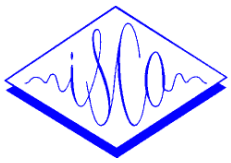

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

REMOTE CONTROL SYSTEM USING SPEECH
-REDUCTION OF KNOWN NOISE-

Tsuyoshi Usagawa, Yuji Morita and Masanao Ebata

Faculty of Engineering, Kumamoto University
2-39-1 Kurokami, Kumamoto 860, JAPAN

ABSTRACT

Surrounding noise seriously affects the performance of speech recognition which can operate under usual noise environment. When we try to make an audio equipment or television set which can be controlled by speech, the sound radiated by the equipment itself also affects the performance of speech recognition. In this study, we propose the method to reduce the latter type noise, a priori known noise. The exponential step normalized LMS algorithm is used for the adaptation of FIR digital filter. Also the method of speech candidate detection is presented using the characteristics of adaptive filter. The experiments are carried out under usual room condition with the several broadcasted sound as noise. The total reduction of noise varies from 15dB to 25dB due to the difference of noise type. There is the correlation between the spectrum of noise and the obtained filter coefficients on adaptive filter.

I. INTRODUCTION

It is widely recognized that one of the key techniques for human-machine interface is speech recognition and speech production. Rapid advancement of present semiconductor technique realizes one or a few chips can construct the speech recognition system, even though, which can work for specific person and also specific isolated words[1]. Now we are expecting that such a word recognition system is going to be used in many fields. Before to make a practical speech recognition system, we need to find a way out of several problems. Surrounding noise is one of the most serious ones. Word recognition chips on a market assume clean or high SNR speech input, so not only the recognition error but also the error of speech segment detection are often observed when input speech signal is degraded by noise[2]. On the other hand, the application of word recognition system aims at so-called "hands-free" system or "voice control" system, for example a remote control of television set using speech. In these application field, not only surrounding noise but also the noise generated by the equipment itself seriously affect the performance of speech recognition.

Many approaches to solve the noise problem are reported using a variety of techniques: speech enhancement techniques, special purpose directional microphone systems, noise reduction in recognition level, and so on[3,4]. In this study, we assume that the dominant noise in observed signal itself or correlated one is available and we propose the method to reduce this kind of noise using adaptive digital filter. Our model system is a remote controller of television set using speech. For the adaptation of Finite-Impulse-Response(FIR) digital filter, we use the Exponential-Step Least-Mean-Square (ES-LMS) algorithm proposed by S.Makino, et al[5], which is the expansion of Normalized LMS algorithm using the characteristics of room acoustics. The ES-LMS algorithm is said that it has double the convergence speed but the same computational load as the conventional NLMS.

The detection of speech candidate in degraded input signal is also very important to get sufficient performance in speech recognition system. As the first level of speech candidate detection, we propose the detection method based on the characteristics of output of adaptive filter.

The experiments of noise reduction are carried out for kinds of broadcasted sound as noise. Also the performance of the detection of speech candidate are examined. The results show the effectiveness of proposed method to reduce a priori known noise and they shows that the amount of noise reduction is affected by the spectral characteristics of noise.

II. REDUCTION OF A PRIORI KNOWN NOISE

2.1 Basic Configuration

The observable signal, $y(k)$, in the presence of broadcasted sound from the television set is described as follows,

$$y(k) = h(k) * x(k) + s(k) + n(k) \quad (1)$$

where $x(k)$ is the broadcasted sound which is directly observable, $s(k)$ is the target speech to be recognized and $n(k)$ is surrounding noise. Also $h(k)$ is the impulse response of room and $*$ means the convolution. The both components of $h(k)*x(k)$ and $n(k)$ must be reduce to recognize speech correctly. Under usual listening condition, the

broadcasted sound is the dominant noise in the observed signal so that surrounding noise can be neglected. So the estimation of speech, $\hat{s}(k)$, is given as follows,

$$\hat{s}(k) \doteqdot y(k) - \hat{h}(k) * x(k) \quad (2)$$

where $\hat{h}(k)$ is the estimated impulse response of room.

2.2 Configuration Using Adaptive Filter

As mentioned in previous section, the reduction of a priori know noise can be realized based on the estimated impulse response of room. If this impulse response is stational, it can be derived using direct measurement methods, such as the cross spectral technique or the cross correlation technique, or the direct transfer function measurement using white noise or impulse. For the practical usage, the impulse response of room is hardly stational, so that it should be estimated using adaptive algorithm. The configuration of the proposed system is depicted in Fig.1, in which the block surrounded by dotted line represents the adaptive digital filter. The configuration of the proposed noise reduction system is similar to that of an acoustic echo canceller. Recently many adaptive algorithm are reported to get faster convergence speed and more ERLE(Echo Return Loss Enhancement). In those algorithm, the ES-LMS algorithm proposed by S. Makino, et al, has a significant advantage for our purpose. The ES-LMS algorithm is the expansion of NLMS using the characteristics of a room acoustics. As generally known, the impulse response of room is exponentially decayed and the ratio of decay corresponds to the reverberant energy decay curve. Based on this property of room acoustics, the step size of ES-LMS algorithm is varied for each filter coefficient and it decreases exponentially, where the one of conventional NLMS algorithm is constant for every filter coefficient. It is said that the ES-LMS algorithm has double the convergence speed but the same computational load as the conventional NLMS.

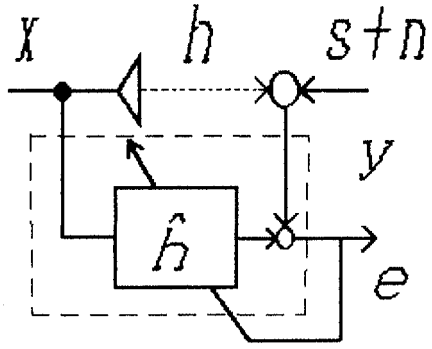

Fig.1 Configuration of proposed system

2.3 Detection Of Speech

Under usual noise environment, the input signal contents not only the a priori know noise and speech to be recognized but also ambient noise. Generally ambient noise can be classified into stational noise such as traffic noise or fun noise from air conditioner, and non-stational one such as noise due to door opening and closing or conversation not to be recognized by word recognition system. When the input signal contents the word expected to be recognized or an ambient noise, the output level of the adaptive filter rapidly increases but the adaptation must not work while such a speech or an ambient noise is presented. On the other hand if the increase of output level is due to the change of the characteristics of a room acoustics (a propagation path of a priori known noise), in this case the increase is also drastic, the adaptation must work to get the new optimum coefficients of filter as soon as possible. To realize the remote control system using speech, the candidate word must be distinguished in such situation.

The candidate word can be picked up using the proposed configuration of the adaptive filter as follows. Now, let the absolute value of the filter output be $|e|$, the minimum value of $|e|$ in last 1s multiplied by constant(k) be θ, the standard deviation of $|e|$ in last 1s be σ. The situations mentioned above are classified into three categories as shown in Table.1.

In case I, the adaptation works to get the optimum coefficients and both θ and σ are updated. In case II, the output of filter is treated as a candidate word to be recognized. The adaptation is stopped and only θ is updated while the condition $|e|>\theta$ is satisfied. Because this strategy cannot make no distinction between speech and non-stational ambient noise, the classification is left for the word recognition level. In case III, there are two possible reasons of the increase of output level, i.e. the presence of the stational ambient noise and the change of the propagation path. When the increase is assigned to the former reason, $|e|$ will not decreases even if the

Table.1 Classification of input signal based on parameters derived from adaptive filter output.

case	*condition*	*Input contents*
I.	$\|e\|<\theta$ ---------->	only a priori known noise.
II.	$\|e\|>\theta$ and $\sigma>\alpha$ ----->	speech or non-stational ambient noise.
III.	$\|e\|<\theta$ and $\sigma<\alpha$ ----->	stational ambient noise or change of a propagation path.

where α is the constant

adaptation works. On the other hand, when it is assigned to the latter one, $|e|$ will decrease step by step and the new optimum coefficients are given. Just after the propagation path is changed or an ambient stational noise is started, there is no way to distinguish case III from case II. In this case, the adaptation stops once until the deviation of $|e|$ comes to small so that $\sigma<\alpha$. Due to the fluctuation of $|e|$, the averaging over 10ms is required to obtain stable $|e|$. Furthermore, the transient point from case I to case II is somewhat delayed from the beginning of speech. The delay is mainly due to the small power level of the first consonant of word, so the pretriggering function is equiped in the proposed filter.

III. EXPERIMENTS

3.1 Experiment Model

The experiments are carried out in a laboratory room whose dimension is 9.0m(L) x 5.5m(W) x 2.8m(H). The reverberation time of this room is approximately 300ms and the sampling frequency of the digital filter is set to 10kHz, so that the number of filter taps is set to 4096. The distance between noise source and pickup microphone is 1.4m and spoken words are recorded in the same room without noise. Noise and speech are numerically mixed on the computer with certain SNR. Two series of experiments are carried out; the reduction of kinds of broadcasted noise and the discrimination of speech from an ambient noise.

3.2 Results I

In the following sections, the effectiveness of the proposed method is evaluated using the Noise Reduction Level (NRL) which is defined as follows after ERLE in an echo canceller system.

$$NRL = 10 \log_{10} \frac{\text{power of } y(k)}{\text{power of } e(k)}$$

The speech "kumamoto" uttered by an adult male is mixed with white noise and 3 broadcasted programs; male voice with percussion, piano solo, organ solo. All of the filter coefficients are set to be zero and NRL values are presented after 3s (30000 samples). Table 2 shows the obtained NRL

Table.2 Reduction of a priori known noise using the proposed method.

	Obtained NRL(dB)	
Type of Noise	LMS	ES-LMS
White Noise	15.6	17.2
Male Voice and percussion	12.0	15.6
Piano Solo	16.6	22.0
Organ Solo	14.6	16.5

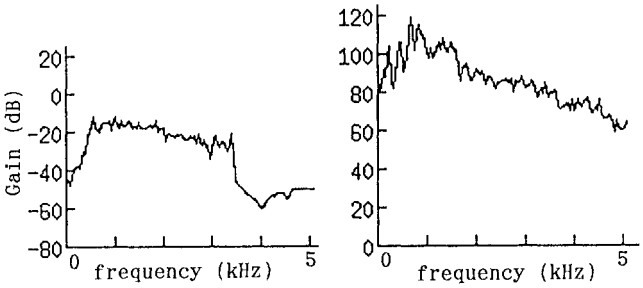

Fig.2 Experiment result when the noise is male voice with percussion.
(a) Transfer function of adaptive filter.
(b) spectrum of noise.

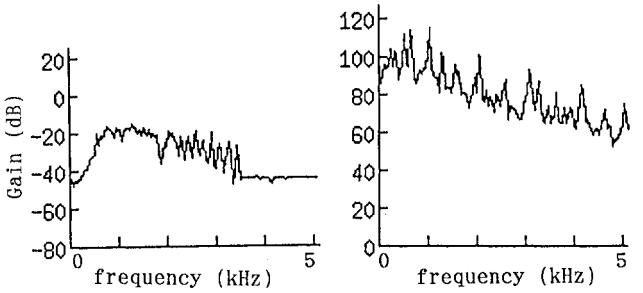

Fig.3 Experiment result when the noise is piano solo.
(a) Transfer function of adaptive filter.
(b) spectrum of noise.

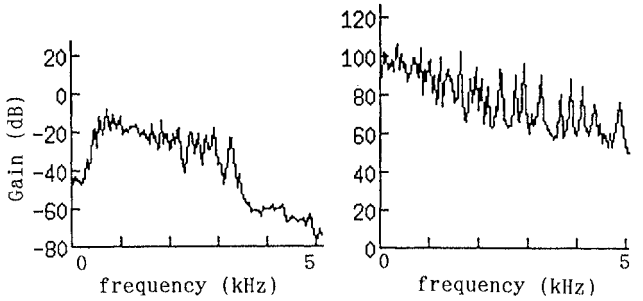

Fig.4 Experiment result when the noise is organ solo.
(a) Transfer function of adaptive filter.
(b) spectrum of noise.

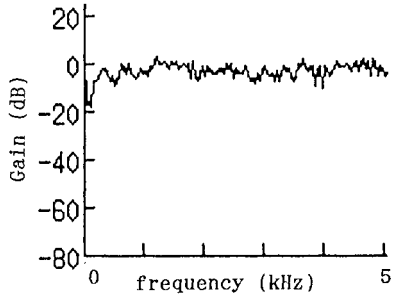

Fig.5 Transfer function obtained by the cross spectrum technique.

using the same configuration except adaptation algorithm, LMS and ES-LMS. Shown in Table 2., NRL comes up 12dB to 16dB using LMS and more than that using ES-LMS at the same convergence time. Also the obtained NRL is changed due to the type of the known noise. To show this evidence in frequency domain, we translate the obtained coefficient of digital filter into spectra as shown in Fig.2(a), 3(a), and 4(a). In these figures, the power spectra of a priori known noise are also presented as figures (b) for the comparison with spectra pattern of the obtained transfer function. Figure 5 shows the transfer function between $x(k)$ and $y(k)$ which is measured by the cross spectral technique using white noise. By the comparison with the transfer function given by the cross spectrum technique shown in Fig.5, the obtained transfer functions on the adaptive filter shown in Fig.2-4 have correlations with the spectrum structure of noise. On the practical usage for a television set, this affection might be serious problem for the effective noise reduction because the broadcasted source is rapidly and constantly changed.

3.3 Results II

Figure 6 and 7 show the results of the determination of speech candidate within the white noise whose level is set at 0dB and +10dB, respectively, to the male speech "video". In each figure, the top most shows the input signal wave form y, the second is the output of the digital filter e, and the third line which has only two level is the status of the adaptation runs or not, and the lowest line is the absolute value of the filter output $|e|$ shown with the threshold. Also in these figures, the obtained candidate ranges are presented as broken lines just above the line of $|e|$. In case of 0dB SNR, the errors of the beginning and the ending points of candidate speech are no more than 70 sample points (7ms). And in case of -10dB SNR, the error of the ending point is much larger and is almost 1000 sample points (100ms), however the error of the beginning point is about 270 sample points (27ms). The accuracy of the proposed determination algorithm is sufficient as a first step to determine a candidate speech.

IV. CONCLUSION

This paper proposes the method to reduce the effect of a priori known noise to realize a remote control system using spoken command words. The given reduction of noise in the experiments is more than 15dB.

REFERENCES

[1] For example, μPD7764 in "Signal Processing LSI Data Book," pp.532-562, NEC (1989)

[2] T.Usagawa, Y.Sumi, M.Ebata and J.Okda, "A Consideration on Improvement of Speech Recoginition in Noisy Environment," Proc.IECIE,J70-A, pp.1854-1857 (1987) (in Japanese)

[3] J.S.Lim,*Speech Enhancement*, Prentice-Hall,(1983)

[4] Y.Kaneda and J.Ohga,"Adaptive microphone-array system for noise reduction," IEEE Trans. ASSP, ASSP-34(6), pp.1391-1400 (1986)

[5] S.Makino and Y.Kaneda, "Acoustic Echo Canceller Algorithm Based on the Variation Characteristics of a Room Impulse Response," Proc. ICASSP90, pp.1133-1136, (1990)

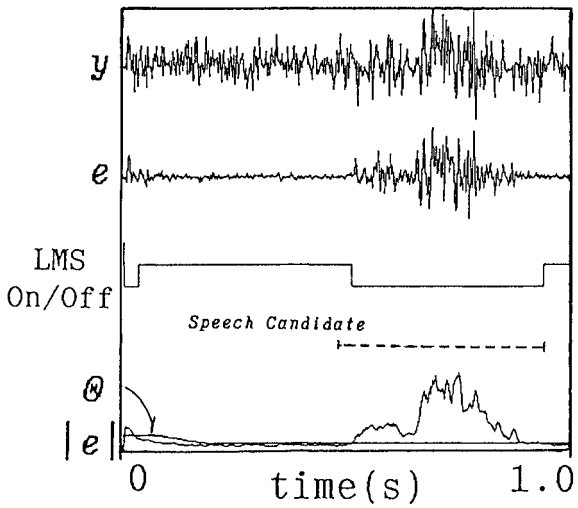

Fig.6 Determination of speech segment (SNR=0dB)

y : Speech signal with noise.
e : Noise reduced signal.
LMS : Adaptation on/off status.
Θ : Threshold for speech detection.
$|e|$: Averaged absolute value of e.

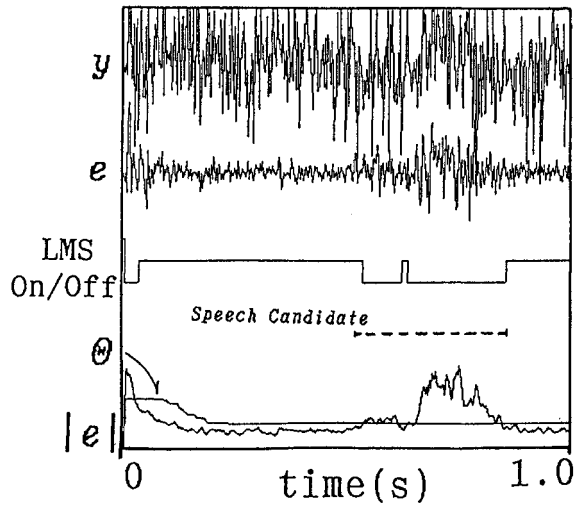

Fig.7 Determination of speech segment (SNR=-10dB)

y : Speech signal with noise.
e : Noise reduced signal.
LMS : Adaptation on/off status.
Θ : Threshold for speech detection.
$|e|$: Averaged absolute value of e.