STUDIA PHONOLOGICA XVI (1982)

Real Time Noise Canceling by Bandpass Filter

Yasuo ARIKI and Toshiyuki SAKAI

SUMMARY

In this paper we describe a hardware system based on a channel vocoder for reducing, in real time, noises superimposed on speech. Furthermore, we describe two algorithms to be performed on the hardware system for reducing classical random noises and acoustic pulsive noises with resonances and long duration such as a telephone bell. Noises in environments degrade the quality of speech so that reduction of noises from noisy speech is essentially required for speech communication by telephone line or for speech recognition by a computer. From the view point of speech communication, we developed the noise reduction hardware system and two algorithms which are performed, in real time, for several kinds of noises including random noises.

1. INTRODUCTION

In a case where speech alone is transmitted, processed or recognized, the main theme is to analyze observed speech waveform based on the speech model. In natural environments, however, speech waveform alone is hardly observed due to noises in the environments. It is not enough, therefore, to analyze speech waveform based only on the speech model in actual speech communication, processing or recognition. Adaptive noise reduction is required which separates speech from noise and recovers or enhances the speech, regarding both a speaker and his environments as information source of speech patterns. At present state of arts, however, such an adaptive reduction for various kinds of environmental noises is difficult to be implemented. Therefore, many conventional studies restrict the type of environmental noises and attempt to reduce them. Our attention is payed to this adaptive reduction of unspecific noises and we developed a real time noise reduction system. In this paper, we describe our system and its algorithms.

Two actual purposes may be considered for reducing noises and enhancing the speech. One is to increase the recognition rate of a speech recognition system. It is generally known that speech recognition rate decreases when environmental noises are superimposed on speech. Therefore, robust recognition algorithms and noise reduction algorithms as preprocessing are required. Another purpose is to

Yasuo ARIKI (有木康雄) : Assistant Professor, Department of Information Science, Kyoto University.
Toshiyuki SAKAI (坂井利之) : Professor, Department of Information Science, Kyoto University.

enhance the speech for communication. Example is the speech degraded by environmental noises through telephone line or a hearing aid. Another example is a playback of the speech recorded under noisy environments. For easier listening, it is desired to reduce the noises and to improve the intelligibility, that is, enhance the speech. In this paper, we describe the speech enhancement by reducing environmental noises from the latter view point.

2. RELATED STUDIES

The type of noises in the environments may be considered in two aspects. One aspect is their own properties. The other is the correlation of noises with the speech. Concerning with their own properties, many types of noises exist. The most typical one is random noise such as white noise. Others are pulsive noises such as hum, and noises with resonances such as sounds of a bell. On the other hand, for the correlation with speech, there exist uncorrelated noises (additive) and correlated noises with speech. Here, we breifly survey noise reduction systems which deal with additive and random noises.[1)]

Noise reduction systems may be classified mainly into four types according to what kind of parameter domain is used for analysis and what kind of difference of property is employed for processing between speech and noises. The four groups are frequency domain method, time domain method as the method based on parameter domain, and methods based on speech periodicity and speech model as the method based on properties to be employed.

2.1 Method based on Short-time Spectral Amplitude

2.1.1 Spectral Subtraction

It is generally known that short-time spectrum is more important than the phase for intelligibility and quality of speech, because the auditory system is perceptually not so sensitive to the phase. Therefore, the speech waveform can be produced only if the short-time spectrum of the undegraded speech can be estimated.

Now, let $y(t)$ denote the observed noisy speech waveform produced by adding the uncorrelated noise $n(t)$ to the speech waveform $s(t)$. $y(t)$ can be expressed as follows:

$$y(t)=s(t)+n(t) \tag{1}$$

The short-time power spectrum through spectral analysis can be obtained from the expression (1) as follows:

$$|Y(\omega)|^2=|S(\omega)|^2+|N(\omega)|^2+S(\omega)N(\omega)^*+S(\omega)^*N(\omega) \tag{2}$$

where $Y(\omega)$, $S(\omega)$ and $N(\omega)$ denote the short-time spectra of $y(t)$, $s(t)$ and $n(t)$ respectively, and * indicates the complex conjugate. The third and fourth terms in the expression (2) equal zero due to uncorrelation of the noise with the speech so that the expression (2) results in the following expression:

$$|Y(\omega)|^2=|S(\omega)|^2+|N(\omega)|^2 \tag{3}$$

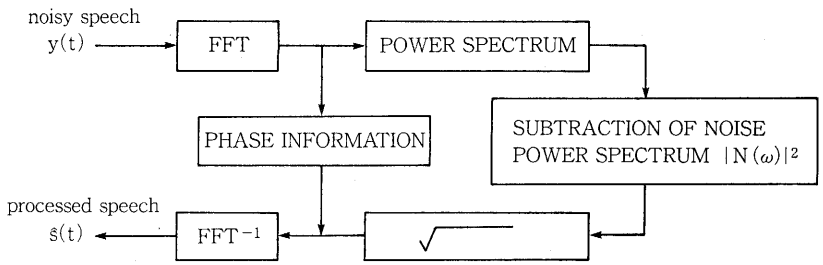

Fig. 1. Spectral subtraction method.

If the short-time power spectrum of the noise is *a priori* known or can be estimated in silence intervals, the short-time power spectrum of the speech can be estimated from the expression (3) as follows:

$$|\hat{S}(\omega)|^2=|Y(\omega)|^2-|N(\omega)|^2 \tag{4}$$

The speech waveform can be produced using the short-time spectral amplitude $|\hat{S}(\omega)|$ from the expression (4) and the phase information of $\hat{S}(\omega)$. As the phase information of $\hat{S}(\omega)$, that of $Y(\omega)$ can be used because of unsensitivity of the phase to auditory system. Fig. 1 shows this spectral subtraction method based on FFT. The expression (4) can be expanded as follows:

$$|S(\omega)|^a=|Y(\omega)|^a-k|N(\omega)|^a \quad (a, k \text{ are constant}) \tag{5}$$

From the above expression, various kinds of systems can be achieved by deciding a and k.[2)]

2.1.2 Wiener Filter

N. Wiener considered the linear filter to extract signal from noisy signal and devised the optimum filter based on minimization of mean-square error.[3)] The constraint to design the optimum linear filter based on minimization of mean-square error is that the noise n(t) and the signal s(t) is stationary with time and statistically independent each other. Under this constraint, the optimum filter $H(\omega)$ can be obtained as follows by denoting $Pn(\omega)$, $Ps(\omega)$ as the power density spectra of the noise and signal respectively.

$$H(\omega)=\frac{Ps(\omega)}{Ps(\omega)+Pn(\omega)} \tag{6}$$

This optimum filter, however, can not be applied directly to the separation of the noise and speech due to the unstationarity of the speech. To apply the optimum filter to the separation of noise and speech, apploximated filter is proposed which uses short-time power spectrum in stead of power density spectrum.[4)] The apploximated filter results in the adaptive Wiener Filter with dynamic characteristics as the following expression by using the denotation in the expression (4)

$$H(\omega)=\frac{|S(\omega)|^2}{|S(\omega)|^2+|N(\omega)|^2}=\frac{|S(\omega)|^2}{|Y(\omega)|^2} \tag{7}$$

The problem in the adaptive Wiener Filter lies in the estimation of the short-time

power spectrum of the speech. An approach often used is to estimate it by the expression (4) described in spectral subtraction. From the expression (7), adaptive Wiener Filter has zero phase so that the short-time spectrum of the speech $\hat{S}(\omega)$ is estimated using the phase $Y(\omega)$ as follows.

$$\hat{S}(\omega)=H(\omega)Y(\omega) \tag{8}$$

Another types of adaptive filters have been studied which are not based on minimization of mean-square error on a short-time basis.[5]

2.2 Method based on Autocorrelation[6]

The autocorrelation function of speech can be perceived as well as speech waveform due to the fact that the autocorrelation function is associated with power spectrum and auditory system is unsensitive to the phase change. Let $\rho(\tau)$ denote the autocorrelation function of the noisy speech. If the noise is uncorrelated with the speech, $\rho(\tau)$ can be expressed as follows:

$$\rho(\tau)=\rho_s(\tau)+\rho_n(\tau) \tag{9}$$

where $\rho_s(\tau)$ and $\rho_n(\tau)$ are the autocorrelation functions of the speech and noise waveforms respectively. If the noise is random, the value of $\rho_n(\tau)$ dominates around the origin ($\tau=0$) so that the noise can be reduced by taking out the interval which corresponds to the pitch period from $\rho(\tau)$ as the output. This method is shown in Fig. 2. The short-time autocorrelation function $\rho_1(\tau)$ is computed from noisy speech waveform y(t) at starting time t1. After that, the pitch period T1 is computed. Next, output signal g(t) is produced from $\rho_1(\tau)$ as the interval with pitch period excepting origin. Similarly, the short-time autocorrelation function $\rho_2(\tau)$ and pitch period T2 are computed from noisy speech waveform y(t) at starting time t2=t1+T1. The output is produced from $\rho_2(\tau)$ as the interval with pitch period T2 excepting origin. These steps are repeated and the output of each step is spliced.

2.3 Method based on Periodicity of Speech[7]

Waveforms of voiced sounds are periodic with a period which corresponds to pitch frequency as shown in Fig. 3(a). In the frequency domain, this property can be seen as the harmonics as shown in Fig. 3(b). Therefore, a comb filter as shown in Fig. 3(c) can pass the harmonics of voiced sounds and consequently reduce the noises between harmonics. The problems of this comb filtering lie in the accuracy of pitch extraction and noise reduction in the unvoiced sound. For these problems, many techniques are developed such as multiple pitch extractors or noise removal in the unvoiced interval by using the noise spectrum estimated in the voiced interval.

2.4 Method based on Speech Model

If the system parameters of a speech model can be directly estimated from noisy speech, the high performance speech recognition or the high quality speech communication can be realized. When we assume an all-pole model for the vocal tract, the speech waveform s(n) can be represented as follows:

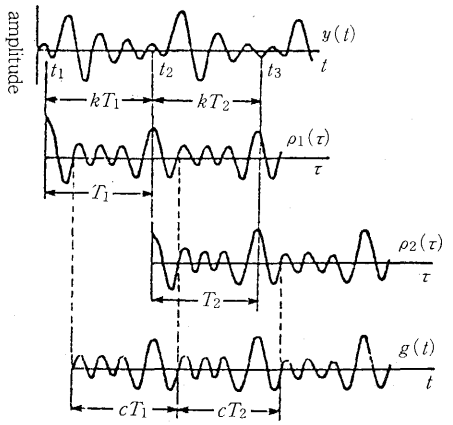

Fig. 2. Method of splicing of short-time autocorrelation function.

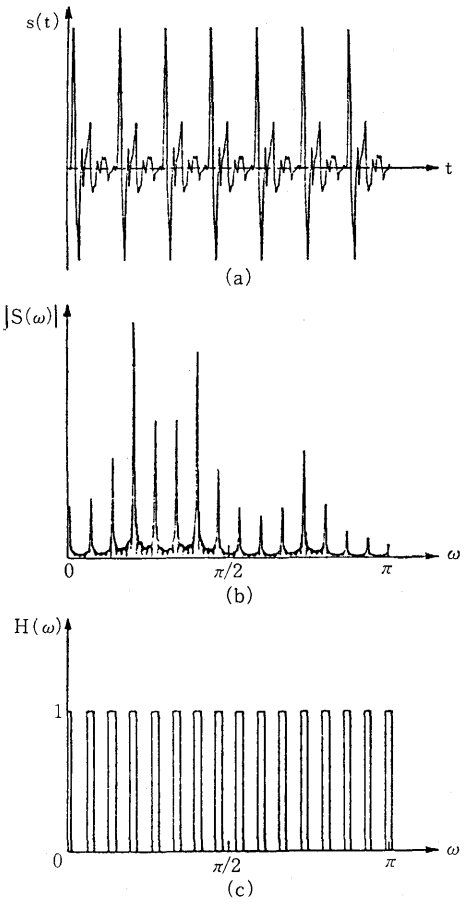

Fig. 3 Method of comb filtering

(a) Periodic time waveform.
(b) Magnitude spectrum of the time waveform.
(c) Comb filter.

$$s(n)=\sum_{i=1}^{p} a_i \cdot s(n-i)+u(n) \tag{10}$$

where u(n) is the excitation source and "P" represents the order of the all-pole model. On the other hand, noisy speech y(n) is represented in terms of the same system parameters a_i as follows:

$$y(n)=\sum_{i=1}^{p} a_i \cdot s(n-i)+u(n)+n(n) \tag{11}$$

If the system parameters a_i can be estimated from the noisy speech waveform y(n), the undegraded speech waveform s(n) can be synthesized from the expression (10). The system parameters a_i are generally estimated by linear predictive analysis after subtracting noises in frequency domain or autocorrelation function.

3. Hardware Organization for Real Time Noise Canceling

3.1 Specification of Hardware

For speech recognition and speech communication by telephone or hearing aid, real time reduction of environmental noises is required. In addition to that, low cost of the system is also required. From this point of view, we decided the system specification for the noise reduction as follows:

(1) Various types of environmental noises are reduced.
(2) Environmental noises are reduced in real time.
(3) Hardware organization is simple.

The methods based on a comb filtering or speech model described in the previous section can not satisfy the real time processing of this specification at the present state of arts. The method based on autocorrelation function seems to have the difficulty in reducing noises other than random noise. One of the possible methods for real time noise reduction by simple hardware is to reduce the noises by analysis and synthesis using bandpass filters in stead of FFT in the method based on the short-time spectral amplitude.

3.2 Hardware Organization

Fig. 4 shows our hardware organization of the noise reduction system by bandpass filters. Noisy speech waveform is passed into 15 channel filter-bank. The center frequencies of the 15 channels used increase in order by a factor $2^{1/3}$. These frequencies are shown in Fig. 4. After they are full-wave-rectified and smoothed, the output waves are sampled at every 16 ms frame interval and digitized with an accuracy of 8 bits. In addition, frame power is also extracted at the same time. The short-time spectral amplitude of undegraded speech is estimated by a micro processor by using the short-time spectral amplitude and frame power of the noisy speech. During this processing, waveform of each channel before rectification is delayed by BBD (Bucket Brigade Device). These delayed waveforms are attenuated by DCAA (Digitally Controlled Audio Attenuator) so that the short-time spectral

amplitude of the delayed waveform is equal to the estimated one. Then, the attenuated waveforms are added to synthesize the speech waveform. In this process, the attenuated waveforms of all channels are passed into the same filter bank as that of analysis to smooth the envelope of processed spectral amplitude. DCAA can be controlled by 4 bits (16 levels: they are referred to as attenuation coefficients) and waveforms are attenuated by 3 dB per one level. In this hardware, noises are reduced in real time if the short-time spectral amplitude of speech can be estimated within 16 ms interval by the micro processor. Besides, the frame interval and the delay time of BBD can be changed by 4, 8, 16, 32 and 64 ms manually. Therefore, the present waveforms of channels can be attenuated at DCAA by using the short-time spectral amplitude of up to four future frames, if the frame interval is adjusted to 16 ms and the delay time of BBD to 64 ms. In the following section, we describe the algorithms to reduce random noise or acoustic pulsive noise with resonances occuring in the environments by the micro processor of this hardware.

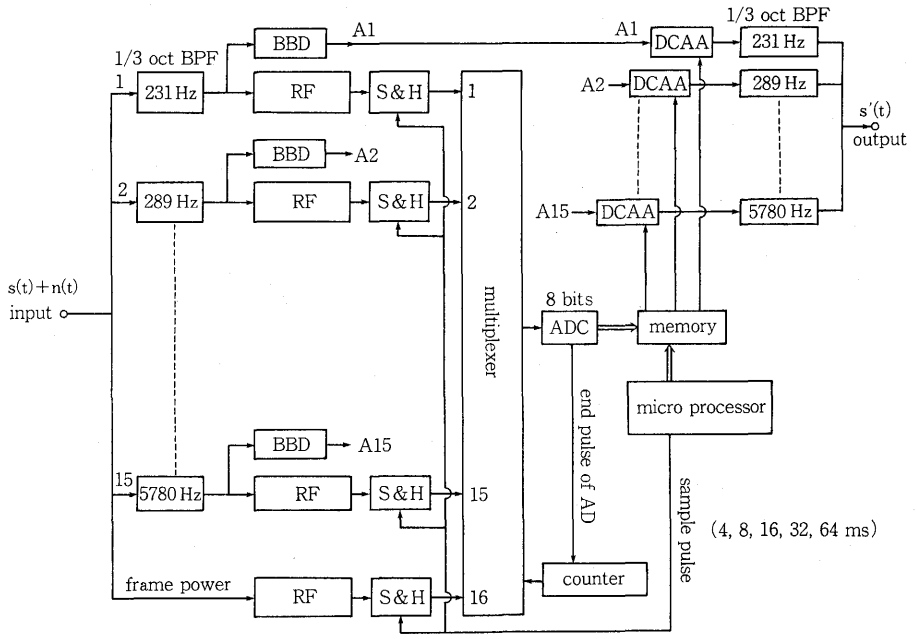

Fig. 4. Hardware organization of real time noise canceling by bandpass filter.

4. Canceling Algorithm for Random Noise[8)]

4.1 Estimation of the Short-time Spectral Amplitude of Random Noise

A sequence of frames is segmented into two kinds of time interval. One is silence interval where the frame power is lower than a certain threshold. The other is sound interval. The short-time spectral amplitude Ni of the noise (i is channel number) is estimated as that in the silence interval.

4.2 Estimation of Short-time Spectral Amplitude of Speech

4.2.1 Technique A—Spectral Subtraction

The estimation of the short-time spectral amplitude of speech $\hat{S}i$ can be obtained by denoting Yi as the short-time spectral amplitude of the present frame of the noisy speech.

$$\hat{S}i = Yi - Ni \tag{12}$$

The above expression corresponds with that of (5) in the case where $k=1$ and $a=1$. The waveforms of the channels are attenuated at DCAA so that the short-time spectral amplitude of waveforms equals that of $\hat{S}i$. The attenuation coefficient j $(0 \leq j \leq 15)$ of DCAA is obtained by the following expression:

$$\frac{\hat{S}i}{Yi} = \frac{1}{\sqrt{2}^{j}} \quad \therefore j = \frac{-1}{\log\sqrt{2}} \cdot \log\frac{\hat{S}i}{Yi} \tag{13}$$

From the above expression, if $\hat{S}i$ equals Yi (no noise), j becomes 0 and if $\hat{S}i$ is estimated to minimum value (noise is strong), j becomes 15. To realize the real time processing the log mapping table is practically used in stead of computing the log function in the expression (13).

4.2.2 Technique B—Selection of Speech Channel

The effect of spectral subtraction for noise reduction is to increase signal-to-noise ratio by suppressing the waveforms of channels with low signal-to-noise ratio. This indicates the another possibility of processing that reduces noise by only passing the waveforms of channels with good signal-to-noise ratio. The practical processing is as follows:

(i) Signal-to-noise ratio of each channel $\hat{S}i/Ni$ is computed according to the expression (12).

(ii) The waveforms of channels whose signal-to-noise ratio is high are passed with attenuation coefficient $j=0$ at DCAA.

(iii) The waveforms of channels whose signal-to-noise ratio is low are suppressed with attenuation coefficient $j=15$ at DCAA.

4.3 Result of Experiment

It is generally known that consonants are important for the intelligibility of speech even though their energy is relatively smaller than that of vowels. The techniques A and B described above suppress the waveforms of channels with low signal-to-noise ratio so that consonants tend to be degraded and consequently the

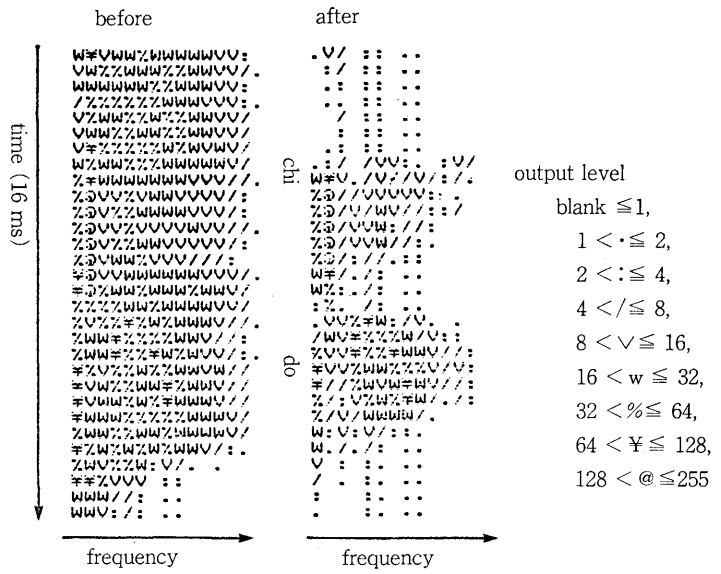

Fig. 5. Spectra before and after processing.

intelligibility of the speech rarely increases. Reduction of listener fatigue is, however conspicuous so that both techniques seem to be useful for practical situation. Fig. 5 shows the spectra before and after processing by technique A. The speech is a part of "Moichido" and signal-to-noise* ratio is 0 dB. The problem of this processing is that the new noise is introduced by the processing due to destruction of smoothness of spectrum in time. Therefore, it is required to interpolate the estimated short-time spectral amplitude of speech between frame intervals and control DCAA in more short time.

5. Canceling Algorithm for Acoustic Pulsive Noise

Here, acoustic pulsive noise indicates the noise with resonances and long duration such as sounds of telephone bell, airplain or bell rings. This type of noise must be dealt with by two different processes, that is, detection and reduction of the noise because the property and time it occures in the environments are *a priori* unknown.

5.1 Detection of Acoustic Pulsive Noise

The basic component of Japanese speech is syllable such as V or CV. On the other hand, acoustic pulsive noise has no such a component and its spectrum is far different from those of vowels, even though it is similar to those of consonants. Therefore, speech and acoustic pulsive noise can be separated by detecting the existence of vowels in noisy speech.

5.1.1 Detection of Noise Section in Time

Spectral change is defined by the following expression:

* In this paper, the signal-to-noise ratio is defined as that of averaged power of speech and noise.

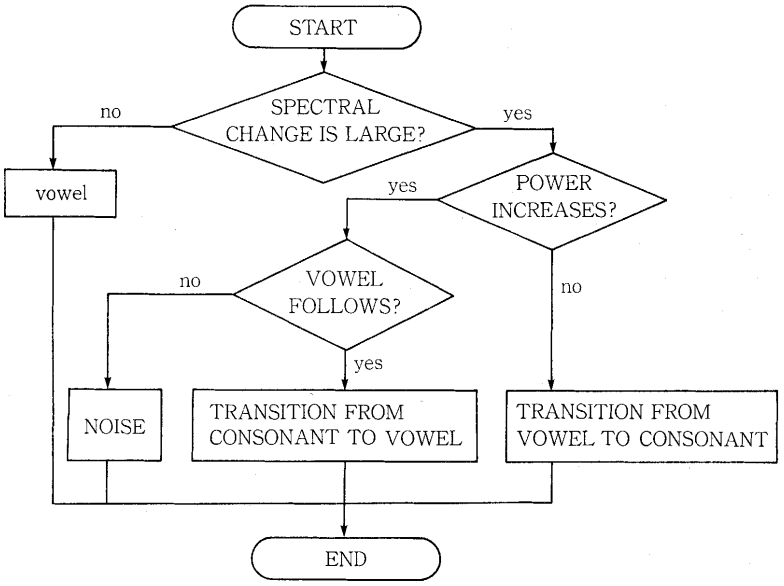

Fig. 6. Detection of noise section in time.

$$\sum_{i=1}^{15} |Yi(t)-Yi(t+1)| \tag{14}$$

where Yi(t) is the short-time spectral amplitude of channel i at time t. Spectral change increases in transition from consonants to vowels, vowels to consonants and the time when acoustic pulsive noise superimposes on speech. When acoustic pulsive noise superimposes, frame power also increases and vowels do no follow. Based on this fact, starting time of noise section can be located by using spectral change, frame power and detection of vowels according to the algorithm illustrated in Fig. 6. Vowel detection can be performed by spectrum matching based on city distance between the short-time spectral amplitude normalized by frame power at present frame and those of standard five vowels /a/, /i/, /u/, /e/, /o/.

5.1.2 Detection of Noisy Channel

The power of acoustic pulsive noise concentrates on a certain frequency band and that frequency band tends to be constant with time. In addition, the power of that frequency band rises up rapidly. According to this property of the noise, the noisy channel can be detected by finding out the channel whose power rises up rapidly at the starting frame of noise section in time.

5.2 Reduction of Acoustic Pulsive Noise

Noise is reduced by suppressing the waveforms of noisy channels in the noise section detected by the process described in the section 5.1. Three levels suppression at DCAA is employed not to degrade the speech when the noise exists on the formant frequency band. Namely, waveforms are passed at the speech channel and they are suppressed by half at the noisy channel on the formant frequency band and they

are fully suppressed at the noisy channel on the anti-formant frequency band.

5.3 Result of Experiment

The acoustic pulsive noises as well as random noise superimposed on the silence interval between the speech can be canceled without degrading the quality of the speech. On the other hand, when such noises are superimposed on the speech, the noises can be reduced with a little degradation of speech.

6. Concluding Remarks

We described the real time noise reduction hardware system and the algorithms to reduce random and acoustic pulsive noises. As evident from our experiment, it is clear that our technique based on a channel vocoder is useful for reducing, in real time, the acoustic pulsive noises which are often heard in real world as well as classical random noises. The future problem is to evaluate the performance of noise reduction, especially, for acoustic pulsive noises and to develop the real time algorithms for reducing noises other than those described in this paper.

References

1) J. S. Lim and A. V. Oppenheim: Enhancement and Bandwidth Compression of Noisy Speech, Proc. IEEE, Vol. 67, No. 12, pp. 1586–1604, 1979.
2) S. F. Boll: Suppression of Acoustic Noise in Speech Using Spectral Subtraction, IEEE Trans., Vol. ASSP-27, No. 2, pp. 113–120, April, 1979.
3) N. Wiener: Extrapolation, Interpolation and Smoothing of Stationary Time Series, with Engineering Application, John Wiley & Sons, 1949.
4) M. W. Callahan: Acoustic Signal Processing Based on the Short-time Spectrum, Ph. D. Disertation, Department of Computer Science, Univ. Utah, 1976.
5) M. R. Sambur: Adaptive Noise Cancelling for Speech Signals, IEEE Trans., Vol. ASSP-26, No. 5, pp. 419–423, Oct., 1979.
6) J. Suzuki: Speech Processing by Splicing of Autocorrelation Function, in Proc. IEEE Int. Conf. ASSP, pp. 113–116, Apr., 1976.
7) T. W. Parsons: Separation of Speech from Interfering Speech by Means of Harmonics Selection, J. Acoust. Soc. Am., Vol. 60, No. 4, pp. 911–918, 1976.
8) M. M. Sondhi, C. E. Schmidt and L. R. Rabiner: Improving the Quality of a Noisy Speech Signal, The Bell System Technical Journal, Vol. 60, No. 8, Oct., 1981.

(Aug. 31, 1982, received)