Simultaneous Adaptation of Echo Cancellation and Spectral Subtraction for In-Car Speech Recognition

Osamu Ichikawa and Masafumi Nishimura

Tokyo Research Laboratory, IBM Japan Ltd., Yamato-shi, Kanagawa-ken, Japan, 242-8502
{ICHIKAW,NISIMURA}@jp.ibm.com

Abstract

For noise robustness of in-car speech recognition, most of the current systems are based on the assumption that there is only a stationary cruising noise. Therefore, the recognition rate is greatly reduced when there is music or news coming from a radio or a CD player in the car. Since reference signals are available from such in-vehicle units, there is great hope that echo cancellers can eliminate the echo component in the observed noisy signals. However, previous research reported that the performance of an echo canceller is degraded in very noisy conditions. This implies it is desirable to combine the processes of echo cancellation and noise reduction. In this paper, we propose a system that uses echo cancellation and spectral subtraction simultaneously. A stationary noise component for spectral subtraction is estimated through the adaptation of an echo canceller. In our experiments, this system significantly reduced the errors in automatic speech recognition compared with the conventional combination of echo cancellation and spectral subtraction.

1. Introduction

Automatic speech recognition is widely used in cars to input commands for car navigation and hands-free telephone dialers. However, as most of the current systems are based on the techniques of multi-conditional training and spectral subtraction, they rely on the assumption that there is only a stationary cruising noise. Therefore, the recognition rate is degraded when there is music or news coming from a radio or a CD player in the car.

Music and news are actually non-stationary noises. However, we may have a chance to eliminate those components by using an echo canceller, because we can utilize the reference signals from such devices.

Previous research reported that an echo canceller works well in a quiet environment. However its performance is poor for low signal to noise ratios [1]. There has been a lot of research on ways to improve the performance of echo cancellers along with noise reduction [2-5]. However, many of the target uses were for teleconference and hands-free telephones, where auditory intelligibility has the highest priority. Our objective is to find a solution for automatic speech recognition with high performance echo cancellation and noise reduction. Our second objective is to retain practical compatibility with the current acoustic model trained with stationary cruising noise and spectral subtraction. In this paper, we assume the cruising noise can be treated as stationary.

2. Conventional Methods

In order to improve the performance of an echo canceller in a noisy environment, the background noise should be reduced before echo cancellation. If many microphones are available, a beamformer can be used to reduce the noise before or at the same step as the echo cancellation [6][7].

Since we assume a single microphone, we need to consider one-channel noise reduction instead of using a beamformer approach. For automatic speech recognition, spectral subtraction is often used because of the computational cost and the performance. As the output is not for humans, the annoying side effect known as musical noise is acceptable. However, the problem for this application is that we cannot place the noise reduction stage before the echo canceller because of the nonlinearity in the echo path [1]. Therefore, the conventional combination is echo cancellation first and noise reduction second, as shown in Fig. 1.

If the echo canceller is implemented using spectral subtraction, the noise reduction stage can be placed before or at the same step as the echo canceller, and we can expect better performance. However, the question is how to estimate the stationary noise spectrum for the noise reduction under the influence of the echo. If the application is a telephone, we can expect noise-only periods in which no one is speaking, so that we may use "spectral minima tracking" [8]. However, we cannot expect such a period in our application, because a car-radio or a car-CD produces sound continuously. Therefore, we propose a new method that estimates the stationary noise spectrum during the adaptation of the echo canceller using spectral subtraction.

Dreiseitel and Gustafsson placed a time-domain echo canceller before the combination of noise reduction and echo canceller in spectral subtraction form [4][8]. By preprocessing the input using the echo canceller, the stationary noise can be estimated more reliably at the noise reduction stage. Our proposed method can also work as an echo suppressor with this type of preprocessing for further improvement.

Since the reverberation in a car is longer than the processing frame, it degrades the performance of frame-based echo cancellation using spectral subtraction. In order to solve this problem, Sakauchi et al. introduced a second term, a scaled echo component estimated in the previous frame [5]. However, their scaling factor should be preset depending on the reverberation in the room. In contrast, our system does not require any a priori knowledge about the room reverberation, because we extended the echo cancellation to refer to the last several frames, and the factors can be determined through the adaptation. The structure is similar to the taps of an adaptive filter in the time domain. In this way, our echo canceller can be adapted to cancel the echoes including the reverberations from past frames.

3. Proposed Method

Fig. 2 shows a block diagram of our proposed Method 1 (without preprocessing), and Fig. 3 shows our proposed

Method 2 (with preprocessing). As the preprocessing stage is a standard N-LMS echo canceller in the time domain, we describe our method after the preprocessor.

The echo canceller stage and the spectral subtraction stage are integrated into the same stage. This estimates both the stationary noise power $\overline{N_\omega}$ and the echo power $Q_\omega(T)$. They are subtracted from the observed noise power $X_\omega(T)$ with the subtraction weights α_1 and α_2, respectively. The compensated output $Y_\omega(T)$ is written as Equation (1).

$$Y_\omega(T) = X_\omega(T) - \alpha_2 \cdot Q_\omega(T) - \alpha_1 \cdot \overline{N_\omega} \qquad (1)$$

Here, T is a frame number. The index ω is a bin number of the DFT corresponding to the sub-band frequency, and the process described in this section should be performed for each ω.

In general, flooring is an essential technique for spectral subtraction. The floored output $Z_\omega(T)$ is given by Equations (2a) and (2b).

$$Z_\omega(T) = Y_\omega(T) \quad \text{if} \quad Y_\omega(T) \geq \beta \cdot \overline{N_\omega} \qquad (2a)$$

$$Z_\omega(T) = \beta \cdot \overline{N_\omega} \quad \text{if} \quad Y_\omega(T) < \beta \cdot \overline{N_\omega} \qquad (2b)$$

Here, β is a flooring coefficient. α_1 and β should be set to the same value with which the acoustic model was trained. The value of α_2 can be larger than α_1 for over-subtraction in order to cancel more of the echo component, which has a large effect on the performance of automatic speech recognition. We introduce an over-subtraction factor γ as Equation (3).

$$\alpha_2 = \gamma \cdot \alpha_1 \qquad (3)$$

Next we describe how to estimate $\overline{N_\omega}$ and $Q_\omega(T)$. The value of $Q_\omega(T)$ is estimated as the weighted sum of the reference signal power $R_\omega(T)$ for the present and the most recent L frames so as to cope with reverberation that lasts longer than the processing frame.

$$Q_\omega(T) = \sum_{l=0}^{L-1} W_\omega(l) \cdot R_\omega(T-l) \qquad (4)$$

For convenience, $\overline{N_\omega}$ is formulated as a product of an arbitrary constant C_0 and its weight.

$$\overline{N_\omega} = W_\omega(L) \cdot C_0 \qquad (5)$$

Although we only consider the stationary cruising noise of a car, the stationary noise power may fluctuate around the average in the frame-wise observation, so $\overline{N_\omega}$ can be estimated as an averaged value. Fig. 4 shows the concept of the estimation.

Therefore, our goal is to estimate the non-negative adaptive weights $W_\omega(l)$ where l ranges from 0 to L. They should be set so as to minimize Equation (6) during non-speech periods with the subtraction weights α_1 and α_2 set to 1.

$$\Phi_\omega = E\left[\{D_\omega(T)\}^2\right] \qquad (6)$$

Here, $D_\omega(T)$ is the error signal as defined in Equation (7). $E[\ ]$

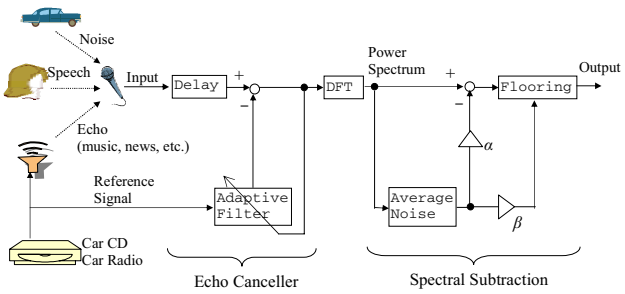

Figure 1: Conventional combination of EC and SS

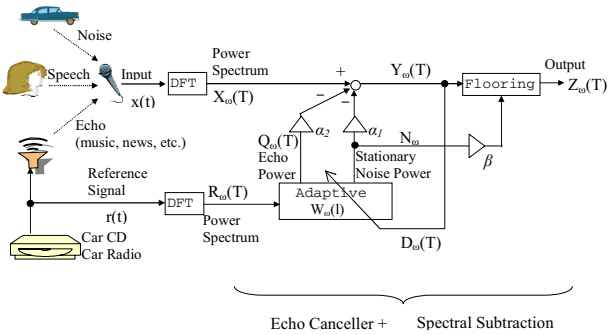

Figure 2: Proposed Method 1 (without preprocessor)

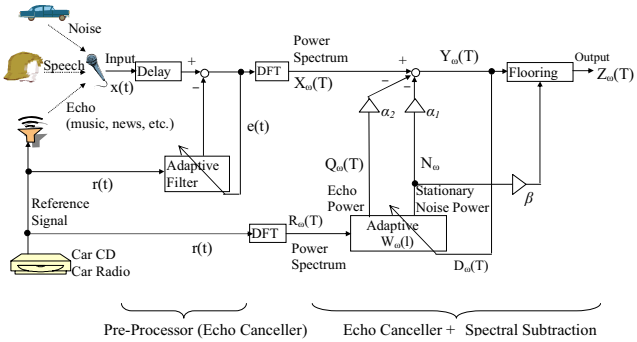

Figure 3: Proposed Method 2 (with preprocessor)

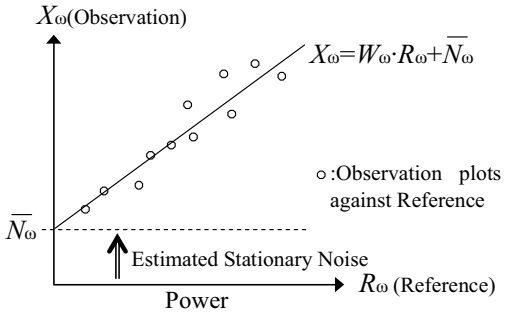

Figure 4: The concept of the estimation of $\overline{N_\omega}$ through the adaptation of W_ω. Here, L=1 for the simplicity.

Table 1: Signal to Noise Ratio

(dB)	Stationary	City Drv.	Highway
S/N_{cruise}	10.5	4.5	2.6
S/N_{music}	10.1	6.5	9.8
S/N_{all}	6.4	1.1	1.2

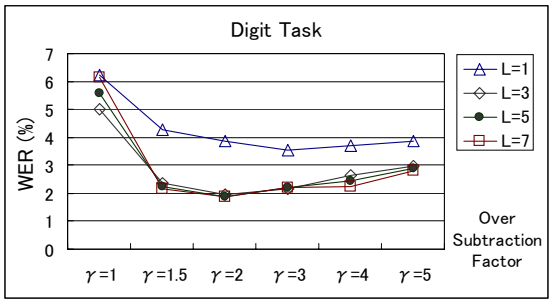

Figure 5: WER using the proposed Method 1 for various values of the over-subtraction factor γ and the length of the adaptive weights L, for the Digit Recognition Task

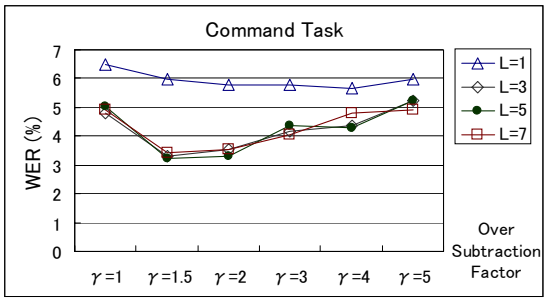

Figure 6: WER using the proposed Method 1 for various values of the over-subtraction factor γ and the length of the adaptive weights L, for the Command Recognition Task

denotes the expectation operator and we calculate it as the frame-wise average during non-speech periods.

$$D_\omega(T) = X_\omega(T) - Q_\omega(T) - \overline{N_\omega}$$

$$= X_\omega(T) - [R_\omega(T), \quad \cdots \quad R_\omega(T-L+1), \quad C_0] \cdot \begin{bmatrix} W_\omega(0) \\ \vdots \\ W_\omega(L-1) \\ W_\omega(L) \end{bmatrix} \tag{7}$$

The values of $W_\omega(l)$ can be determined from $\frac{\partial \Phi_\omega}{\partial W_\omega(l)} = 0 \cdot$

This can be expressed in a matrix and vectors as Equations (8) to (11).

$$\mathbf{C}_\omega = \mathbf{A}_\omega \cdot \mathbf{B}_\omega \tag{8}$$

$$\mathbf{A}_\omega = \begin{bmatrix} \sum_T R_\omega(T) \cdot R_\omega(T) & \cdots & \sum_T R_\omega(T-L+1) \cdot R_\omega(T) & \sum_T C_0 \cdot R_\omega(T) \\ \vdots & \ddots & \vdots & \vdots \\ \sum_T R_\omega(T) \cdot R_\omega(T-L+1) & \cdots & \sum_T R_\omega(T-L+1) \cdot R_\omega(T-L+1) & \sum_T C_0 \cdot R_\omega(T-L+1) \\ \sum_T R_\omega(T) \cdot C_0 & \cdots & \sum_T R_\omega(T-L+1) \cdot C_0 & \sum_T C_0 \cdot C_0 \end{bmatrix} \tag{9}$$

$$\mathbf{B}_\omega = \begin{bmatrix} W_\omega(0) \\ \vdots \\ W_\omega(L-1) \\ W_\omega(L) \end{bmatrix} \tag{10}$$

$$\mathbf{C}_\omega = \begin{bmatrix} \sum_T R_\omega(T) \cdot X_\omega(T) \\ \vdots \\ \sum_T R_\omega(T-L+1) \cdot X_\omega(T) \\ \sum_T C_0 \cdot X_\omega(T) \end{bmatrix} \tag{11}$$

The values of $W_\omega(l)$ can be determined from Equation (12).

$$\mathbf{B}_\omega = \mathbf{A}_\omega^{-1} \cdot \mathbf{C}_\omega \tag{12}$$

Since this off-line form requires the inverse matrix, it has considerable computational cost. We can formulate the adaptive weights $W_\omega(l)$ so as to be successively updated in each non-speech frame using Equations (13a), (13b) and (14). The parameter θ is an updating factor and ε is a constant for stability.

$$\Delta W_\omega(l) = \theta \cdot \frac{R_\omega(T-l) \cdot D_\omega(T)}{\sum_T R_\omega(T-l) \cdot R_\omega(T-l) + \varepsilon} \quad (\text{if } l < L) \tag{13a}$$

$$\Delta W_\omega(l) = \theta \cdot \frac{C_0 \cdot D_\omega(T)}{\sum_T C_0 \cdot C_0 + \varepsilon} \quad (\text{if } l = L) \tag{13b}$$

$$W_\omega(l)^T = W_\omega(l)^{T-1} + \Delta W_\omega(l) \tag{14}$$

This on-line form has a weak dependency on the C_0 value.

4. Experiment

A microphone was installed on the visor in a car. The subject speakers were 12 females and 12 males. Each speaker read 13 Japanese sentences for the digit recognition task and for the command recognition task in a car at each of three speeds (stationary, city driving, or highway speed). The total number of utterances was 936 for each test subject over all of the tasks. They were recorded with a sampling frequency of 22 KHz. The cruising noise in the recorded data was almost constant.

The music playing from the in-vehicle loud speakers was recorded separately by a microphone, along with a reference signal. The music was up-tempo popular music with a female vocalist. The in-vehicle loud speakers are stereo, but the music source was monaural in this experiment. The recorded music was mixed with the recorded utterances to generate the test data. The averaged SNRs are shown in Table. 1. The noise power and the signal power were measured by the average in the non-speech and speech periods respectively in the recorded data.

The digit recognition task involves connected digits with no grammar constraints on the length. Therefore, it is sensitive to insertion errors, mostly occurring in the non-speech periods, and this allows measuring the amount of residual echo.

The command recognition task is a set of commands used in a car. As the grammar only allows 1 command per utterance, we do not have to worry about insertion errors. Therefore, this allows measuring the amount of distortion of the speech possibly caused by the echo canceller.

The acoustic model used for this automatic speech recognition was a speaker independent model trained with various cruising noises including idling, city driving and highway driving. The acoustic model was trained using spectral subtraction with the subtraction weight set to 1.0. Since the training data was sampled at 11 KHz, the test data was down-sampled before recognition. In this experiment, we

Table 2: Detailed Word Error Rates for the conventional methods and the proposed methods

Digit Task WER (%)

	Stationary	City Drv.	Highway	Average
Case 1	0.5	0.6	1.1	0.8
Case 2	3.1	14.1	12.1	9.8
Case 3	1.4	2.2	3.6	2.4
Case 4	1.0	2.0	2.6	1.9
Case 5	1.0	1.2	1.5	1.2

Case 1: SS only (reference without music)
Case 2: SS only
Case 3: Echo Canceller + SS

Command Task WER (%)

	Stationary	City Drv.	Highway	Average
Case 1	2.6	1.0	3.5	2.4
Case 2	3.5	11.9	12.5	9.3
Case 3	4.2	1.9	4.8	3.6
Case 4	3.2	2.6	4.2	3.3
Case 5	2.9	1.0	3.2	2.4

Case 4: Proposed Method 1 (L=5, α_1 =1.0 , α_2 =2.0)
Case 5: Proposed Method 2 with Preprocessor (L=5, α_1 =1.0 , α_2 =2.0)

did not use a speech-silence detector for the automatic speech recognition and the complete utterances were decoded in order to measure the front-end performance accurately.

On the other hand, the performance of speech-silence detector is critically important for the front-end processing including echo cancellation, spectral subtraction and the proposed method. In this experiment, we used the oracle speech-silence information for the front-end processing. This was prepared using the data without adding the music. In order to get the most reliable speech-silence information, we installed two additional microphones to do the speech-silence detection based on the coherence between the microphone outputs [9].

Fig. 5 and Fig. 6 show the resulting WERs (Word Error Rates) depending on the various over-subtraction factors γ and the lengths of the adaptive weights L, for the proposed Method 1. This used the on-line formula with the parameters $C_0=10^3$, $\theta=0.1$, and $\varepsilon=10^4$. The WERs are averaged values for the three speeds and the 24 subject speakers. Based on the results, the over-subtraction of the echo improved the recognition accuracy. The optimum factor was around 1.5 to 2.0. Also, introducing a sufficient length of adaptive weights improved the recognition accuracy. In the following experiment, we select γ=2.0 and L=5 as the default setting.

Table 2 shows performance comparisons with the conventional methods. Case 1 is for reference. Music was NOT mixed into the test data. It was processed by conventional spectral subtraction and decoded. Automatic speech recognition performs very well for the stationary cruising noise. Case 2 and the following cases have music mixed into the test data. Case 2 processed the test data only with conventional spectral subtraction. Since there is no echo cancellation, the recognition performance was severely degraded. Case 3 processed the test data by using the conventional combination of echo cancellation and spectral subtraction as shown in Fig. 1. The echo canceller was N-LMS in the time domain with a tap length of 2,048. The recognition performance was much improved from Case 2 as a result of the echo canceller. Case 4 processed the test data using the proposed Method 1 with the parameters γ=2 and L=5. L was selected so to be comparable with the tap length in Case 3. This shows performance superior to Case 3. Case 5 processed the test data by the proposed Method 2 with the parameters γ=2 and L=5. The tap length of the preprocessing echo canceller was 512. The performance is improved in favor of the preprocessing.

5. Conclusions

In order to reduce both background noise and echo effectively for automatic speech recognition in a car, we proposed a new method that adapts echo cancellation and spectral subtraction simultaneously. The stationary noise component is estimated through the adaptation of an echo canceller. As the echo canceller is also implemented using spectral subtraction, the echo component can be further reduced by introducing over-subtraction. We can still use the existing acoustic model trained only with the background noises and spectral subtraction, since we kept the subtraction weight the same as for the stationary noise and introduced over-subtraction only for the echo. The performance can be improved even more by introducing a shot-tap echo cancellation as a preprocessor. In our experiment, this method showed superior recognition accuracy compared to the conventional combination of echo cancellation and spectral subtraction.

6. References

[1] F. Basbug, K. Swaminathan and S. Nandkumar, "Integrated Noise Reduction and Echo Cancellation for IS-136 Systems", *Proc. ICASSP 2000*, Vol. 3, pp.1863-1866, Jul. 2000.

[2] R. Martin and P. Vary, "Combined Acoustic Echo Control and Noise Reduction for Hands-Free Telephony-State of the Art and Perspectives", *Proc. EUSIPCO '96*, pp.1107-1110, 1996.

[3] B. Ayad, G. Faucon and R.L. Bouquin-Jeannes, "Optimization of a Noise Reduction Preprocessing in an Acoustic Echo and Noise Controller", *Proc. ICASSP '96*, Vol.2, pp.953-956, 1996.

[4] P. Dreiseitel and H. Puder, "A Combination of Noise Reduction and Improved Echo Cancellation", *Proc. IWAENC '97*, pp.180-183, 1997.

[5] S. Sakauchi, A. Nakagawa, Y. Haneda and A. Kataoka, "Implementing and Evaluating an Audio Teleconferencing Terminal with Noise and Echo Reduction", *Proc. IWAENC 2003*, pp.191-194, 2003

[6] M. Dahl, I. Claesson and S. Nordebo, "Simultaneous Echo Cancellation and Car Noise Suppression Employing a Microphone Array", *Proc. ICASSP '97,* Vol. 1, pp.239-242, 1997

[7] K. Kobayashi, K. Furuya and A. Kataoka, "A Microphone Array System with Echo Canceller", *Trans. IEICE (in Japanese),* Vol.J87-A, No.2, pp.143-152, 2004

[8] S. Gustafsson, R. Martin and P. Vary, "Combined Acoustic Echo Control and Noise Reduction for Hands-Free Telephony", *Signal Processing* 64, vol.1, pp.21-32, 1998

[9] Agaiby and T.J. Moir, "Knowing the wheat from the weeds in noisy speech", *Proc. EUROSPEECH'97*, Vol.3, pp.1119-1122, 1997