IN-CAR SPEECH RECOGNITION USING DISTRIBUTED MICROPHONES - ADAPTING TO AUTOMATICALLY DETECTED DRIVING CONDITIONS -

Hideki Banno, Tetsuya Shinde, Kazuya Takeda and Fumitada Itakura

Center for Integrated Acoustic Information Research
Nagoya University
464-8603 Furo-cho, Nagoya, Japan

ABSTRACT

In this paper, we describe a multichannel method of noisy speech recognition that can adapt to various in-car noise situations during driving. The method allows us to estimate the log spectrum of speech at a close-talking microphone based on the multiple regression of the log spectra (MRLS) of noisy signals captured by multiple distributed microphones. Through clustering of the spatial noise distributions under various driving conditions, the regression weights for MRLS are effectively adapted to the driving conditions. The experimental evaluation shows an average error rate reduction of 43 % in isolated word recognition under 15 different driving conditions.

1. INTRODUCTION

Array-microphone signal processing is effective for spatially selective signal capture, hence, noisy speech recognition when the locations of the speaker and noise sources are predetermined. However, when the spatial configuration of the speaker and noise sources is unknown or changes continuously, it is not easy to steer the directivity adaptively to the new environment [1], [2], [3].

In order to improve the robustness to a small perturbation of the spatial distribution of the source and noise signals, we have proposed multiple regression of log spectra (MRLS), using log spectra of the signals captured by distributed microphones to approximate that of the close-talking microphone, through linear regression [4]. In the previous study, we implemented MRLS for in-car speech recognition and showed its effectiveness in improving the accuracy of noisy speech recognition. Through the experiments, we also found that further improvement of the recognition accuracy can be achieved if the regression weights are trained for each speaker and/or a particular in-car sound condition that is mainly governed by car conditions, e.g., music playing, open window, and fan noise, as well as driving speed. However, while training regression weights for a speaker at enrollment is not difficult, changing the weights in order to adapt to the driving conditions is not easy.

The aim of this study is to improve the MRLS so that regression weights can be changed adaptively to the in-car noise conditions. For this purpose, we make use of the advantage of *distributed* microphones for capturing the *spatial distribution* of noise sounds.

The rest of the paper is arranged as follows. First, in Section 2, we describe the in-car speech corpus recorded using distributed microphones. The basic idea of MRLS and its extension to the adaptive method are described in Section 3 and Section 4, respectively. In Section 5, experimental evaluations and their results are discussed. Section 6 is a summary of this paper.

2. MULTIPLE REGRESSION OF LOG SPECTRA

The basic idea of MRLS is to approximate the log power spectrum of speech recorded using a close-talking microphone, by a linear combination of the log power spectra of distributed distant microphones[4]. The approximation is given by the following procedure.

Suppose that $X_0(k)$ is the spectrum of the speech obtained by the close-talking microphone at the k^{th} spectral channel, and $X_i(k)$, *i=1,…, N*, are the spectra of the speech obtained by the distant microphones located at N different positions. The log spectral regression is given by

$$\log|X_0(k)| = \sum_{i=1}^{N} \bar{w}_i(k) \log|X_i(k)|, \qquad (1)$$

where $\bar{w}_i(k)$ are the real numbers that give the minimum regression error, i.e.,

$$\bar{w}_i(k) = \arg\min_{w_i(k)} E\left[d^2\right], \qquad (2)$$

where

$$d^2 = \sum_{k=1}^{K} \left\{ \log|X_0(k)| - \sum_{i=1}^{N} w_i(k) \log|X_i(k)| \right\}^2 . \qquad (3)$$

Here, the expectation, $E[]$, is calculated over all training utterances.

Note that the minimization of regression error $E\left[d^2\right]$ is equivalent to minimizing the cepstral distance between the approximated and the target spectrum, because of the orthogonality of the discrete time cosine transform (DCT) matrix. Therefore, the MRLS has the same form as the maximum likelihood optimization of the filter-and-sum beam former proposed in [5]. Applying the regression analysis in the log spectrum domain has two further merits: (1) the spectrum flooring due to over-subtraction can be avoided, and (2) the target spectrum for a wider range of intensity can be approximated.

3. ADAPTING MRLS TO THE AUTOMATICALLY DETECTED NOISE CONDITIONS

In the previous report[4], we found that changing regression weights adaptively to the driving conditions is effective in improving the recognition accuracy. In this section, we propose a method of discriminating in-car noise conditions, which is mainly affected by driving conditions, using *spatial distribution* of noise signals, and of controlling the regression weights for MRLS. The basic procedure of the proposed method is as follows. 1) Cluster the noise signals, i.e., short-time nonspeech segments preceding utterances, into several groups. 2) For each noise group, train optimal regression weights for MRLS, using the speech segments. 3) Finally, for unknown input speech, find a corresponding noise group from background noise, i.e., the nonspeech segments, and perform MRLS with the optimal weights for the noise cluster.

If there is a significant change in the sound source location, it greatly affects the relative intensity among distributed microphones. Therefore, in order to cluster the spatial noise distributions, we have developed a feature vector based on the relative intensity of the signals captured at the different positions to that of the nearest distant microphone, i.e.,

$$\mathbf{R} = [R_3(k), R_4(k), R_5(k), R_7(k)] \;\; k = 4, 5, \cdots 24,$$

where $R_i(k) = X_i(k)/X_6(k)$ is the relative power at the k^{th} mel-filterbank (MFB) channel calculated from the i^{th} microphone signal. We do not use the lower frequency channel because the spectra of stationary car noise is concentrated in the lower frequency region. Thus, $\mathbf{R}$ is a vector with 84 elements. As shown in Figure 1, the 6^{th} microphone is the one nearest to the driver. Finally, the 84 elements are normalized so that their mean and variance across elements are 0 and 1.0, respectively. Prototypes of noise clusters are obtained by applying the k-means algorithm to the feature vectors extracted from the training set of noise signals.

Table 1. Distributions of the noise samples in the four clusters.

	(1)NORMAL	(2)MUSIC	(3)FAN LO.	(4)FAN HI.	(5)OPN WIN.
cluster 1					
idle	**545**	10	0	0	232
city	**784**	69	130	8	100
express way	**895**	111	190	0	40
cluster 2					
idle	328	**873**	7	0	3
city	109	**827**	1	2	1
express way	3	**777**	5	2	0
cluster 3					
idle	24	15	**890**	**900**	28
city	0	2	**769**	**886**	5
express way	1	3	**695**	**898**	2
cluster 4					
idle	3	2	3	0	**637**
city	7	2	0	0	**794**
express way	1	9	2	0	**858**

In Table 1, an example of the clustering results are listed. The table shows how many samples of each driving condition each noise class contains when four clusters of noise are learned. As seen from the table, clusters are naturally formed for 'normal', 'music playing', 'fan' and 'open window' situations, regardless of the driving speeds. From the results, it is expected that the relative power of the sound signals at different microphone positions can be a good cue for controlling MRLS weights.

4. DISTRIBUTED MICROPHONE IN-CAR SPEECH CORPUS

The distributed microphone speech corpus is a part of the CIAIR (Center for Integrated Acoustic Information Research) in-car speech corpus collected at Nagoya University [6], which contains 800 speaker's speeches (isolated word utterances, read phonetically balanced sentences and dialogues) while driving. The data collection is performed using a specially designed data collection vehicle that has multiple data acquisition capabilities of up to 16 channels of audio signals, three channels of video signals and other driving related information (car position, speed, engine speed, brake and acceleration pedals and steering handle). Five microphones are placed around the driver's seat, as shown in Figure 1, where the top view and side view of the driver's seat is illustrated. In Figure 1, microphone positions are marked by the black circles. Microphones #3 and #4 are located on the dashboard; #5, #6 and #7 are attached to the ceiling. Microphone #6 is the one nearest to the speaker. In addition to these distributed microphones, the driver wears a headset with a close-talking microphone (#1).

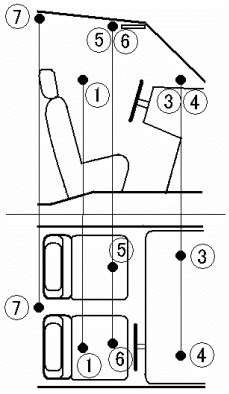

Fig. 1. Microphone positions for data collection: Side view (top) and top view (bottom).

In the most of the corpus, the speaker is driving in a city area near Nagoya University, however, a part of the corpus that we use in this study was collected under carefully controlled driving conditions, i.e., combinations of three car speeds (idle, driving in a city area and driving on an expressway) and five car conditions (fan on (hi/lo), CD player on, open window, and normal driving condition). For this part of the corpus, 50 isolated word utterances of 20 speakers were recorded under all combinations of driving speeds and car conditions.

5. EXPERIMENTAL EVALUATIONS

5.1. Experimental Setup

Speech signals used in the experiments were digitized into 16 bits at the sampling frequency of 16 kHz. For the spectral analysis, 24-channel mel-filterbank analysis is performed by applying the triangular windows on the FFT spectrum of the 25-ms-long windowed speech. This basic analysis is realized through HTK standard MFB analysis [7]. The regression analysis is performed on the logarithm of MFB output. Since the power of the in-car noise signal is concentrated in the lower frequency region, the regression analysis is performed for the range of 250-8kHz, i.e., 4^{th} to 24^{th} spectral channels of the MFB. Then DCT is executed to convert the log-MFB feature vector into the MFCC vector for the speech recognition experiments.

Three different HMMs are trained: 1) "close-talking HMM" is trained using the close-talking microphone speech, 2) "distant microphone HMM" is trained using the speech at the nearest distant microphone, and 3) "MRLS HMM" is trained using MRLS results. The regression weights optimized for each training sentence are used for generating the training data of MRLS HMM.

The structure of the three HMMs is fixed, i.e., 1) three-state triphones based on 43 phonemes that share 1000 states; 2) each state has 16-component mixture Gaussian distributions; and 3) the feature vector is a 25 (12 MFCC + 12 Δ MFCC + Δ logpower)-dimensional vector. The total number of training sentences is about 8,000. 2,000 of which were uttered while driving and 6,000 in an idling car.

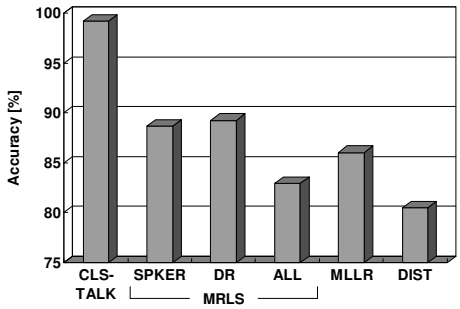

Fig. 2. Recognition performances averaged over various driving conditions. Close-talking (CLS-TALK), MRLS with optimized weights for a speaker (SPKER), with optimized weights for a driving condition (DR), with optimized weights for all training data (ALL), MLLR and distant microphone (DIST), from left to right.

5.2. Baseline Performance of MRLS

For the evaluation of the baseline performance of MRLS, five recognition experiments are performed: (1) recognition of close-talking speech using close-talking HMM ("CLS-TALK"), (2) recognition of MRLS output optimized for each speaker using MRLS HMM ("MRLS SPKER"), (3) recognition of MRLS output optimized for each driving condition using the MRLS HMM ("MRLS DR"), (4) recognition of MRLS output optimized for all training data using MRLS HMM ("MRLS ALL") and (5) recognition of nearest distant microphone speech by the distant microphone HMM ("DIST."),

The obtained recognition accuracies are listed in Table 2, and the average accuracies over fifteen driving conditions are shown in Figure 2. It is found that MRLS outperforms the nearest distant microphone result even in "MRLS ALL", where a set of *universal* weights are used for all conditions. This result confirms the robustness of the MRLS to the change of the location of the noise sources, because the primary noise locations are different depending on driving conditions. It is also found that the improvement is greater when the performance of the distant microphone is lower.

5.3. MRLS Performance with Weight Adaptation

To evaluate the performance of MRLS with weight adaptation, optimal regression weights for the four noise clusters described in Section 3 are trained. Using a 200 ms non-

Table 2. MRLS results obtained under various driving conditions.

	(1)CLS-TALK	(2)MRLS SPKER	(3)MRLS DR	(4)MRLS ALL	(5)DIST.
		NORMAL			
idle	99.67	99.67	99.56	99.89	99.56
city	99.78	98.67	98.78	98.33	98.22
ex. way	99.56	96.56	97.00	92.56	92.44
		MUSIC PLAY			
idle	99.33	88.78	95.22	90.89	84.00
city	99.00	90.56	93.22	90.22	85.56
ex. way	99.78	91.56	92.89	88.89	86.89
		FAN LO.			
idle	98.56	98.11	98.33	97.00	95.00
city	99.89	97.89	97.44	95.00	95.11
ex. way	99.44	95.33	95.33	89.44	90.78
		FAN HI.			
idle	98.89	75.22	76.22	59.44	53.89
city	98.55	78.79	79.58	65.51	61.38
ex. way	98.78	76.78	77.67	61.00	56.89
		OPEN WINDOW			
idle	99.56	95.67	95.44	92.56	88.33
city	98.89	86.22	85.56	77.11	75.78
ex. way	99.00	60.56	56.78	46.33	43.33

speech segment preceding the utterance, the nearest prototype of the noise cluster is searched, then, the utterance is recognized after MRLS with the regression weights optimized for the corresponding noise cluster. The same MRLS HMM is used. The results of the experiments are shown in Figure 3. As seen in Figure 3, the performance of the MRLS using adaptive regression weights is as high as the results of using the optimally trained weights for each driving condition. Furthermore, the MRLS outperforms the MLLR adaptation (five-word supervised adaptation) applied to the close-talking speech [8]. Therefore, the effectiveness of the proposed method is confirmed.

6. SUMMARY

In this paper, we described a multichannel method of noisy speech recognition that can adapt to various in-car noise conditions during driving. The method allows us to estimate the log spectrum of speech at a close-talking microphone based on the multiple regression of the log spectra (MRLS) of noisy signals captured by multiple distributed microphones. Through clustering of the spatial noise distributions under various driving conditions, the regression weights for MRLS are effectively adapted to the driving conditions. The experimental evaluation shows an error rate reduction of 43 % in isolated word recognition under various driving conditions.

Acknowledgement: This work has been supported by a Grant-in-Aid for COE Research (No. 11CE2005).

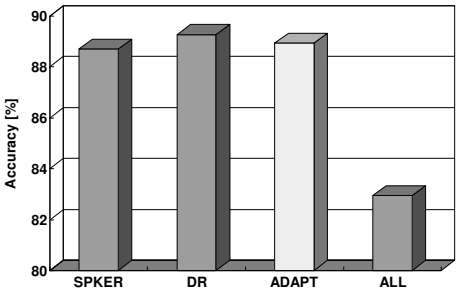

Fig. 3. Recognition performance of MRLS with optimized weights for a speaker (SPKER), with optimized weights for a driving condition (DR), proposed method (ADAPT), with optimized weights for all training data (ALL), from left to right.

7. REFERENCES

[1] Widrow, B. et al., “Adaptive Noise Cancelling: Principles and Applications”, Proc. IEEE, Vol.63, No.12, (1975.12).

[2] Kaneda, Y. and Ohga, J., “Adaptive Microphone-Array System for Noise Reduction”, IEEE Trans. Acoustics Speech and Signal Processing, 34 (6): 1391-1400, (1986).

[3] Yamada, T., Nakamura, S. and Shikano, K. ”Distant-talking speech recognition based on a 3-D Viterbi search using a microphone array”, IEEE Transactions on Speech and Audio Processing, Vol.10, No.2, pp.48-56, February 2002

[4] T.Shinde K. Takeda and F. Itakura, “Multiple regression of Log Spectra for in-car speech recognition”, Proc. International Conference on Spoken Language Processing, Vol.II, pp.797-800, 2002 (ICSLP2002, Denver)

[5] Michael L. Seltzer, Bhiksha Raj, and Richard M. Stern, “Speech Recognizer-based microphone array processing for robust hands-free speech recognition”, Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol.I, pp.897-900, 2002 (ICASSP2002, Orlando)

[6] Kawaguchi, N., Takeda, K., et al., “Construction of Speech Corpus in Moving Car Environment”, Proc. International Conference on Spoken Language Processing, pp.1281-1284, 2000 (ICSLP2000, Beijing, China).

[7] Young, S. et al. “The HTK Book”

[8] C.J.Leggetter and P.C.Woodland, “Flexible speaker adaptation using maximum likelihood linear regression,” Proc. of the ARPA Spoken Language Technology Workshop, 1995, Barton Creek