Speech Recognition under Noisy Environments using Multiple Microphones Based on Asynchronous and Intermittent Measurements

Kohei Machida* and Akinori Ito*
*Graduate School of Engineering, Tohoku University, Sendai, Japan, 980-8579 JAPAN
E-mail: {machida, aito}@spcom.ecei.tohoku.ac.jp Tel: +81-22-795-7084

***Abstract*—We propose a robust speech recognition method under noisy environments using multiple microphones based on asynchronous and intermittent observation. In asynchronous and intermittent observation, the noise spectrum is estimated by the environmental noise observed in fragments from multiple microphones, and spectral subtraction is performed by this estimated noise spectrum. In this paper, we consider the case of estimating the noise spectrum from the noise observed by another microphone just before speech input. However, the noise spectrum needs to be compensated because of the difference in the location of the microphone in this case. Then, we examined compensating the noise spectrum by using the estimated LSFL on the log spectrum. By compensating the noise spectrum, the recognition rate improved compared with the case without compensation.**

I. Introduction

Speech recognition is expected to become more and more popular because of its easiness to use. However, we have to deal with the problem of the environmental noise in order to use speech recognition in a real environment [1]. Without a proper noise-robust speech recognition algorithm, speech recognition performance degrades remarkably even when speech recognition is used indoor because of indoor noise, such as the sound of an air conditioner or a television.

Methods for noise robustness that use multiple microphones such as microphone array have been proposed so far, and have shown high efficiency [2]. The microphone array can use spatial information such as the phase difference to estimate the direction of the sound source and to emphasize sound from a particular direction. However, the methods using multiple microphones to date, including the microphone array, assume synchronous and continuous observation of the signal.

Synchronous observation assumes that sampling of all microphones are performed synchronously. To realize the synchronous observation of many microphones, we need an expensive A/D converter. Continuous observation assumes that all microphones always continue observing sounds. Considering the speech recognition using embedded device, the consumption of energy will be a problem when the continuous observation is assumed. To save the energy consumption, a voice-operated switch (VOX) is commonly used. When no speech is observed, the VOX turns off the speech recognition logic and the recognizer becomes in a sleeping state. The problem of using VOX is that it is impossible to observe the non-speech signal just before the speech signal, which is needed for noise-reduction method as a sample of environmental noise.

Therefore, in this paper, we propose a speech recognition method under noisy environments using multiple microphones based on asynchronous and intermittent observation, which do not require synchronous and continuous observation.

II. Asynchronous and Intermittent Measurements

Generally, by estimating the environmental noise superimposed on the speech, speech recognition accuracy under noisy environments can be improved using techniques such as the spectral subtraction [3] or HMM composition [4], [5].

In our study, energy consumption and cost of equipment in the conventional synchronous and continuous observation are problems. Therefore, method of performing recognition by observing the noise and sound in the environment with less energy by using an inexpensive equipment is asynchronous and intermittent observation. We assume that microphones are arranged in multiple places in the room and observe the environmental noise independently and intermittently. And then, the environmental noise is estimated using the signals observed those microphones.

We assume the following assumptions for asynchronous and intermittent observation.

- Two or more microphones are used for observation, all of which are distributed in the room.
- On each microphone, the sound is sampled independently (asynchronous observation). We also assume that the rough sampling time of the signal observed in each microphone can be known.
- Each microphone observes the environmental noise intermittently (intermittent observation), for example one second in every ten seconds. The microphone that is not observing the sound is in the power-saving state.
- The positions of all microphones and sources of the noises are almost fixed.
- The voice-operated switch is used to record speech, which means that only the speech part (without preceding and following silence parts) is extracted and recognized.

Since each microphone is asynchronous, we cannot calculate the phase difference of the observed signal and thus methods for microphone array cannot be used. In recent years, there have been a couple of researches of microphone array using asynchronous multiple microphones [6], [7]. By estimating and correcting the difference in sampling time between microphones, the estimation of microphone position and sound source position can be achieved. Thus, the same processing as the microphone array is possible in asynchronous. However, it is still difficult to synchronize the signals observed by such microphones in real-time. Furthermore, because continuous observation of each microphone is required in these asynchronous microphone array systems, these systems are not suitable for intermittent observation. Therefore, we do not consider the synchronization of the microphones in our study.

The intermittent observation assumes that each microphone observes the environmental noise at times. Fig. 1 shows the concept of the intermittent observation. A noise spectrum is calculated from the noise signal observed in fragments with multiple microphones (the red portion of Fig. 1).

In addition, the position of each microphone and noise source are assumed to be fixed, which enables to estimate the relationship between the signals observed by different microphones.

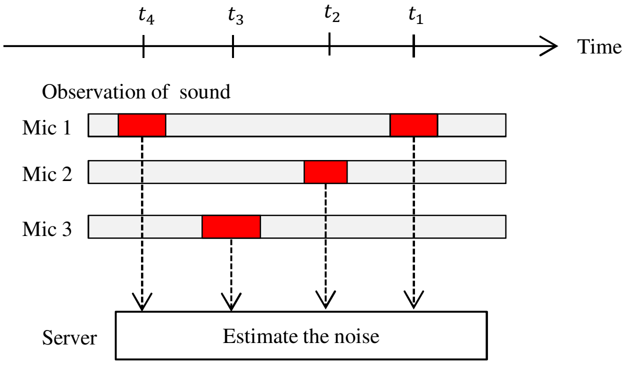


Fig. 1. Concept of asynchronous and intermittent observation

In our study, it is supposed to perform noise-robust speech recognition according to the following procedure. Firstly, the environmental noise observed intermittently by multiple microphones in the room are collected to the server. Secondly, noise in that environment is sequentially estimated by noise collected in the server. Finally, when recognizing the speech, noise reduction by spectral subtraction is performed by estimated noise spectrum.

A. Spectral Subtraction

Spectral subtraction (SS) is a method of noise reduction by subtracting the spectrum of noise from the spectrum of speech with noise in a power spectrum domain. It is a simple algorithm, which is widely used because of its high noise reduction effect. Many improved algorithms of SS have also been proposed [8].

Since the speech and noise are observed together, it is impossible to calculate the true spectrum of the noise superimposed on the speech. Thus, generally under the assumption that the noise is stationary within a short time, the noise spectrum is calculated from the section just before the speech, which contains only the noise signal. Then the observed noise spectrum is subtracted from the observed spectrum for estimating the spectrum of the clean speech.

Let the spectrum of the observed signal at i-th frame be $|X_i(\omega)|^2$ and the estimated noise spectrum be $|\hat{N}_i(\omega)|^2$. Then the power spectrum of the estimated speech signal is calculated as (1).

$$|\hat{S}_i(\omega)|^2 = \begin{cases} |X_i(\omega)|^2 - \alpha|\hat{N}(\omega)|^2 \\ \quad (\text{if } |X_i(\omega)|^2 > \alpha|\hat{N}(\omega)|^2) \\ \beta|X_i(\omega)| \\ \quad (\text{otherwise}) \end{cases} \tag{1}$$

Here, the coefficient α is the subtraction coefficient. It is known that the noise reduction performance increases by overestimating the noise spectrum Moreover, the coefficient $\beta \ll 1$ is the flooring coefficient. When the value has become negative after subtracting the noise spectrum, β is multiplied by the original spectrum.

III. Compensation of Spectrum

If the noise is observed just before the speech inputs, it is desirable to estimate the noise spectrum from that noise. However, for we assume voice-operated switch, we cannot observe the noise signal just before the speech. Therefore, we consider a method of estimating the noise spectrum from the noise captured by other microphone and temporally near to the speech. Fig. 2 shows the assumed situation. In this figure, we observe speech using one microphone and the noise is estimated from the signal of the other microphone.

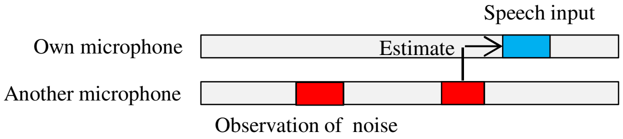


Fig. 2. Situation of estimation by another microphone

When estimating the noise spectrum from the noise observed by another microphone, it is not possible to calculate the noise spectrum as it is. Because characteristics of the observed noise are changed by the difference in the location of the microphone. Therefore, the noise spectrum needs to be compensated. If observations at all the microphones were synchronized, it would be able to calculate the transfer function between two microphones and the spectrum could be compensated exactly. In this study, however, it is not possible to calculate the exact transfer function because of the assumption of asynchrony. Thus, we approximated the log spectrum of the noise using the least squares fitting line (LSFL), and used the estimated LSFL for compensating the noise spectrum. Here, we assume that we can obtain power spectra of the noise signal observed by multiple microphones at the almost same time beforehand (it is not necessarily just before the speech input).

As shown in Fig. 3, we calculate the linear regression line by the least squares method on the log power spectra of the noise signals observed by all microphones.

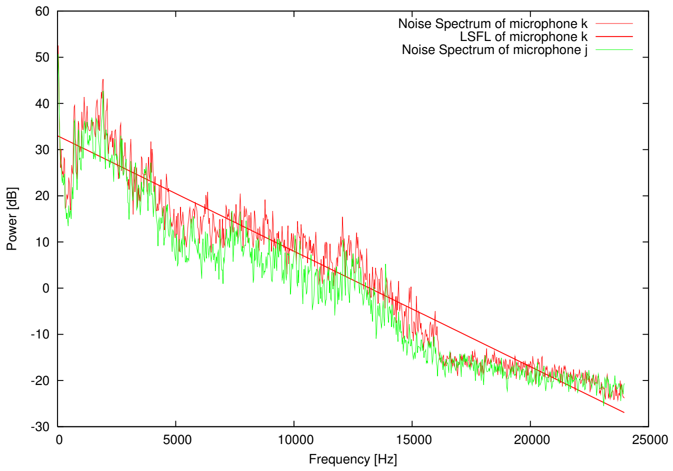


Fig. 3. Compensation by LSFL

Let the noise spectrum observed by microphone j be $|N_i(\omega|j)|^2$, and the linear regression line calculated from the log power spectrum of noise observed by microphone k be $a_k\omega + b_k$. Then the estimated power spectrum of microphone k estimated from microphone j is calculated as (2).

$$\log|N_i(\omega|j{\to}k)|^2 = \frac{a_k}{a_j}(\log|N_i(\omega|j)|^2 - b_j) + b_k \qquad (2)$$

This is equivalent to transforming the noise spectrum calculated by microphone j onto the regression line of microphone k. This processing is performed frame by frame.

IV. Experiment

We investigated whether the proposed compensation method of the spectrum was effective. In the experiment, we performed isolated word recognition using the speech recognition engine Julius [9], and calculated the recognition rate.

A. Experimental Paradigm

The vocabulary of the isolated word recognition was the 30 words of names of electronic appliances. We prepared 20 sets of this 30 words (thus 600 words in total), spoken by one male. The speech was recorded in a soundproof chamber.

We also prepared three kinds of environmental noises: the air conditioner, the microwave oven, and the faucet. Those noises were recorded individually in the experiment. The arrangement of the microphones and the source of noise is shown in Fig. 4. The distance between Mic1 and the source of the noise was about 1.5 m, and the distance between two microphones is 5.0 m. In this experiment, the noise added to the isolated word is recorded with Mic1 of Fig. 4. It is assumed that the target speech is inputted into Mic1. The waveform of each noise are shown in Fig. 5. The noise of microwave oven is not stable as Fig. 5. When each noise was added, SNR was -6.4dB, -10.5dB and -2.8dB, respectively.

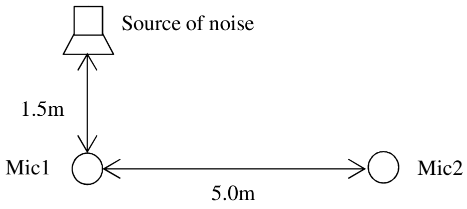
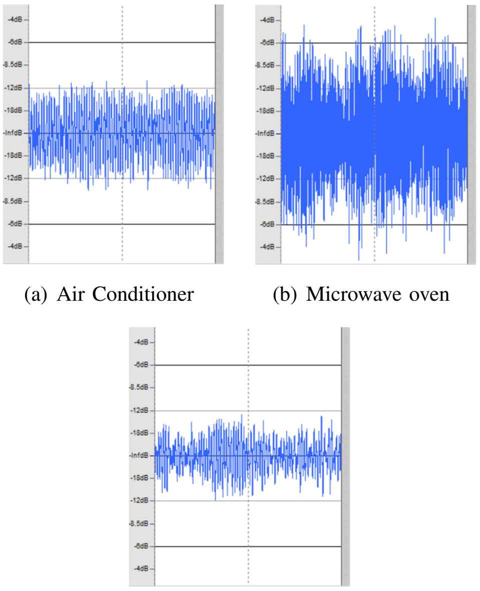


Fig. 4. Recording condition



(a) Air Conditioner

(b) Microwave oven

(c) Faucet

Fig. 5. Waveform of each noise

The section used for estimation of noise spectrum was 0.3 seconds just before the isolated word speech. In this experiment, we compared the following three conditions for noise estimation.

- Estimate the noise spectrum from own microphone Mic1. (Condition 1)
- Estimate the noise spectrum from another microphone Mic2 without compensating the spectrum. (Condition 2)
- Estimate the noise spectrum from another microphone Mic2 with compensating the spectrum. (Condition 3)

Using the noise spectrum estimated under these conditions, we performed spectral subtraction to each isolated word and calculated the recognition rate. If the noise is stationary, the noise spectrum estimated by Condition 1 is considered to be the best. In this experiment, we confirm whether the recognition rate improves from Condition 2 to Condition 3 by

TABLE II
Experimental result

Added noise	Before SS (%)	After SS (%)		
		Condition 1	Condition 2	Condition 3
Air Conditioner	75.00	93.17	91.17	92.67
Microwave oven	28.17	42.96	44.83	43.17
Faucet	96.50	99.17	98.33	99.17

compensating a spectrum and how close the recognition rate of Condition 3 is to the recognition rate of Condition 1.

The other experimental conditions are shown in TABLE I. The subtraction coefficient alpha and the flooring coefficient beta were fixed to the value which had showed the highest performance by the preliminary experiment. The acoustic model is gender independent PTM(Phonetic Tied-Mixture model) trained from the ASJ-JNAS database [10].

TABLE I
Experimental conditions

Subtraction coefficient α	2.0
Flooring coefficient β	0.5
Sampling frequency	16 kHz
Frame width	40 ms
Frame shift	20 ms
Window function	Sine window
Acoustic model	PTM

B. Experimental Result

The experimental result is shown in TABLE II. TABLE II shows the recognition rate before performing Spectral Subtraction (Before SS) and the recognition rate after Spectral Subtraction (After SS) against each added noise. In addition, when we performed isolated word recognition without noise, the recognition rate was 100%.

Comparing Condition 2 with Condition 3, the recognition rate was improved in Condition 3 for the air conditioner noise and the faucet noise. We obtained 0.5 point improvement for the air conditioner noise, and the recognition rate for the faucet noise was same as that under Condition 1. The effect of spectrum compensation was confirmed in the case where these noises were added.

However, when adding the microwave oven noise, the recognition rate of Condition 3 was lower than that of Condition 2. Considering the fact that the recognition rate of Condition 2 was higher than that of Condition 1 in this case, the microwave oven noise was not stable and the noise signal observed by Microphone 2 was similar to the noise superimposed to the speech recorded by Microphone 1 in this experiment. Therefore, it is considered that the recognition rate decreased by the spectral compensation as the result of Condition 3.

V. Conclusions

In our study, we proposed the speech recognition under noisy environments using multiple microphones based on asynchronous and intermittent observation. Each microphone observes environmental noise independently and intermittently.

In this paper, we consider the case of estimating the noise spectrum from the noise observed by another microphone just before speech input. In order to estimate the noise spectrum from another microphone, it is necessary to compensate the spectrum. So, we proposed the compensation method using the estimated LSFL of the log power spectrum as the method of the spectrum compensation in asynchronous observation. Furthermore, we confirmed the effect of this spectrum compensation method by the experiment. In the experiment, the recognition rate using the spectral subtraction and the compensation improved for the stationary noise, compared with the case without compensation. However, we could not improve the accuracy when the noise was not stationary. We need further improvement of noise estimation for nonstationary noise.

As a future work, we are going to study the method of estimating the present noise spectrum from the past noise signal. And we consider the observation interval of the noise.

Acknowledgment

Part of this work was supported by NEC Communication Systems, Ltd.

References

[1] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Communication,* vol. 16, no. 3, pp. 261-291, 1995.

[2] M. Mizumachi and M. Akagi, "Noise reduction by paired-microphones using spectral subtraction," *Proc. Int. Conf. Acoustics, Speech and Signal Processing,* vol. 2, pp. 1001-1004, 1998.

[3] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtracion," *IEEE Trans. Acoust. Speech Signal Process,* vol. ASSP-27, no. 2, pp. 113-120, 1979.

[4] F. Martin, K. Shikano and Y. Minami, "Recognition of noisy speech by composition of speech and noise," *IEEE Trans. Speech & Audion Process,* vol. 4, pp. 352-359, 1996.

[5] M. J. F. Gales and S. J. Young, "Robust continuous speech recognition using parallel model combination," *Proc. Int. Conf. Acoustics, Speech and Signal Processing,* vol. 2, pp. 1001-1004, 1998.

[6] K. Hasegawa, N. Ono, S. Miyabe and S. Sagayama "Blind Estimation of Locations and Time Offsets for Distributed Recording Devices," *LNCS,* vol. 6365/2010, pp. 57-64, DOI: 10.1007/978-3-642-15995-4_8, 2010.

[7] H.Miura, T.Yoshida, K.Nakamura, K.Nakadai, "SLAM-based Online Calibration of Asynchronous Microphone Array for Robot Audition," IEEE/RSJ IROS-2011, pp.524-529.

[8] N. Kitaoka, I. Akahori and S. Hasegawa, "Speech Recognition Under Noisy Environments Using Spectral Subtraction With Smoothing Of Time Direction And Real-Time Cepstral Mean Normalization," *Proc. Int. Workshop on Hands-free Speech Communication (HSC2001),* pp. 159-162, 2001.

[9] A. Lee, T. Kawahara and K. Shikano, "Julius — An Open Source Real-Time Large Vocabulary Recognition Engine," *Proc. Eurospeech2001,* pp. 1691-1694, 2001.

[10] K. Ito, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano and S. Itahashi, "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," *Journal of Acoustic Society Japan (E),* vol. 20, No. 3, pp. 199-206, 1999.