Echo Canceller for Multi-Loudspeakers Based on Maximum Likelihood Using an Acoustic Model

Kentaro Koga*, Tetsuya Takiguchi† and Yasuo Ariki†
* Fujitsu Ten Limited, 1-2-28 Kobe City, Hyogo 652-8510 Japan
E-mail: k-koga@mms.ten.fujitsu.com
† Department of Computer Science and Systems Engineering, Kobe University
1-1 Rokkodai, Nada, Kobe, 657-8501, Japan
E-mail: takigu@kobe-u.ac.jp, ariki@kobe-u.ac.jp

***Abstract*—In this paper, as a key technology for the improvement of a speech recognition system in car environments, we propose a single-microphone-based acoustic echo canceller that selects an optimum cancellation result based on the echo estimation with maximum likelihood using an acoustic model for signals from multi-loudspeakers. The results of experiments conducted on speech superimposed on music show that the proposed canceller can improve the S/N ratio and speech recognition rate, compared to the canceller based on a NLMS algorithm, where the signals from multi-loudspeakers are measured by a single microphone.**

I. Introduction

The current major HMI (Human Machine Interface) system for in-car devices is a touch panel system. But, controlling the touch panel is a distraction for the person driving the car. Taking your eyes off the road while driving can cause traffic accidents. In order to reduce the potential for causing traffic accidents, an audio HMI using a speech recognition system, which enables a driver to operate in-car devices while continuing to look at the road, is necessary.

The car environment contains much noise, which hinders speech recognition by decreasing the S/N ratio of the signal observed at a microphone. In an environment with relatively steady noise, such as road noise, the speech recognition systems have been improved in denoising. But in environments with variable noise, such as music from loudspeakers, current speech recognition systems have difficulty recognizing speech from among the other noises. Therefore, microphone-array-based techniques have been proposed (e.g. [1], [2], and [3]). Array processing can offer the additional advantage of spatial processing, but microphone arrays may not be suitable in a car environment because of their size and cost.

In particular, by using acoustic echo canceller, it is possible to obtain a high-speech recognition rate and improved S/N ratio. The acoustic echoes in a car are output from multi-loudspeakers and measured with a single microphone. The reference signals that are necessary to simulate acoustic echoes are generally composed in two channels or more.

The conventional algorithm of an acoustic echo canceller is achieved utilizing adaptive filter functioning error learning, such as Normalized Least Mean Square (NLMS). In order to cancel acoustic echoes using an acoustic echo canceller in a car environment, the error signals, adaptive filter and reference signals have to be in one channel, because the speech recognition system only uses a one-channel microphone. Since the sounds in a car are reflected in complicated ways by the car interior, which is composed of seats, window glass, and other items, the acoustic echo estimation from a one-channel reference signal cannot obtain adequate improvement of speech recognition because the acoustic echoes in a car are output from multi-loudspeakers and the cancellation result will not converge properly.

In this paper, in a car environment using multi-loudspeakers, we propose an echo canceller based on maximum-likelihood selection among any combination of acoustic transfer characteristics, and describe the resulting improvement in the speech recognition rate obtained with the proposed method.

II. Echo Canceller Model in a Car

A model of the acoustic echo canceller for output from multi-loudspeakers and measured by a single microphone is shown in Fig. 1.

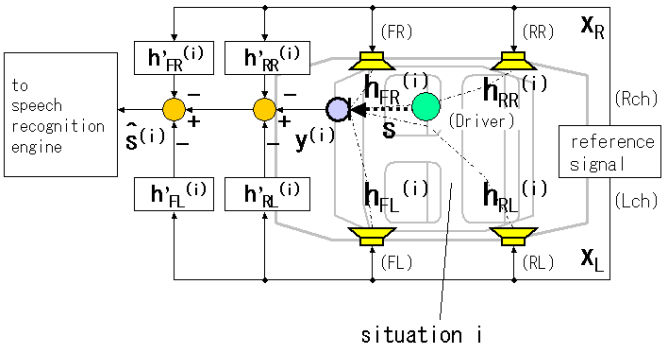

Fig. 1. Configuration of acoustic echo canceller in a car

The observed signal at the microphone $y^{(i)}$ in the car environment (i) is expressed in the time domain as follows:

$$y^{(i)} = s + N^{(i)} \tag{1}$$

where s is the driver's speech. $N^{(i)}$ is an acoustic echo and is expressed as follows:

$$N^{(i)} = \sum x_L (h_{FL}^{(i)} + h_{RL}^{(i)}) + \sum x_R (h_{FR}^{(i)} + h_{RR}^{(i)}) \tag{2}$$

where the case of a two-channel reference signal and four loudspeakers is considered for simplicity. The two-channel reference signals are x_L and x_R, and the transfer characteristics from each loudspeaker to the microphone in the car environment (i) are $h^{(i)}_{FL}$, $h^{(i)}_{RL}$, $h^{(i)}_{FR}$, and $h^{(i)}_{RR}$. The acoustic echo to be estimated with an echo canceller is expressed as follows:

$$N'^{(i)} = \sum x_L(h'^{(i)}_{FL} + h'^{(i)}_{RL}) + \sum x_R(h'^{(i)}_{FR} + h'^{(i)}_{RR}) \quad (3)$$

Therefore, the clean speech signal of the driver $\hat{s}^{(i)}$ is obtained as follows:

$$\hat{s}^{(i)} = y^{(i)} - N'^{(i)} \quad (4)$$

In $\hat{s}^{(i)}$, the estimated acoustic echo $N'^{(i)}$ needs to be optimized in order to minimize the estimated error as the target speech s remains. However, with an adaptive filter (such as the conventional NLMS method), it is impossible to estimate acoustic echoes from multi-loudspeakers, respectively. Therefore, we proposed an acoustic echo canceller for selecting transfer characteristics $h^{(i)}_{FL}$, $h^{(i)}_{RL}$, $h^{(i)}_{FR}$, and $h^{(i)}_{RR}$ to optimize an estimated acoustic echo $N'^{(i)}$ based on maximum likelihood using an acoustic model.

III. Acoustic Echo Canceller Based on Echo Estimation with Maximum Likelihood

Instead of carrying out echo estimation using error learning, we consider a way to select the optimum transfer characteristics to be estimated from a database, which are measured in an actual environment. The procedure we used was as follows:

Step 1 Create acoustic echoes for all assumed transfer characteristics in a car environment to reduce them from observed signals.

Step 2 Select the optimum environment (transfer characteristic) by maximum likelihood estimation, calculating likelihood for the echo-cancelled speech signals using an acoustic model.

After **Step 1**, if the transfer characteristics of a real (test) environment are identical to those of the estimated environment, only clean speech is supposed to exist after cancellation of the acoustic echo. That means we can obtain a high likelihood with an acoustic model for the echo-cancelled signal. On the other hand, if the transfer characteristics of a real (test) environment and those of the estimated environment are mismatched, the clean speech and echo error signal exist in the signal after cancellation. That will give us a lower likelihood with an acoustic model for the echo-canceller-applied signal. Therefore, in **Step 2**, a maximum likelihood criterion is used to estimate the optimal transfer characteristic using echo-canceller-applied signals, where acoustic echoes are created using a database of transfer characteristics.

A. Database of transfer characteristics

In order to make a database of transfer characteristics, we measure impulse resonses in a car. As we can suppose that there will be a variety of situations due to the presence of passengers and articles in a car, we should establish various situations and calculate each transfer characteristic.

In this research, we establish 12 different situations in the passenger locations as shown in Fig. 2. We assume that the car will hold five passengers.

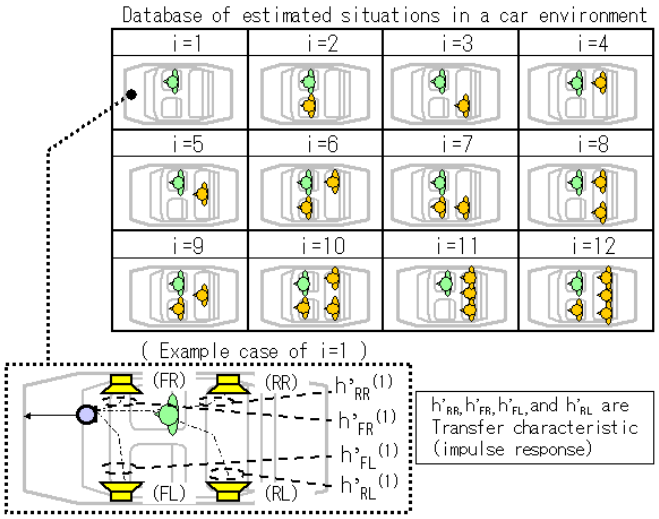

Fig. 2. 12 types of transfer characteristics

In this paper, we do not consider the situations where a driver is absent or 2 passengers in the back seat at the one side because these make no sense as speech-activated situations in a car.

B. Calculation of speech likelihood

A GMM (Gaussian Mixture Model) is prepared in advance as an acoustic model to be referred to when calculating speech likelihood after cancellation, where the MFCC feature (Mel Frequency Cepstrum Coefficient) of several speakers' speech data is calculated to train GMM.

MFCC is a technique used to calculate speech features using discrete cosine transformation of the logarithmic value of the power component through FFT.

We then calculate a GMM (Gaussian Mixture Model) from the obtained acoustic feature (MFCC) of each speaker. Here, the speaker's MFCC is o, and the speech likelihood $P(o)$ is expressed as the sum of weighted normal distributions as Eq. (5). Those parameters are estimated by EM algorithm [4].

$$P(o) = \sum_{w=1}^{W} \lambda_w N(o; \mu_w, \sigma_w) \quad (5)$$

We calculate the MFCC of the cancellation result s and then calculate the speech likelihood using the GMM in order to select the optimal transfer characteristic.

C. Acoustic echo canceller based on echo estimation with maximum likelihood

Fig. 3 shows the configuration of an acoustic canceller using maximum likelihood in a car environment.

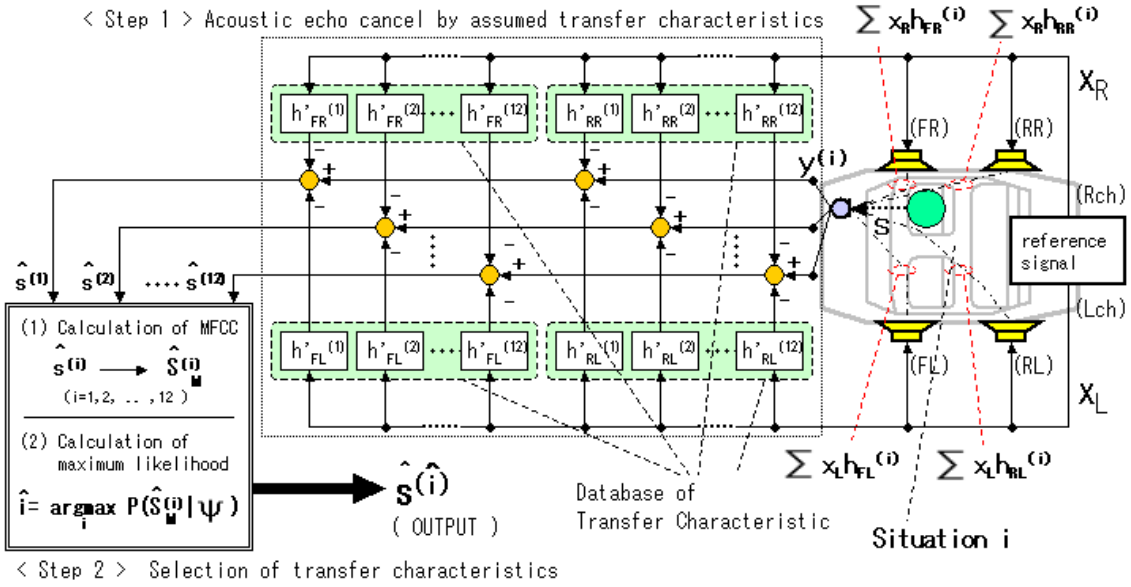

Fig. 3. Configuration of an acoustic canceller using maximum likelihood in a car environment

In Fig. 3, an observed signal $y^{(i)}$ in a car environment (i) is expressed as follows:

$$y^{(i)} = s + \sum x_L h_{FL}^{(i)} + \sum x_L h_{RL}^{(i)} + \sum x_R h_{FR}^{(i)} + \sum x_R h_{RR}^{(i)} \quad (= s + N'^{(i)}) \tag{6}$$

The acoustic echoes $N'^{(1)}, N'^{(2)}, \ldots, N'^{(12)}$ are calculated for twelve different situations as shown in Fig. 2. Then the clean signals are obtained by subtraction of the acoustic echoes from the observed signal according to Eq. (4).

$$\hat{s}^{(1)}, \hat{s}^{(2)}, \ldots, \hat{s}^{(12)} \tag{7}$$

Next, the MFCCs $\hat{S}_M^{(1)}, \hat{S}_M^{(2)}, \ldots, \hat{S}_M^{(12)}$ are calculated. The selection of the acoustic echo is handled in a maximum-likelihood framework using the acoustic model (GMM).

When a set of GMM is represented by $\psi = \{\lambda, \mu, \sigma\}$, the optimal combination $\hat{i}$ is calculated as follows:

$$\hat{i} = \arg\max_i P(\hat{S}_M^{(i)}|\psi) \tag{8}$$

This cancellation result $\hat{s}^{(\hat{i})}$ shows the maximum speech likelihood, in other words, it shows that the acoustic echoes are cancelled to the highest level.

IV. Experiment

Conducting echo cancellation using the proposed method on a speech signal superimposed on music recorded under an actual environment, we are going to show that the proposed method can improve the S/N ratio and speech recognition rate better than NLMS can.

We conducted the experiment using speech signals superimposed on music $y^{(i)}(i = 1, 2, \ldots, 12)$ recorded under a real environment with 12 different situations (the passenger locations shown in Fig. 2). Table I shows the conditions for the evaluation data and Table II shows the conditions for the algorithms.

The speech signals superimposed on music $y^{(o)}$ measured under the environment with passenger locations (o) should be cancelled by using transfer characteristics $h'^{(o)}$ measured with the same passenger locations (o), and the results $\hat{s}^{(o)}$ based on echo estimation with maximum likelihood should be selected. We define the rate of $\hat{s}^{(o)}$ to be selected to $y^{(o)}$ as a selection rate of proper transfer characteristics.

As shown in Fig. 4, the selection rate (output $\hat{s}^{(o)}$ to input $y^{(o)}$) of proper transfer characteristics based on echo estimation with maximum likelihood averages out to 79.8% for speakers' average. Thus, the result shows that the selected transfer characteristics are not 100% for all speakers. However, as shown in Fig. 5 and Fig. 6, the cancellation effect using an acoustic echo canceller based on echo estimation with maximum likelihood was improved for the S/N ratio (9.5 dB higher), and the speech recognition rate was 21.4% higher than that for NLMS, where NLMS applies the one-channel reference signals, that used to be in two channels, as input values for an adaptive filter.

The "ideal" cases in Fig. 5 and Fig. 6 represent what would happen if the selection rate of proper transfer characteristics is 100%. Since the selection rate of proper transfer characteristics does not reach 100%, the S/N ratio decreases by 0.1 dB and the recognition rate decreases by 4.5% compared to the ideal result.

In Fig. 7, the speech recognition rate versus the selection rate is plotted. As shown in Fig. 7, as the selection rate improves, the speech recognition rate also improves.

TABLE I
CONDITIONS FOR EVALUATION DATA

Number of speakers	5
Number of sentences	100 sentences
Sampling frequency	16 kHz
Car model used for measuring impulse responces	will CYPHA

TABLE II
CONDITIONS FOR ALGORITHMS

Tap length of filter	1,200
Number of sentences in GMM training	1,200 sentences
Number of mixtures in GMM	32
Dimensions of MFCC	16
Frame width for MFCC characteristics selection	32 ms
Sift width for MFCC characteristics selection	8 ms

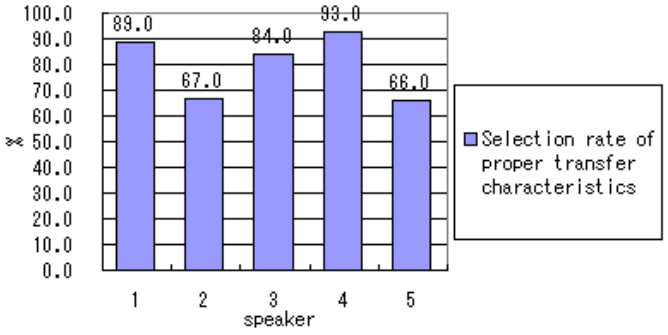

Fig. 4. Selection rate of proper transfer characteristics

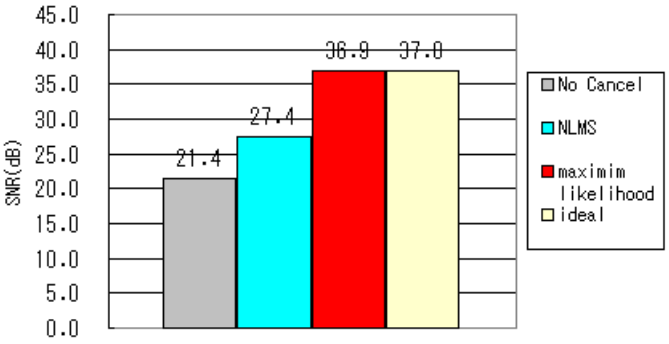

Fig. 5. S/N ratio

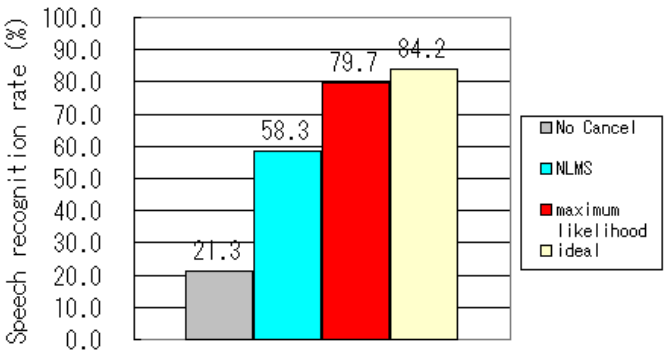

Fig. 6. Speech recognition rate

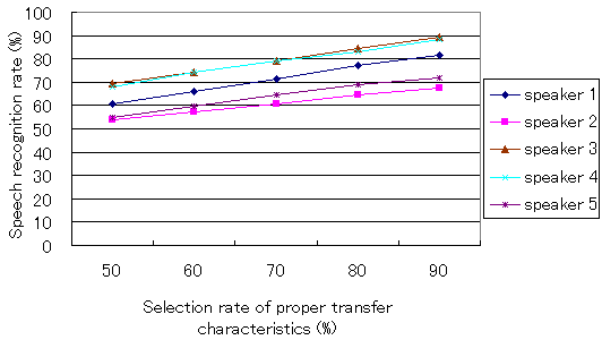

Fig. 7. Relation between speech recognition rate and selection rate of proper transfer characteristics

V. CONCLUSION

In this paper, we have proposed an acoustic echo canceller for selecting the optinum cancellation effect using the maximum likelihood. Moreover, through our experiments, we have confirmed SN improvement in observed signals and improvement in the speech recognition rate compared to NLMS.

In future work, we will need to apply our algorithm to more variable car environments.

REFERENCES

[1] A. Nakagawa, S. Shimauchi, Y. Haneda, S. Aoki, and S. Makino, "Channel-number-compressed multi-channel acoustic echo canceller forhigh-presence teleconferencing system with large display," Proc. of ICASSP, pp. 813-816, 2000.

[2] S. Miyabe, Y. Hinamoto, H. Saruwatari, K. Shikano, and Y. Tatekura, "Interface for barge-in free spoken dialogue system based on sound field reproduction and microphone array," EURASIP Journal on Applied Signal Processing archive Volume 2007 , Issue 1, Article ID 57470, 13 pages, 2007.

[3] H. Buchner and W. Kellermann, "A fundamental relation between blind and supervised adaptive filtering illustrated for blind source separation and acoustic echo cancellation," Joint Workshop on Hands-Free Speech Communication and Microphone Arrays (HSCMA), pp. 17-20, 2008.

[4] X. D. Huang, Y. Ariki and M. A. Jack, "Hidden Markov Models for Speech Recognition," Edinburgh University Press, Sept., 1990.