MULTIPLE REGRESSION OF LOG-SPECTRA FOR IN-CAR SPEECH RECOGNITION

Tetsuya Shinde, Kazuya Takeda and Fumitada Itakura

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

ABSTRACT
This paper describes a new multichannel method of noisy speech recognition, which estimates 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 the distributed microphones. Since the method does not assume the arrangement of sound sources and microphones, it can be applied to in-car speech recognition directly. The experimental evaluation shows an error reduction of up to 15% in isolated word recognition accuracy under various driving conditions.

1. INTRODUCTION
Improving the recognition accuracy under a noisy environment is one of the important issues in extending the application domain of the speech recognition technology[1]. Among the various approaches for noisy speech recognition, speech enhancement based on multichannel data acquisition is currently under extensive research. The most fundamental and important multichannel method is the microphone array beam-former method because the assumption that the target and the interfering signals are not spatially dispersed, and are apart from each other, is reasonable in many situations. In other words, the microphone array beam-former method is effective when the positions of the speaker and the noise sources are predetermined and the positions do not change during the signal acquisition process.

On the other hand, 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. An adaptive microphone array system for noise reduction (AMNOR) is a general framework that provides an optimal directivity pattern to the given spatial distribution of the target signal and noise [2], [3]. However, the AMNOR assumes that 1) the target sound direction is known, and 2) the noise signal that does not contain the target signal can be observed through multiple microphones. In-car speech recognition has some difficulties in taking advantage of the beam-former because neither the speaker position nor the primary noise location is fixed. In such situations, the change of the target and interfering signal locations may even aggravate the recognition performance due to the use of microphone-array-based signal processing. The tolerance of the multichannel system to the change of the spatial configuration is an important issue for in-car speech recognition.

The authors have earlier proposed the feature averaging method for recognizing distant speech captured through distributed microphones. In [4], we experimentally confirmed that using the average cepstrum of the distant speech signals captured through distributed microphones improves the recognition accuracy of the distant speech by about 20%. In this paper, we extend the idea of feature averaging to multiple regression of the log spectrum (MRLS) of the speech captured by the distant microphones so as to approximate the log spectrum of the speech of the close-talking microphone. The expected merits of the proposed method are as follows: 1) The method does not make any assumptions about the positions of the speaker and noise sources with respect to the microphones. Therefore, the system can be trained for various sitting positions of drivers. 2) Since the regression weights can be optimized over a certain length of speech segments, e.g., sentences of speech under particular road conditions, the tolerance to the change of the environment is high.

Because our main objective is in-car speech recognition, we assume that the speaker position does not change within a driver, and that the noise source location is governed mainly by the car conditions such as the conditions of windows, wipers, and the road. Once the regression analyses are performed for each speaker and driving condition, and the optimal regression coefficients are decided, the performance of the enhancement is not affected by small changes in the conditions.

The aim of this paper is to describe the proposed method and evaluate the performance of the method for in-car speech recognition. In Section 2, we introduce the spectral regression method for speech enhancement. In Section 3, the in-car speech database that is used for the evaluation experiments is described. In Sections 4 and 5, the evaluation experiments are presented. Section 6 summarizes this paper.
2. LOG SPECTRAL MULTIPLE REGRESSION

The basic idea of the proposed method is to approximate the log power spectrum of the close-talking microphone speech by a linear combination of the log power spectra of the distant microphones. The approximation is given by the following procedure.

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

\[
\log |X_{\text{ref}}(k)| = \sum_{i=1}^{N} \tilde{a}_i(k) \log |X_i(k)|, \tag{1}
\]

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

\[
\tilde{a}_i(k) = \arg \min_{a_i(k)} E \left[ d^2 \right], \tag{2}
\]

where

\[
d^2 = \sum_{k=1}^{K} \left( \log |X_{\text{ref}}(k)| - \sum_{i=1}^{N} a_i(k) \log |X_i(k)| \right)^2. \tag{3}
\]

Here, the expectation, \( E[] \), is calculated over the training utterances.

Note that the regression error \( \min E \left[ d^2 \right] \) is equal to the cepstral distance between the approximated and the target spectrum because of the orthogonality of the discrete cosine transform (DCT) matrix. Therefore, the method can be considered as an extension of feature average in the cepstrum domain by replacing the average value with the weighted sum. Furthermore, applying the regression analysis in the log spectrum domain has the following two merits: (1) the spectrum flooring due to the oversubtraction can be avoided, and (2) the target spectrum for a wider range of intensity can be approximated.

3. IN-CAR SPEECH CORPUS

The speech data used for the experiments is a part of the CIAIR in-car speech corpus collected at Nagoya University [5]. The corpus consists of speech data collected from 800 speakers for utterances of isolated words, phonetically balanced sentences, dialogues with a human operator, a ‘wizard-of-oz’ system and a fully automated spoken dialogue system for the restaurant guidance task domain. The speech data is collected while the speaker is driving on the city streets near the university. 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 and three channels of video signals. Five microphones are placed around the driver as shown in Figure 1. The driver wears a headset with a close-talking microphone. The speech signal captured by this close-talking microphone is used as the reference in the regression analysis.

The database that is used for the recognition experiments is collected through the same setup but the driving conditions are carefully controlled. The database consists of 50 isolated word utterances recorded for three traffic conditions (idling, driving in a city and driving on an expressway), and five driving conditions (the fan on (hi/lo), CD player on, hazard-light on, an open window, and the normal driving condition).

4. EXPERIMENTS ON SPECTRAL REGRESSION

4.1. Speech Analysis

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-filter-bank (MFB) analysis is performed by applying the triangular windows on the FFT spectrum of the 25 millisecond-long windowed speech. This basic analysis is realized through HTK standard MFB analysis [6]. 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., \( 4^{\text{th}} \) to \( 24^{\text{th}} \) spectral channel of the MFB. Then DCT is executed to convert the log-MFB feature vector into the MFCC vector for the speech recognition experiments.

4.2. Spectral Distortion

First, the effectiveness of the approximation by the regression is evaluated from the viewpoint of the spectral distor-
tion. The signal-to-deviation ratio (SDR) is given by

\[
\text{SDR} = 10 \log \frac{\{ X^{(i)}(k) \}^2}{\{ X^{(i)}(k) - \tilde{X}^{(i)}(k) \}^2},
\]

where \( X^{(i)}(k) \) is the log spectrum of the reference speech captured by the close-talking microphone, and \( \tilde{X}^{(i)}(k) \) is the log spectrum to be evaluated. The evaluated log spectra are that of the distant microphone output and the result of the approximation through multiple regression. The SDR calculation is performed over 8,000 phonetically balanced sentences. In this experiment, the regression weights are calculated for each sentence and the resultant regression error is evaluated. Thus, the result gives the upper bound of the SDR value.

The results are shown in Figure 2. Among the distributed distant microphones, the best SDR value is obtained by the nearest microphone, mic. 6, and the second best result is obtained by mic. 4 which is located far but in front of the driver. The effectiveness of the spectrum regression is revealed from the result that the SDR of the approximated log spectrum is about 6 dB higher than that of the best microphone result.

5. RECOGNITION EXPERIMENT

5.1. Experimental Setup

In this section, the effectiveness of the proposed method is evaluated through speech recognition experiments. Throughout this section, the same structure is used for the set of triphone HMMs, i.e., 1) they 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.

For comparison, the following three different sets of HMM are trained: 1) close-talking model is trained using the
close-talking microphone speech, 2) **distant microphone model** is trained using the speech at the nearest microphone (mic. 6 in Figure 1), and 3) **MRLS model** is trained using the spectra obtained from the MRLS method. For training the MRLS model, the regression weights are optimally determined for each training sentence. The total number of training sentences is about 8,000; 2,000 of them are uttered while driving and 6,000 of them are uttered in the idling car.

The test data includes a isolated word utterances of a 50 word set. Each of 18 speakers uttered the word set under 18 different car conditions. For each utterance, six different versions of the speech data are recognized. They are 1) speech recorded using the close-talking microphone, 2) speech recorded at the nearest microphone, 3) MRLS output with the optimally determined weights for each utterance, 4) MRLS output with the optimally determined weights for each speaker, 5) MRLS output with the optimally determined weights for each driving condition, and 6) MRLS output with the optimally decided weights for all of the training data.

Note that case 3) is an unrealistic case in that the close-talking speech itself can be used for recognition. The results for this case indicate the upper bound of the MRLS. On the other hand, cases 4) and 5) assume that the optimal weights for MRLS are constant for each speaker or driving condition.

### 5.2. Results

For the evaluation, six recognition experiments are performed: 1) recognize close-talking speech by the close-talking model (**close-talking**), 2) recognize nearest microphone speech by the distant microphone model (**distant**), 3) recognize optimal MRLS output by the MRLS model (**MRLS opt.**), 4) recognize MRLS output optimized for each speaker by the MRLS model (**MRLS speaker**), 5) recognize MRLS output optimized for each driving condition using the MRLS model (**MRLS cond.**), and 6) recognize MRLS output optimized for all training data using the MRLS model (**MRLS all**). In addition, the results of spectrum subtraction (**SS**) are also compared where the training and the test speech at the nearest distant microphone are enhanced by the spectrum subtraction [7].

The results are shown in Figure 3 for each car condition. It is found that MRLS outperforms the nearest distant microphone result even in the MRLS:all case. This result suggests the robustness of the method to the change of the location of the noise sources, because the primary noise locations are different between ‘open window’ and ‘fan on’ cases. It is also found that the improvement is larger when the performance of the distant microphone is lower. Furthermore, by optimizing the regression weights for each speaker or driving condition, recognition accuracy can be further improved, but the performance is still not as high as the result of the upper bound.

### 6. SUMMARY

In this paper, we have proposed a method based on multiple regression of the log spectrum (MRLS) for in-car speech recognition. The proposed method is suited to the in-car application because it does not assume the spatial configuration of the sound sources and microphones and the weights for the regression can be trained using a certain length of speech data. From the experimental evaluation, the effectiveness of the method is demonstrated for the in-car speech collected under the real driving conditions. Further improvement is expected by effectively grouping the environmental conditions for adaptively updating the regression weights.

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

### 7. REFERENCES


