.2000.3275, available online at http://www.idealibrary.com on IDEAL®

ESTIMATION OF SPEECH COMPONENTS BY ACF ANALYSIS IN A NOISY ENVIRONMENT

M. KAZAMA

3-4-7, *Shimo-ochiai*, *Shinjuku-ku*, *Tokyo*, 161-0033, *Japan*

AND

M. TOHYAMA

Kogakuin University, 2665-1, *Nakano-machi*, *Hachioji-shi*, *Tokyo*, 192-0015, *Japan*

(*Accepted* 7 *August* 2000)

A speech signal can be decomposed into the fundamental frequency and harmonics, and the autocorrelation function (ACF) is an effective tool for identifying the fundamental frequency and the harmonics. This paper, thus, explains how ACF harmonic analysis can be applied to speech detection and reconstruction when speech communication technologies are used in noisy environments. The dominant sinusoidal components used for the ACF analysis can be picked out from the short-time Fourier spectrum records of a noisy speech signal by using a peak-picking method. Because the number of components usable for speech reconstruction depends on the signal-to-noise (S/N) ratio, we authors developed new methods for peak-picking method and for harmonic sieving. The number of components picked out is adjusted frame by frame depending on the short-time S/N ratio, and harmonics are extracted from the short-time Fourier spectrum record by changing the frame length adaptively according to the fundamental frequency. Consequently, intelligible speech without "musical noise" could be reconstructed from noisy speech signals.

1. INTRODUCTION

The reconstruction of speech from noisy signals is a fundamental issue in research on speech communication technologies that are used in noisy environments, as are hearing aids and speech recognition systems. For speech recognition from a noisy signal, it is important to identify the environmental noise and then to detect a speech signal. The environmental noise in noisy speech signals is generally separated from the speech signals by spectral subtraction [1]: the noise-power spectrum of the 'silent' (noise-only) portions of the signal is identified and then subtracted from the overall signal. This processing can be done in real time by using a short-time Fourier transform (STFT), but it is difficult to determine whether the signals in short-time frames represent silence or speech information. Furthermore, this spectral subtraction introduces a processing noise that is called "musical noise" [2, 3].

The analysis and synthesis of acoustic signals can be based on the sinusoidal model [4], and we have already confirmed that this model enables intelligible speech to be expressed using five high-energy frequency components [5]. We call this approach "peak picking", and an approach of this kind has been used in cochlear implants [6]. When we use a microphone or talk directly to someone in a noisy environment, we usually come closer to the microphone or speak louder than we normally would. We can thus assume that most of

0022-460X/01/110041 + 12 $35.00/0

the peaks in the power spectrum of a noisy speech signal are due to the speech rather than the noise. Consequently, we should be able to synthesize the speech by extracting the high-energy speech components even from the short-time Fourier spectrum of noisy speech. The tonal quality, however, of speech reconstructed from a noisy signal by the peak-picking method is still not good [7].

A vowel can be modelled by the superposition of its fundamental frequency and harmonics, and the autocorrelation function (ACF) is an effective tool to use when looking for a fundamental frequency [8]. Thus, in this paper, we report our investigation of the ways that ACF analysis and a new peak-picking method can be used in the detection and reconstruction of speech. The method we describe here decreases processing distortion (including musical noise) by adjusting the number of frequency components to be extracted from a frame according to the S/N ratio of the signal in that frame. And the fundamental frequencies of vowel sounds are estimated from the ACF of the frequency components picked out of a noisy speech signal. Our speech reconstruction method is suitable for real-time signal processing because it uses a STFT.

This paper is organized as follows. In section 2, we briefly review and summarize our previous work. In section 3, we demonstrate that the fundamental frequency of a vowel can be determined by ACF analysis when only a few sinusoidal components are picked out. In section 4, we describe our new peak-picking and harmonics sieving methods using the ACF and STFT.

2. THE AUTHORS' PREVIOUS WORK

2.1. SPEECH SYNTHESIS BY PEAK PICKING [5]

To find out how many sinusoidal components were needed to represent a speech signal, we used a Japanese female voice sampled every $\frac{1}{16}$ ms. We cut the sample into 32-ms frames, each with 512 data points, and analyzed the power spectrum of the speech sample in every frame by using a STFT (see Figure 1). We used a rectangular-window function to cut the speech data into short frames because we wished to avoid discarding frame signal energy. Discontinuities between successive frames were avoided by having each frame start with the last 256 data points of the previous frame. After analyzing the power spectrum of a frame, we used the peak-picking method to extract the most significant sinusoidal components. The frame signal was then reconstructed from the extracted components. In this reconstruction a triangular window was used in order to smooth out any discontinuity [9].

The original and two reconstructed speech waveforms are shown in Figure 2. One can see that the envelope of the waveform reconstructed from only the most significant sinusoidal components is almost the same as that of the original waveform. This reconstructed speech was generally but not completely intelligible, particularly because the consonants were not clear. Five major sinusoidal components were found to be necessary for expressing intelligible speech including consonant sounds (see Figure 2(c)).

2.2. NOISE REDUCTION BY PEAK PICKING [7]

To investigate the reconstruction of speech from noisy signals, we added white noise to clean speech so that the S/N ratio was 0 dB. If the energy of the signal in a frame was greater than a specified threshold value, we assumed that the frame was a speech frame. The frequency components in it were analyzed by using a STFT, and the dominant components were extracted by a peak-picking method. From these components the waveform

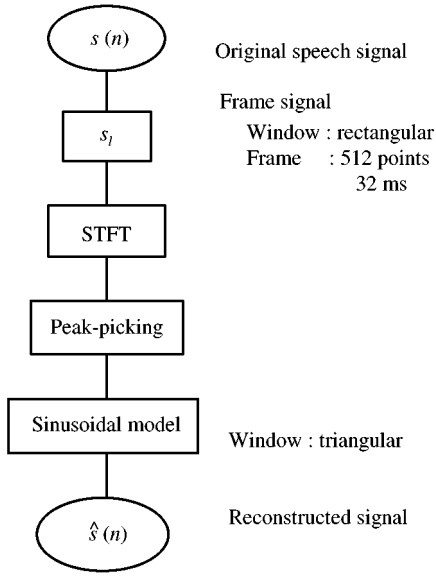

Figure 1. Peak-picking method for speech reconstruction.

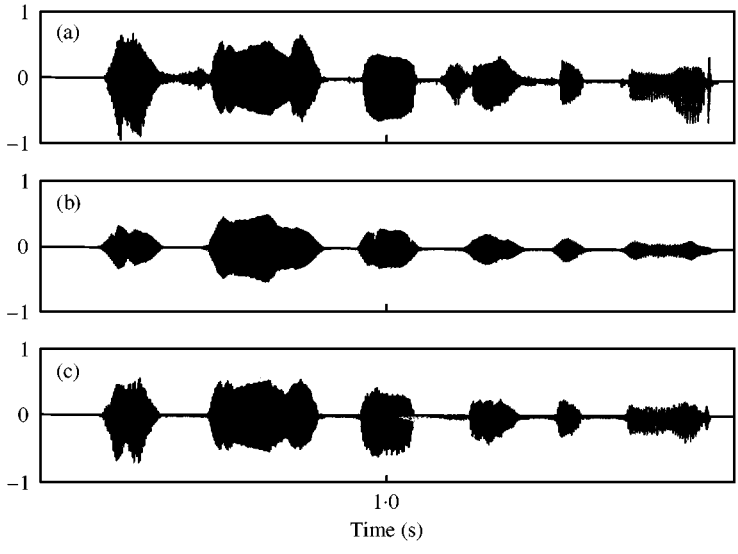

Figure 2. Original and reconstructed waveforms: (a) original; (b) reconstructed from one sinusoidal component; and (c) reconstructed from five sinusoidal components.

corresponding to that frame was then reconstructed according to the sinusoidal model [4]. If the energy was less than the threshold value, the frame was assumed to be a silent portion of the speech signal.

Figure 3(a) and 3(b) show clean and noisy speech waveforms, and Figure 3(c) shows the speech waveform reconstructed from the five frequency components extracted. The residual noise after the reconstructed speech was extracted is shown in Figure 3(d). The reconstructed waveform is clearly almost the same as the original one. The noise reduction

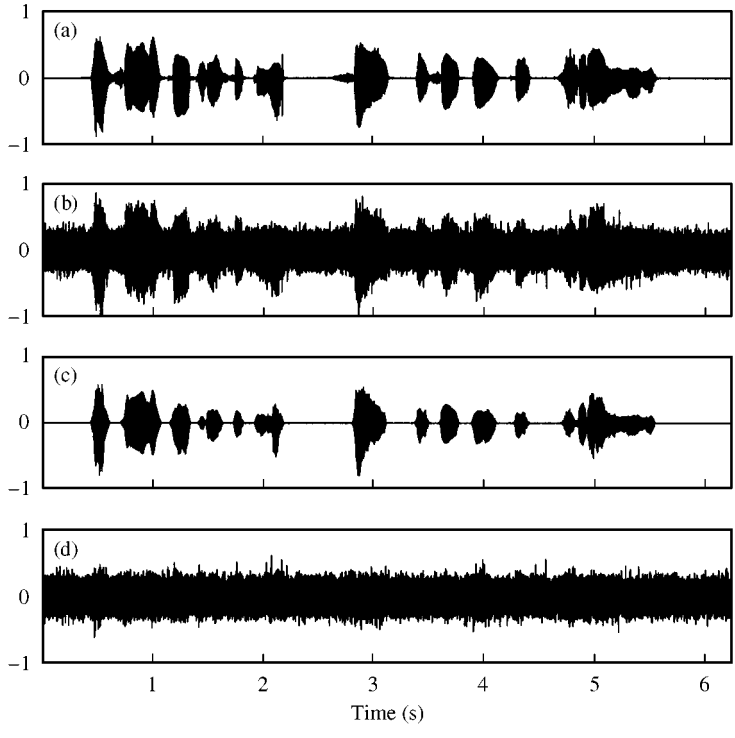

Figure 3. Original and reconstructed waveforms: (a) original; (b) original plus noise (S/N 0 dB); (c) reconstructed; and (d) residual noise.

effect can be estimated by calculating a signal-to-deviation ratio (*SDR*) defined by

$$SDR = 10 \times \log \left(\frac{\sum_n s(n)^2}{\sum_n (\hat{s}(n) - s(n))^2} \right) (\text{dB}),$$

where $s(n)$ is the original speech signal and $\hat{s}(n)$ is the reconstructed speech signal. The peak-picking method increased the *SDR* by about 10 dB when the S/N ratio in the noisy speech signal was 0 dB. The reconstructed speech did not sound good, however, because processing distortion (musical noise) had been introduced.

In the experiment described above the number of extracted components used in the reconstruction was fixed at five for all the frames, but the number can also be changed frame by frame. Figure 4 shows how the *SDR* changes with changes in the number of frequency components extracted and used for reconstruction. We found that the number of components should be increased when the S/N ratio of the signal is high and should be decreased when it is low. In our experiments, the effects of noise were suppressed best when 5–10 sinusoidal components were extracted and used for reconstruction.

3. SPEECH REPRESENTATION BASED ON ACF ANALYSIS

The expression of a vowel can be based on the harmonic structure estimated from ACF analysis [8]. So we conducted a series of experiments investigating the feasibility of using ACF analysis to estimate the fundamental frequency from only a few dominant frequency components and then using that estimate for sieving the harmonic components.

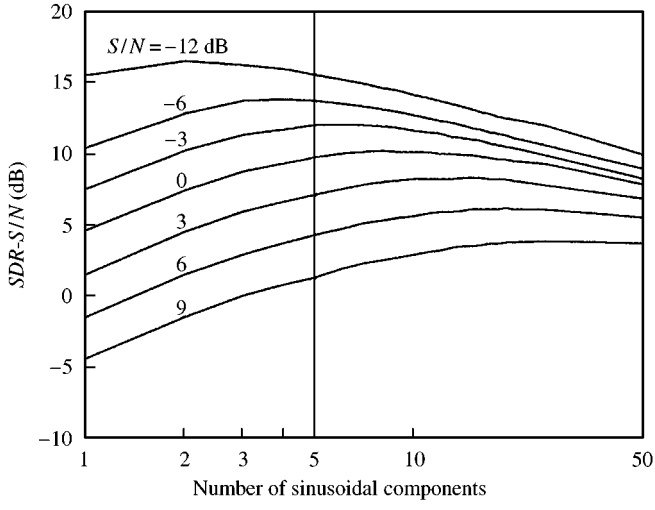

Figure 4. The noise reduction effect.

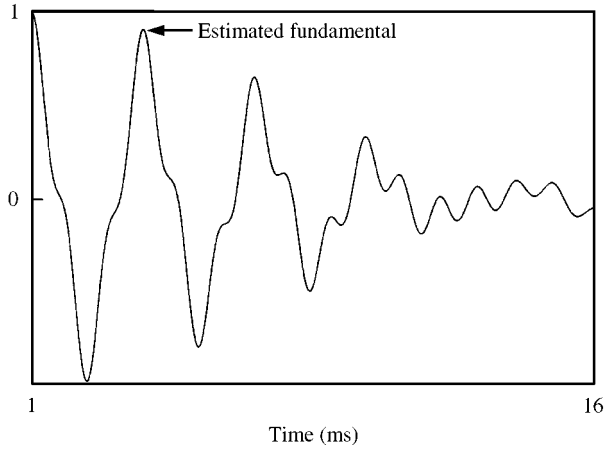

Figure 5. ACF example.

3.1. SPEECH RECONSTRUCTION PROCEDURE

The speech sample we used was the same one used in the peak-picking experiments described in section 2.1. When the frequencies of three dominant components picked out were under 1 kHz, the frame was assumed to be a vowel frame. We estimated its fundamental frequency by using the cyclic ACF of the components. Assuming the waveform of the components extracted to be periodic, the ACF was obtained as the inverse Fourier transform of the power spectrum record. To increase the precision with which the fundamental frequency was estimated, we calculated the ACF by appending zeros to the power spectrum record until the record length became 20 times the original length. Figure 5 shows an example of the ACF which estimates the vowel frame fundamental frequency.

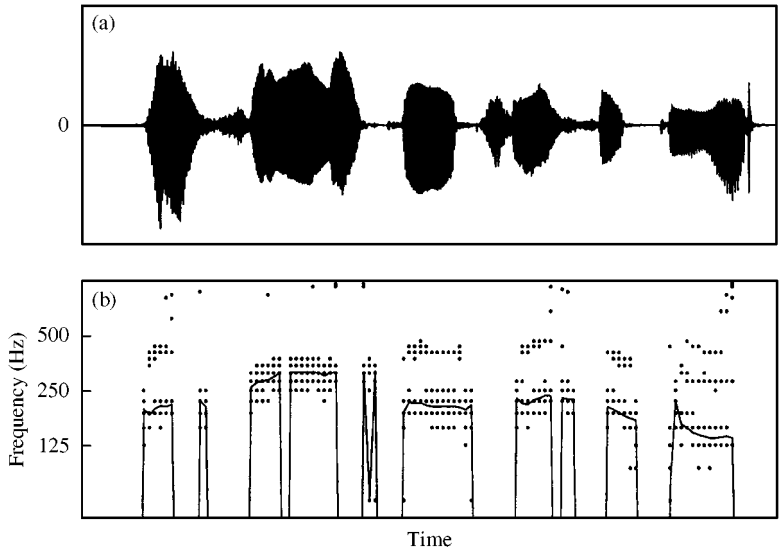

Figure 6. (a) Speech signal; (b) frequencies extracted from vowel frames; and the corresponding fundamental frequencies estimated.

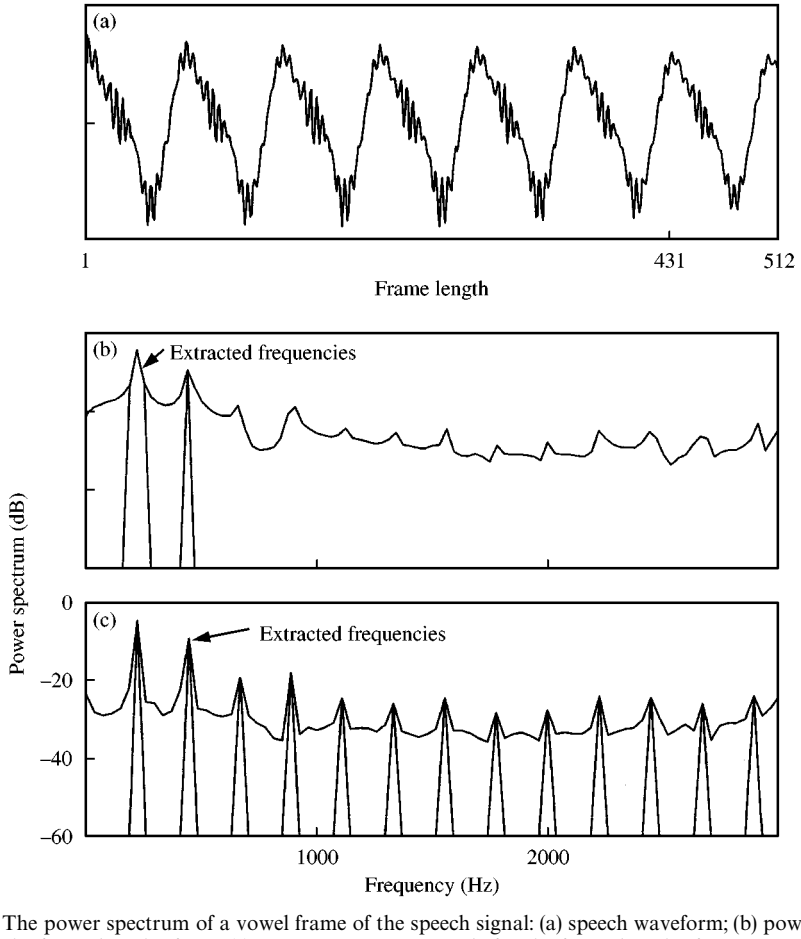

Figure 7. The power spectrum of a vowel frame of the speech signal: (a) speech waveform; (b) power-spectrum example for the frame length of 512; (c) power spectrum example for the frame length of 431.

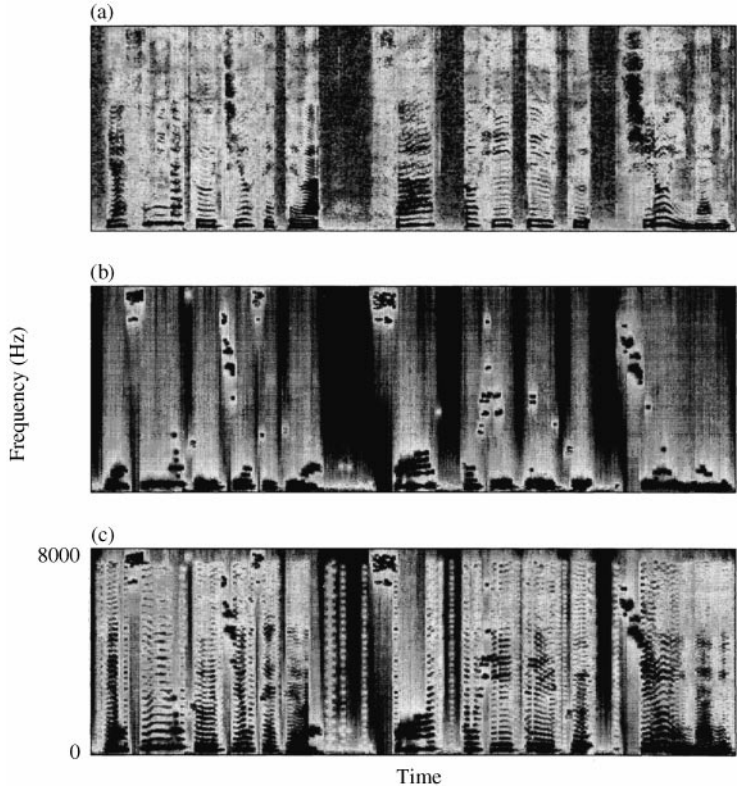

Figure 8. (a) original and (b,c) reconstructed speech signal spectrograms; (b) reconstructed using five components; and (c) reconstructed from fundamental and harmonic frequencies.

The frame length used in sieving the harmonic components by STFT (DFT) analysis was adaptively selected between 257 and 512 points at every frame after estimating the fundamental frequency for that frame [10]. Vowel frames were thus reconstructed using the fundamental and the harmonic frequencies extracted by sieving. Frames were assumed to be consonant frames when the frequencies of two dominant components picked out were higher than 1 kHz. The waveforms in those frames were reconstructed using only the components extracted by peak picking. As in our earlier work, a triangular window was used in reconstructing the speech signal in order to smooth out any discontinuities [9].

3.2. RECONSTRUCTED SPEECH

An original speech waveform, the five frequencies extracted from frames that were assumed to be vowel frames, and the corresponding fundamental frequencies estimated are shown in Figure 6. The estimated fundamental frequencies were distributed between 150 and 350 Hz. The five major sinusoidal components extracted from the vowel frame of the speech waveform shown in Figure 7(a) were in a low-frequency band and did not seem to reflect the original formants clearly. The extracted frequencies as shown in Figure 7(b) are mostly distributed around the spectrum peaks because the analysis frame length is not adjusted to the fundamental frequency. Figure 7(c) shows the power spectrum of Figure 7(a) and also shows the fundamental and harmonic frequencies extracted. The time frame length

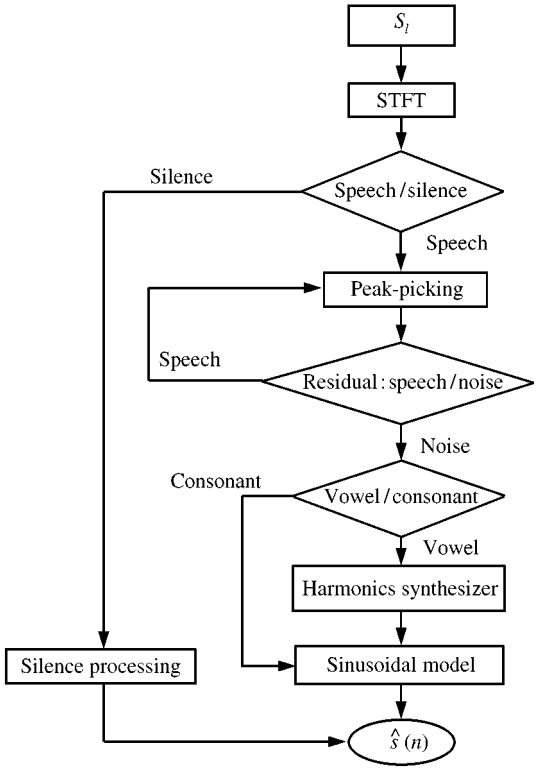

Figure 9. Variable peak-picking method for noise reduction.

for STFT power spectrum analysis, 431 data points, was selected to be equal to the period of the frame fundamental frequency. As shown in Figure 7(c), the fundamental frequency can be estimated accurately from the five extracted frequency components by using the peak-picking method and ACF analysis.

The original and reconstructed speech spectrograms are shown in Figure 8. The mid-frequency energy in Figure 8(c) is more intensive than that in Figure 8(b), and the improvement in the speech quality was confirmed by informal listening tests. Consequently, we have confirmed that the peak-picking method and ACF analysis can be used to identify the fundamental frequency of a vowel frame and that the harmonics can be identified in a sieving procedure using frame lengths adjusted according to the fundamental frequencies.

4. REDUCING NOISE BY VARIABLE PEAK PICKING AND HARMONIC-SIEVING

4.1. VARIABLE PEAK PICKING

The variable peak-picking method using ACF analysis to reduce noise is shown in Figure 9. The noisy speech used in the present experiments was the same as that described in section 2.2. If the energy of the signal in a frame was greater than a specified threshold value, the frame was assumed to be a speech frame, and we continued picking out peaks until the energy of the residual frequency components had a distribution like that of the Gaussian noise. The distribution of the power spectrum given by the squared sum of the real

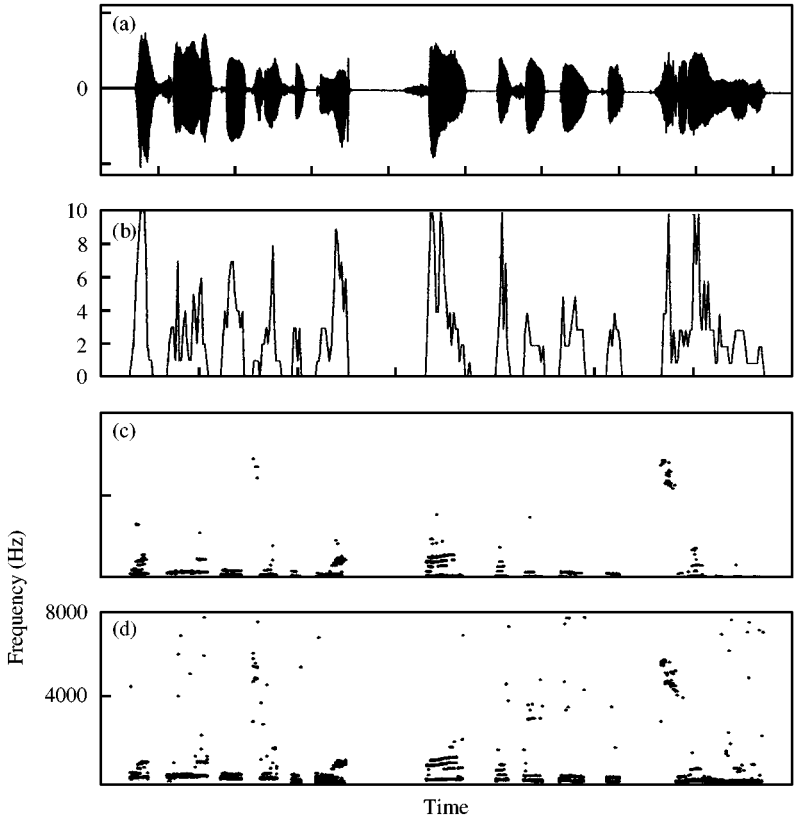

Figure 10. Frequency components distributions: (a) clean speech signal; (b) number of extracted components; (c) distribution of extracted components; and (d) distribution when five components were extracted.

and imaginary parts of a Gaussian noise spectrum is exponential. Thus, we continued peak-picking until the ratio of the residual components with energy higher than the average energy to all the residuals became 0·3, and the maximum number of extracted components was limited to 10. We could therefore expect that musical noise would not be produced during the reconstruction because a nearly optimum number of components was extracted from every frame without also extracting high-energy noise components. The vowel and consonant frames were discriminated as described in section 3.1, and the reconstructed speech signal of a vowel frame was synthesized from the fundamental frequency and the harmonics by using the method described in section 3.1. The squared magnitudes of the harmonics, however, were weighted by -6 dB/oct.

4.2. SPEECH RECONSTRUCTION RESULT

Figure 10(a) illustrates a clean speech signal, and Figure 10(b) shows the distribution of the number of frequency components extracted. If one compare Figures 10(a) and 10(b), one can see that the number of frequency components extracted decreases as the amplitude of the speech signal decreases. The distribution of the frequency components extracted by variable peak picking is shown in Figure 10(c), and the distribution obtained when five components were always extracted is shown in Figure 10(d). The frequency components in

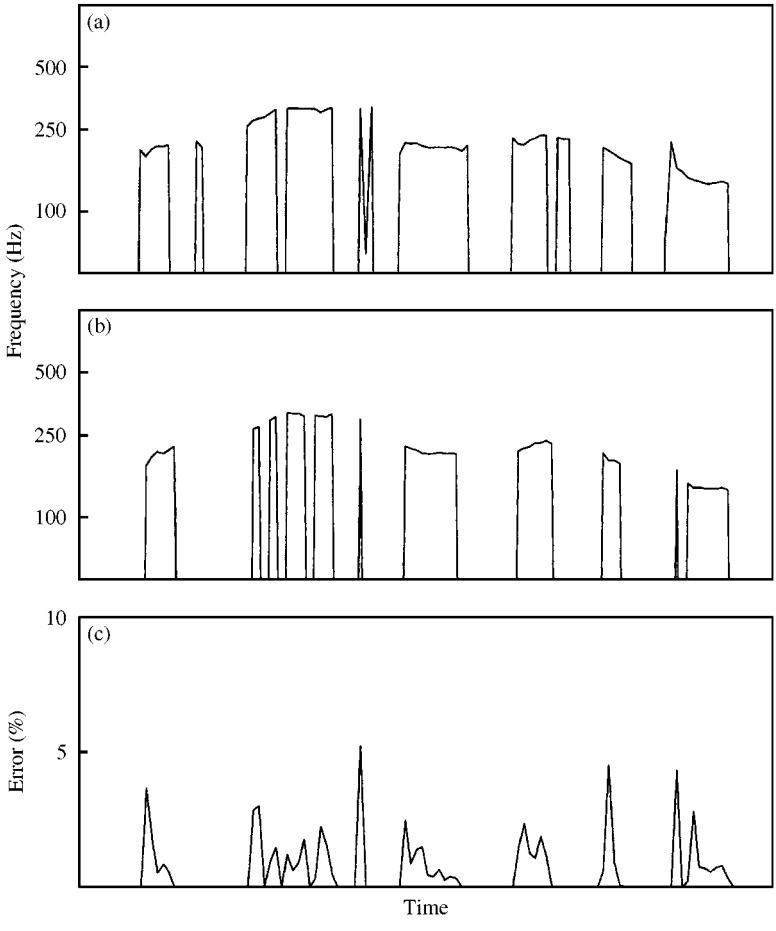

Figure 11. Fundamental frequencies estimated from clean (F_c) and noisy (F_N) samples: (a) F_c; (b) F_N; (c) errors when estimated from noisy speech: $\{((F_c) - (F_N))/(F_c)\} \times 100$; $(F_c) \neq 0$, $(F_N) \neq 0$.

Figure 10(c) are less scattered than those in Figure 10(d). We confirmed by listening that the scattered frequency components are one source of musical noise. Figure 11 shows the fundamental frequencies estimated from clean and noisy speech samples. Figure 11(c) shows the estimation errors derived from the difference between Figure 11(a) and 11(b) when the fundamental frequencies of the clean and noisy speech signals could both be estimated. The fundamental frequency could be estimated from a noisy sample to be within 5% whenever the vowel frames could be identified. Some vowels, however, were hidden by noise and not identified.

Samples of energy time-envelopes for the frequency components of noisy, reconstructed, and clean speech signals are shown in Figure 12. For each frequency component we can see the energy time-envelope recovered from noisy speech is almost the same as the energy time-envelope of clean speech.

We confirmed by listening that the speech reconstructed in this experiment sounded better than that reconstructed in the experiment described in section 2.2 (i.e., reconstructed from five sinusoidal components only) because it contained less musical noise. The noise reduction in the two experiments, however, was similar: about 10 dB.

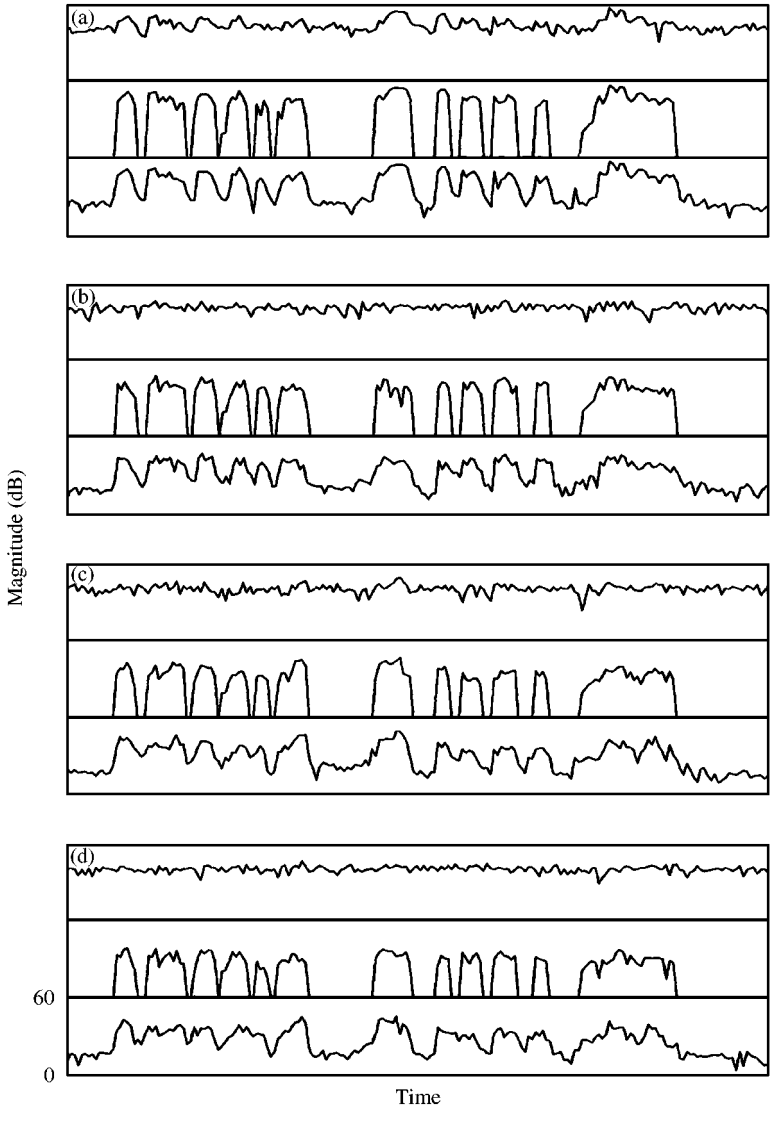

Figure 12. Energy-time envelopes for frequency components of noisy speech (top traces), reconstructed speech (middle traces), and clean speech (bottom traces): (a) 250 Hz; (b) 500 Hz; (c) 1000 Hz; and (d) 2000 Hz.

5. CONCLUSION

The method described here is based on picking peaks out of the power spectrum and on harmonic analysis using the ACF. The fundamental frequency of even a noisy speech signal can be identified by using a few major sinusoidal components extracted by our variable peak-picking method. And the harmonics can be extracted from a noisy signal by using a STFT-based sieving procedure in which the frame length for STFT processing is adjusted according to the fundamental frequency of the signal in that frame. Experimental results confirmed that this method reduces the noise level by 10 dB without introducing processing distortion. After developing our proposed method for real-time-based processing, we shall do a hearing test towards the speech enhancement in a hearing aid. It might be reasonable

to do intelligibility tests with normal hearing subjects first, since informal quality judgments and simple error measures cannot replace intelligibility tests.

ACKNOWLEDGMENT

The authors wish to thank A. Morita for his cooperation and fruitful discussions. This research project is partly supported by NHK HOSO BUNKA FUND.

REFERENCES

1. S. F. Boll 1979 *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **ASSP-27**, 113–120. Suppression of acoustic noise in speech using spectral subtraction.
2. Z. Goh, K. C. Tan and B. T. G. Tan 1998 *IEEE Transactions on Speech and Audio Processing*, **6**, 287–292. Postprocessing method for suppressing musical noise generated by spectral subtraction.
3. Y. Ephraim and H. L. V. Trees 1995 *IEEE Transactions on Speech and Audio Processing*, **3**, 251–266. A signal subspace approach for speech enhancement.
4. T. F. Quatieri and R. J. Mcaulay 1986, *IEEE Transactions on Acoustics, Speech, and Signal Processing* **ASSP-34**, 1449–1464. Speech transformations based on a sinusoidal representation.
5. M. Kazama, M. Tohyama and T. Ohnishi 1997 *The Fifth International Congress on Sound and Vibration*, 2079–2086. Speech signal enhancement based on a sinusoidal model.
6. R. V. Shannon, F. G. Zeng, V. Kamath, J. Wygonski, M. Ekelid 1995 *Science*, **270**, 303–304. Speech recognition with primarily temporal cues.
7. M. Kazama, M. Tohyama and A. Morita, 1999 *137th Meeting of the Acoustical Society of America and the 2nd Convention of the European Acoustics Association*, 2pSPa9. Speech reconstruction from a received noisy signal.
8. Y. Ando, S. Sato and H. Sakai 1999 in *Computational Acoustics in Architecture* (J. J. Sendra, editor), 63–97 *Southampton, Boston, WIT Press*, Fundamental subjective attributes of fields based on the model of auditory-brain system, Chapter 4.
9. J. Laroche and M. Dolson 1999 *Journal of the Audio Engineering Society* **47**, 928–936. New phase-vocoder techniques for real-time pitch shifting, chorusing, harmonizing, and other exotic audio modifications.
10. T. Ohnishi, M. Kazama and M. Tohyama 1997 *The Fifth International Congress on Sound and Vibration*, 2167–2174. Acoustic signal processing using multi-windowed STFT and harmonics sieving.