VOICE CONTROL SYSTEM FOR TELEVISION SET USING MASKING MODEL AS A FRONT-END OF SPEECH RECOGNIZER

Tsuyoshi Usagawa, Makoto Iwata and Masanao Ebata

Department of Electrical Engineering and Computer Science
Kumamoto University, 2-39-1 Kurokami, Kumamoto 860, JAPAN
E-mail : tuie@eecs.kumamoto-u.ac.jp (T.Usagawa)

ABSTRACT

Surrounding noise often affects the performance of speech recognition system when it is used in office or home. Especially situation is more serious when colored and nonstational noise such as an sound from television or other audio equipment is introduced. The authors proposed a voice control system for television set using an adaptive noise canceler, and it works well even if sound of television set has comparable level of speech. In this paper, a new front-end of speech recognition is introduced for the voice control system. This font-end utilizes a simplified masking model to reduce the effect of residual noise. According to experimental results, 90% correct recognition is achieved even if the level of television sound is almost 15dB higher than one of speech.

1 INTRODUCTION

To make a speech recognition system robust enough for daily life usage, we should overcome several problems. Surrounding noise is one of the most serious ones because it affects the performance even if the noise level is low enough for human being to understand speech. Especially, colored and nonstational noise, such as human voice and music sound, affects the performance of speech recognition system seriously. In this paper, we revise the previous voice control system for television set to improve the robustness against noise[1]. This system utilizes noise-robust parameter extraction method based on simplified masking model. The proposed method does not depend on noise level unlike the method of noise injection to templates, and it can work properly for nonstationary noise unlike spectral subtraction method. Also because of simplicity of model, the proposed method requires small extra computational load.

2 CONFIGURATION OF VOICE CONTROL SYSTEM

As shown in Fig. 1, the proposed voice control system is composed of three components; 1) reduction of television sound by means of adaptive noise canceler, 2) determination of speech candidate and 3) speech recognition using masking model.

2.1 Reduction of Television Sound

As shown in Fig. 1, the observable signal $y(k)$ in the presence of television sound $x(k)$ is described as follows,

$$y(k) = h(k) * x(k) + s(k) + n(k),$$

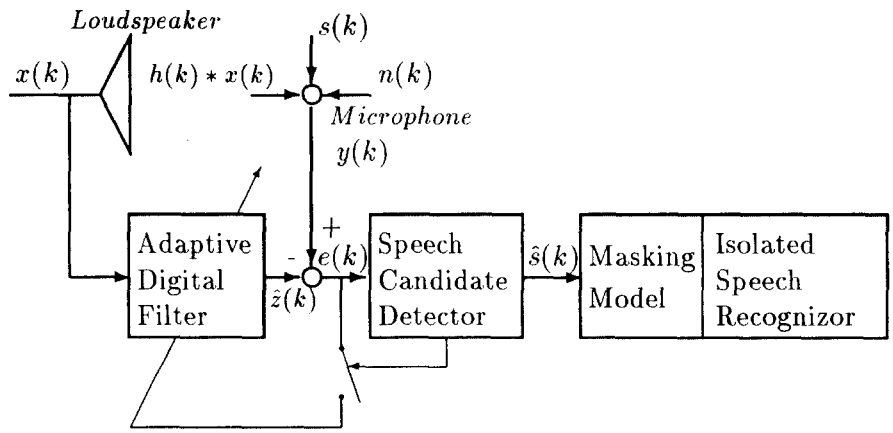


Figure 1: Schematic Block Diagram of Voice Control System of Television Set

where $s(k)$ is a target speech to be recognized and $n(k)$ is an ambient noise. And $h(k)$ is an impulse response from the loudspeaker of television set to the microphone and $*$ means the convolution. In case of usual situation, the component of $h(k) * s(k)$ can be treated as dominant in comparison with $n(k)$, so that the estimation of speech $\hat{s}(k)$ can be obtained as a residual error $e(k)$ using an estimation of impulse response of room as follows,

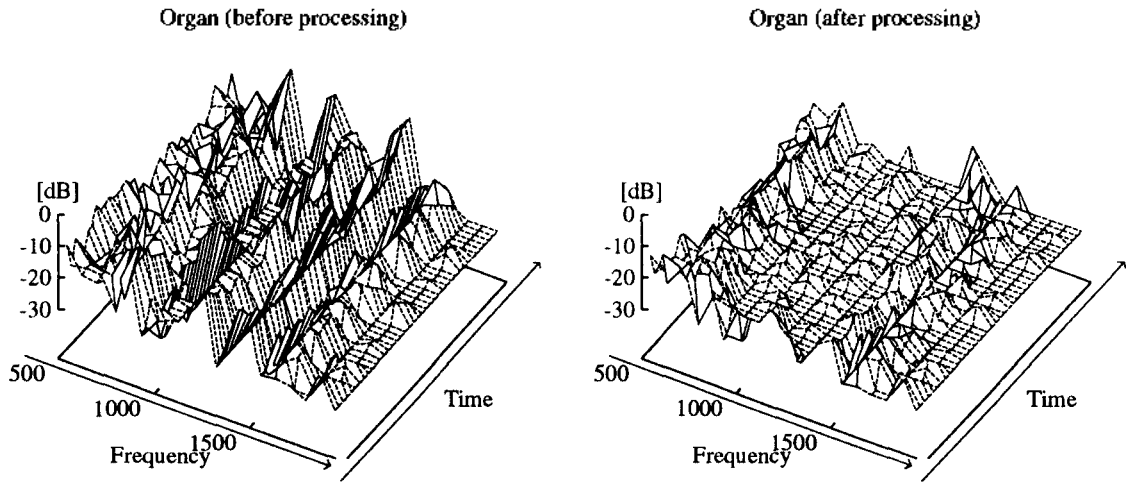
$$\hat{s}(k) \approx e(k) = y(k) - \hat{h}(k) * x(k).$$

This relationship means the reduction of television sound achieved by means of the estimation of impulse response. As well known, such an impulse response is easily changed so that the adaptive digital filter technique is adequate. The requirements to the adaptive algorithm are similar to ones for an acoustic echo canceler. That is the stability and good convergence speed even if the introduced signal is colored and nonstational one. There are many adaptive algorithms concerned with the characteristics for colored input signal. We use the variable tap length LMS (VT-LMS) algorithm which has very good convergence property for colored noise[2]. Figure 2 (a) and (b) show the running power spectra of observed and noise reduced signals, respectively. The signal source is organ solo, so that many harmonic components are observed in Fig. 2(a). On the other hand, the running spectrum of residual noise is flattened as shown in Fig. 2(b). This characteristics of residual signal is very attractive to use masking model for further noise reduction at speech recognition.

2.2 Determination of Speech Candidate

Determination of speech candidate is necessary not only for speech recognition but also for control of adaptive digital filter. When speech is observed, the adaptation has to be suspended like a double-talk condition of an acoustic echo canceler. On the other hand when the propagation path of television sound is changed, the adaptation to estimate new impulse response is required. In both cases, the output level of adaptive digital filter is increased so that the discrimination of these two cases is important issue.

In this paper, the determination is carried out based on the correlation between $\hat{z}(k)$ and $e(k)$[3]. At first, we need to determine the output level of adaptive digital filter is changed



(a) Observed Signal (b) Residual Noise

Figure 2: An Example of Running Power Spectra when Organ Solo is used as sound from television set: (a) Observed Signal, (b) Residual Noise

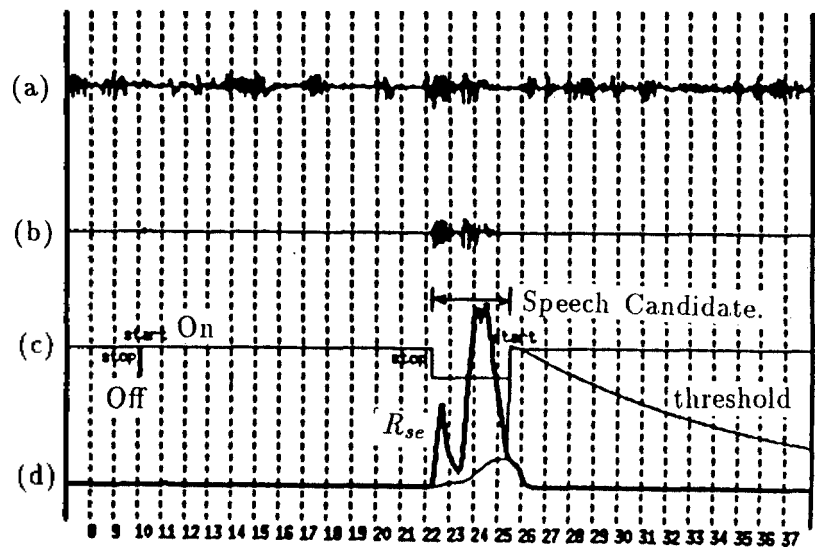


Figure 3: An Example of Determination of Speech Candidate. (a) $y(n)$, (b) $e(n)$, (c) On/Off of Adaptation and (d) R_{se} (thick) with threshold (thin).

or not. An index R_{se} as follows,

$$R_{se} = \log_{10} \frac{E[e^2(k)]}{E[\hat{z}^2(k)]},$$

shows the relative output level so that it can be used to detect the change of situation. R_{se} increases in both cases; change of impulse response and presence of speech. The determination of these two case is performed using short term cross correlation between $\hat{z}(k)$ and $e(k)$. Figure 3 shows an example of determination of speech candidate. Lines

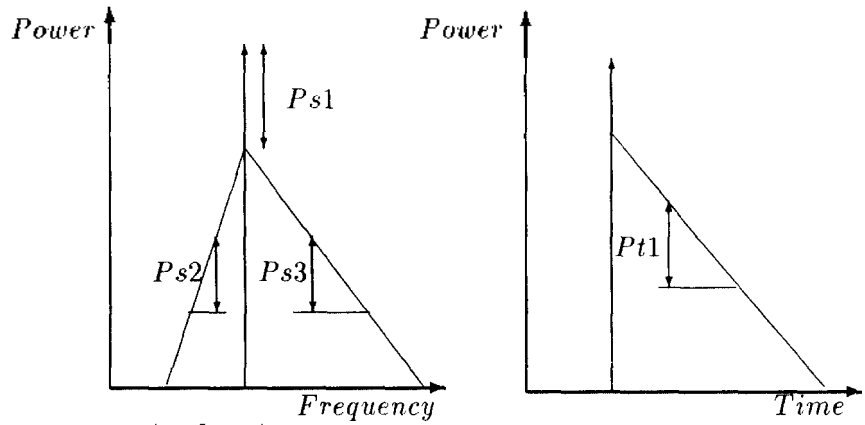


Figure 4: Schematic Representation of Masking Model

(a), (b), (c) and (d) show observed signal $y(n)$, residual of noise reduction $e(n)$, status of adaptation ON/OFF and R_{se} (thick) with the threshold (thin), resplectively. The abscissa shows time in frame with 100ms interval. At 10th frame, the echo path was changed but the adaptation is not suspended for a long term. On the other hand, speech is added from 22th to 25th and it is detected as shown in line (c) as stop of adaptation.

2.3 Speech Recognition using Masking Model

Usually speech parameters are extracted from the power spectrum of observed signal. Extracted speech parameters by conventional method are affected by even low level noise. On the other hand, as known as masking phenomenon, we can not detect an acoustic signal whose level is lower than the masking threshold generated by other acoustic signals. When we listen to speech under noisy environment, noise is suppressed by masking mechanism. To utilize masking mechanism as a font-end of speech recognition system, we need a model of masking mechanism which is sufficiently simple for practical usage.

There are three types of masking; forward, simultaneous and backward masking. The effect of backward masking is rather small so that our model concerns only simultaneous and forward masking. To make the model simple, the spectral shape of simultaneous masking threshold is modeled as a triangle; i.e. a slope of spectrum is linear against power level along with logarithmic frequency scale. Also forward masking is modeled by a constant decay ratio for each analyzing frame[4]. Figure 4 schematic models of simultaneous and forward masking, respectively. And parameters of masking model are decided according to masking audiogram derived by psychoacoustical experiments and obtained from a literature[5][6]. Frequency range is divided into 4 frequency bands.

This masking model is introduced between power spectrum observation and LPC analysis. At first like other font-end, the proposed method obtains the power spectrum of input signal using FFT. Based on the power spectra of current and previous analyzing frame, the masking threshold level is decided. Figure 5 shows an example of observed and masked power spectra. Thin line shows the directly observed power spectrum and thick line shows the masked one. As parameters for speech recognition, 15 LPC cepstrum coefficients and power level are obtained for each analyzing frame. And LPC cepstrum coefficients are

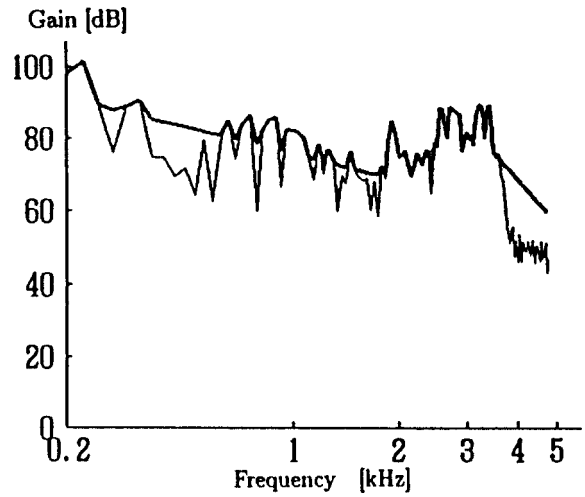


Figure 5: Example Power Spectrum. Thin line is observed power spectrum and thick line is masked one.

extracted from an autocorrelation function, which is derived by inverse FFT of observed masked power spectrum.

3 EXPERIMENTAL RESULTS

The masking model is implemented as a font-end of speaker dependent isolated word speech recognition system. The vocabulary words are selected to simulate the voice control system of television set such as volume control, channel control, and so on. Three kinds of signal are used as a television sound; white noise as a reference, male voice and organ solo as examples of colored and nonstational noise.

Figures 6, 7 and 8 show the experimental results as the ratio of correct recognition versus SNR for white noise, male voice and organ solo, respectively. The results labeled *Direct* are ones obtained by conventional method, and the results labeled *LMS* are obtained by the previous system which has the adaptive noise canceler. The results labeled *With Masking* show the performance given by the proposed system which uses the proposed font-end and the adaptive noise canceler. As shown in this figure, the overall improvement of SNR is almost $20dB$ for each case, and the font-end based on the masking model improves almost $10dB$ in SNR. As a result, this speech recognition system can work even if SNR is less than $-15dB$.

4 CONCLUSION

In this paper, we proposed the voice control system for television set using masking model. Combined with noise reduction by adaptive digital filter, the proposed method is very effective for not only steady noise but also colored and nonstational noise. From the results of experiment, the proposed system can work even if SNR is less than $-15dB$.

References

[1] T.Usagawa, Y.Morita and M.Ebata, J.Acoust.Soc.Jpn.(E), 13(5), pp.295-300 (1992)
[2] T.Usagawa, H.Matsuo, Y.Morita and M.Ebata, IECIE Trans. Fundamentals, E75-A(11) (1992)

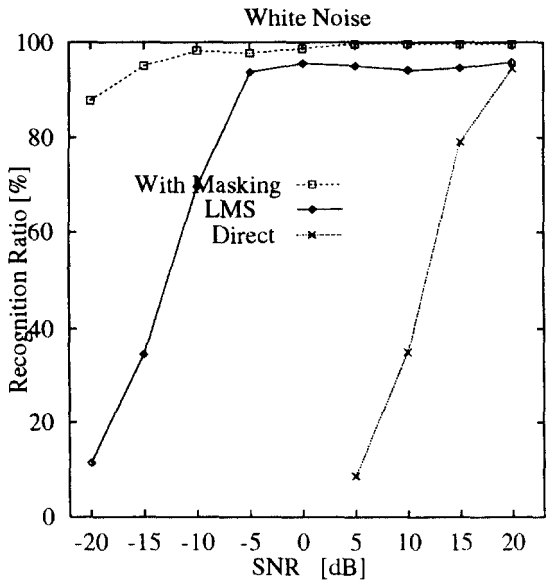


Figure 6: Results (White Noise)

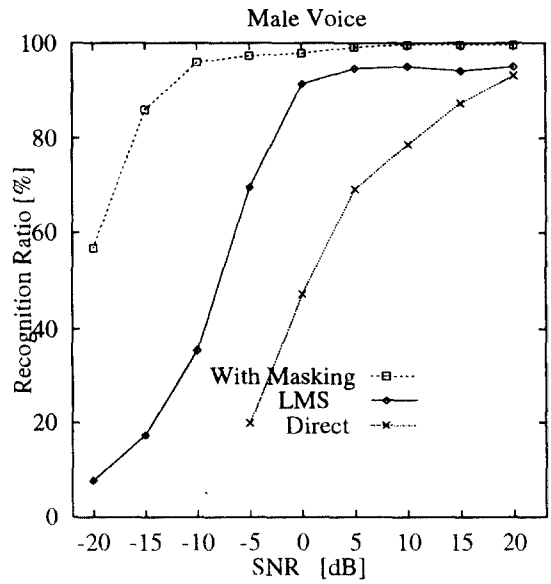


Figure 7: Results (Male Voice)

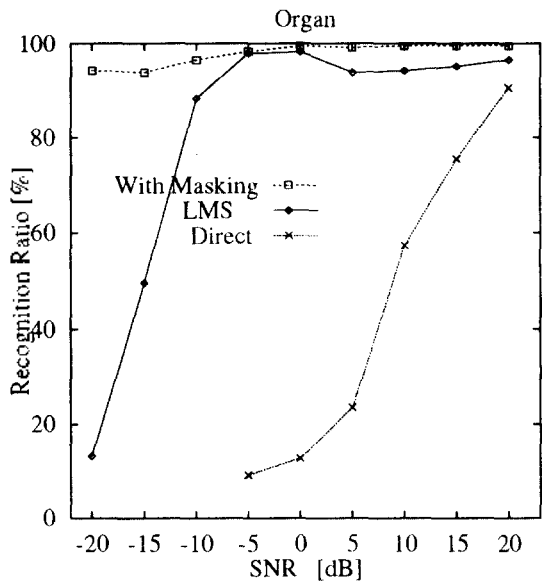


Figure 8: Results (Organ Solo)

[3] K.Morise, K.Utsumi, T.Usagawa and M.Ebata, Tech. Report of IEICE, EA93-55 (1993)

[4] H. Matsuo, T. Usagawa and M. Ebata, Technical Report of Inst. Elect. Info. Com. Eng., EA93-23, pp.9-15 (1992) (in Japanese)

[5] H. Feltcher, *Speech and Hearing in Communication* (D. Van Nostrand Co. Inc., 1953) pp.68-69,

[6] M. Tagawa, T. Usagawa and M. Ebata, Proc. Spring Meeting of Acoustic Society of Japan, No.1-6-5 (1992) (in Japanese)