Improvement of In-Car Speech Recognition by Acoustic Echo Canceller with Maximum Likelihood

Kentaro Koga †, Shinji Fukuda †, Tetsuya Takiguchi ‡, Yasuo Ariki ‡
† Fujitsu Ten Limited, 1-2-28 Kobe City, Hyogo 652-8510 Japan, +81-78-682-0534
‡ Research Center for Urban Safety and Security Kobe University
E-mail:k-koga@mms.ten.fujitsu.com

Abstract

In this paper, as a key technology for improvement of speech recognition system in car environments, we propose an acoustic echo canceller for selecting an optimum cancellation result based on the echo estimation with maximum likelihood using transfer characteristics measured prospectively.

The results of experiments conducted to speech superimposed on music show that the proposed canceller can improve S/N ratio and speech recognition rate, compared to the canceller based on the algorithm of NLMS.

1 Introduction

The current major system of HMI for in-car devices is a touch panel system. But, controlling the touch panel leads a driver to inattentive driving. Looking away in driving can cause traffic accidents. In order to reduce the cause of traffic accidents, it is required to have audio HMI using speech recognition system, which enables a driver to operate in-car devices without looking away from the forward.

The environment in a car is full of much noise, disturbing speech recognition with decreasing S/N ratio of signal observed at a microphone. Under the environment with relatively steady noise such as road noise, the speech recognition is improving in denoising. But under the environment with not-steady noise such as music from loudspeakers, the speech recognition is not improving with difficulty in denoising.

In particular, in order to obtain the high recognition rate, S/N ratio is to be improved by using acoustic echo canceller. The acoustic echoes in a car are output from multi-loudspeakers and measured by 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 acoustic echo canceller is realized with the adaptive filter functioning error learning such as Normalized Least Mean Square (NLMS) [1]. In order to cancel acoustic echoes using acoustic echo canceller in a car environment, error signals, adaptive filter and reference signals have to be in one channel, because the speech recognition uses one channel microphone only. Since the sounds in a car are reflected complicatedly by the interior composed of seats, window glasses and others, the acoustic echo estimation by reference signals put into one channel cannot have enough improvement of speech recognition, due to the cancellation result not converged properly.

In this paper, under the environment in a car using multi-loudspeakers, taking the method by selection with compiling a database not by estimating echoes, we consider an echo canceller using two channel reference signals respectively and show the improvement effect in speech recognition rate with the proposed method.

2 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.

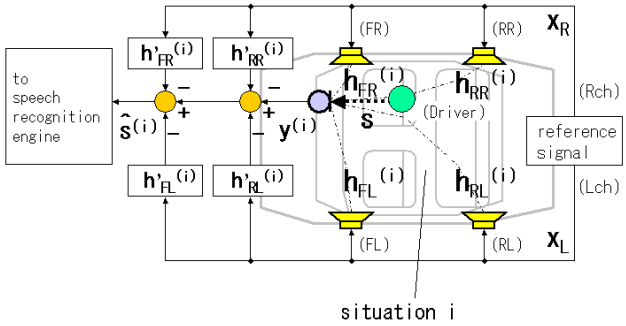

Fig. 1 Configuration of acoustic echo canceller in a car

The observed signal at the microphone $y^{(i)}$ is expressed as follows:

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

where $N^{(i)}$ is an acoustic echo and expressed as follows:

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

where 2 channel reference signals are x_L, x_R, transfer characteristics from each loudspeaker to microphone in the car environment (i) are $h_{FL}^{(i)}$, $h_{RL}^{(i)}$, $h_{FR}^{(i)}$, and $h_{RR}^{(i)}$, and driver's speech is s. The acoustic echo to be estimated with an echo canceller is expressed as follows:

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

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

$$\hat{s}^{(i)} = y^{(i)} - N'^{(i)} \tag{1}$$

In $\hat{s}^{(i)}$, the estimated acoustic echo $N'^{(i)}$ is requested to be optimized in order to minimize the estimated error as the target speech s remains.

However, with the adaptive filter such as NLMS as a conventional method, it is impossible to estimate acoustic echoes from multi-loudspeakers respectively. Since the model in Fig. 1 has only one estimated error, when applying error learning with an adaptive filter, the whole environment in a car shall be regarded as one system. Since the reference signals are in two channels, they need to be put into one channel. But the estimation by deeming multiple acoustic echoes as one influences the optimization of $N'^{(i)}$ negatively.

Therefore we proposed an acoustic echo canceler for selecting transfer characteristics $h_{FL}^{(i)}$, $h_{RL}^{(i)}$, $h_{FR}^{(i)}$, and $h_{RR}^{(i)}$ to optimize an estimated acoustic echo $N'^{(i)}$.

3 Acoustic echo canceller based on echo estimation with maximum likelihood

Instead of the echo estimation by error learning, we consider the idea to select the optimum transfer characteristics to be estimated from the database, which are measured in an actual environment. Here are the procedures:

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

Step.2 We have sought to select the optimum estimated environment by maximum likelihood estimation calculating likelihood using acoustic models after cancellation.

After **Step.1**, if the transfer characteristics of real environment are identical to those of the estimated environment, only clean speech is supposed to exist after cancellation of acoustic echo. However, if the transfer characteristics of real environment and those of the estimated environment have mismatch, the clean speech and echo error signal exist in the signal after cancellation.

The above **Step.2** is to be done for selecting clean speech.

3.1 Database of transfer characteristics

The transfer characteristics are calculated by means of impulse response measurement in a car [2]. As we can suppose various situations by allocation of passengers and articles in a car, we should establish various situations and calculate each transfer characteristic.

In this paper, we establish 12 different situations in the passenger locations as shown in Fig. 2. We assume that the passenger capacity is five.

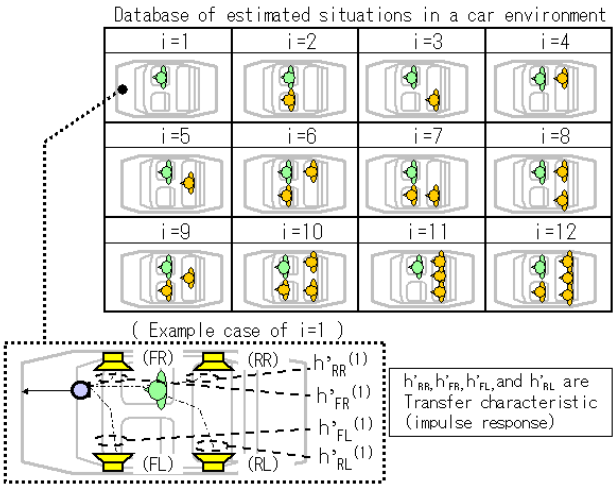

Fig. 2 12-way transfer characteristics

The part with a length T within the transfer characteristics measured and calculated under the environment (i) is prescribed as ${h'_{FL}}^{(i)}, {h'_{FR}}^{(i)}, {h'_{RL}}^{(i)}, {h'_{RR}}^{(i)}$ $(i = 1, 2, \ldots, 12)$.

In this paper, we do not consider the situations of a driver being absence and 2 passengers in the rear seat at the one side because these make no sense as speech-activated situations in a car.

3.2 Calculation of speech likelihood

GMM (Gaussian Mixture Model) is prepared in advance as an acoustic model to be referred when calculating speech likelihood after cancellation.

Then MFCC feature (Mel Frequency Cepstrum Coefficient) is calculated using prepared several speakers' speech data (for learning).

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

We then calculate GMM (Gaussian Mixture Model) from the obtained acoustic feature (MFCC) of each speaker. Here speaker's MFCC is o, the speech likelihood $P(o)$ is expressed as the sum of weighted normal distributions as Eq. (2). Here the number of distribution is W. The mean vector of element wth in the normal distributions W is μ_w and the variation is σ_w. λ_w is a weighting coefficient expressed as $\sum_1^W \lambda_w = 1$. Those parameters are estimated by EM algorithm [3].

$$P(o) = \sum_{w=1}^{W} \lambda_w N(o; \mu_w, \sigma_w) \tag{2}$$

We calculate MFCC S of the cancellation result s and then calculate the speech likelihood using the GMM.

3.3 Acoustic echo canceller based on echo estimation with maximum likelihood

Fig. 3 shows a 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:

$$y^{(i)} = s + \sum x_L h_{FL}^{(i)} + \sum x_L h_{RL}^{(i)} + \sum x_R h_{FR}^{(i)} + \sum x_R h_{RR}^{(i)} \quad (= s + N'^{(i)}) \tag{3}$$

The acoustic echo $N'^{(1)}, N'^{(2)}, \ldots, N'^{(12)}$ are prepared from twelve combinations of transfer characteristics by using $y^{(i)}$ as estimating every environment. Then the signals subtracted the acoustic echoes from the observed signal are calculated according to Eq. (1):

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

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

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

$$\hat{i} = \arg\max_i P(\hat{S}_M^{(i)} | \psi) \tag{4}$$

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.

4 Experiment

In [4], conducting echo cancellation by this proposal method to simulated speech signals superimposed on music, we showed that the proposal method has a prospect to improve S/N ratio better than NLMS. In this paper, conducting echo cancellation by this proposal method to the speech signal superimposed on music recorded under an actual environment, we are going to show that the proposal method can improve S/N ratio and speech recognition rate better than NLMS.

We conducted experiment using speech signals superimposed on music: $y^{(i)}(i = 1, 2, \ldots, 12)$, measured under the actual environment with 12 different situations in the passenger locations shown in Fig. 2.

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 $\hat{s}^{(o)}$ to be selected to $y^{(o)}$ as a selection rate of proper transfer characteristics.

Fig. 4, Fig. 5, and Fig. 6 respectively show the selection rate of proper transfer characteristics, S/N ratio and recognition rate, when using cancellers by NLMS and by the proposal method: echo estimation with maximum likelihood. For references, the following four values are also added: S/N ratio average and recognition rate of the original measured signal $y^{(i)}$, and ideal values (when the output $\hat{s}^{(o)}$ is obtained to all $y^{(o)}(o = 1, 2, \ldots, 12)$; the selection rate of proper transfer characteristics is 100%) of the proposal method. NLMS applies the one-channel reference signals that used to be in two channels, as input values for an adaptive filter. Table. 1 shows the conditions for evaluation data and Table. 2 shows the conditions for algorithms.

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 is 79.8 (%) in speakers' average, and thus, the result shows that the selected transfer characteristics are not proper in 100% for all speakers. However, as shown in Fig.5 and Fig.6, the cancellation effects by acoustic echo canceller based on echo estimation with maximum likelihood are improved in S/N ratio by 9.5(dB) and in speech recognition rate by 21.4(%), compared to NLMS.

Since the selection rate of proper transfer characteristics does not reach 100%, S/N ratio is short by 0.1(dB) and the recognition rate is short by 4.5(%) compared to the ideal result.

Table 1 Conditions for evaluation data

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

Table 2 Conditions for algorithms

Tap length of filter T	1200
Number of sentences in GMM training	1200 sentences
Number of mixture in GMM	32
Dimensions of MFCC	16
Frame width for MFCC characteristics selection	32ms
Sift width for MFCC characteristics selection	8ms

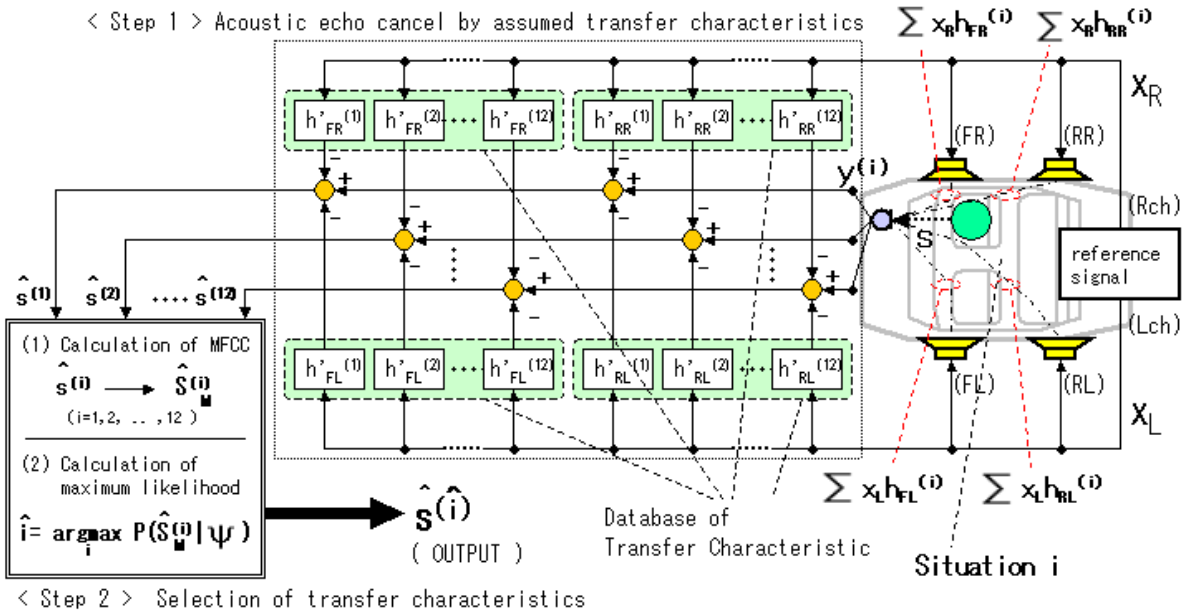

Fig. 3 Configuration of an acoustic canceller using maximum likelihood in a car environment

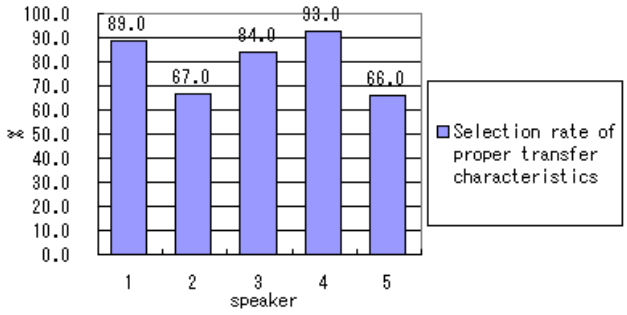

Fig. 4 Selection rate of proper transfer characteristics

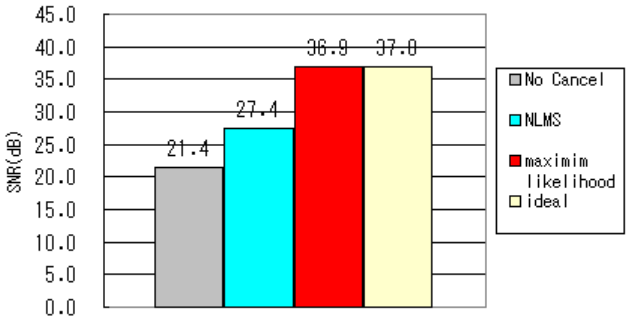

Fig. 5 S/N ratio

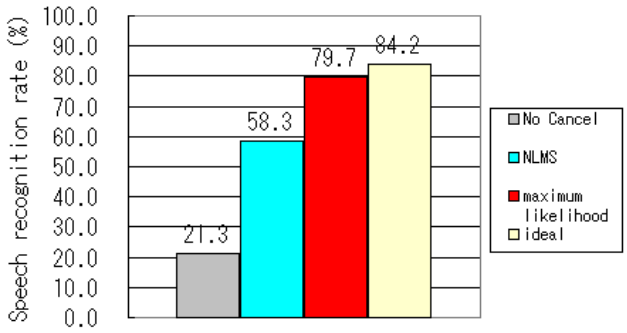

Fig. 6 Speech recognition rate

5 Conclusion

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

References

[1] Ohga and others, “Acoustic System and Digital Processing” Institute of Electronics, Information and Communication Engineers 1995.

[2] Satoh, pp669-676, Vol. 58, No.10, Journal of Acoustical Society of Japan 2002.

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

[4] Koga and others, Acoustical Science and Technology Spring, 2008 3-P-6, pp.849-850, 2008-03.