Optimizing Regression for in-car Speech Recognition using Multiple Distributed Microphones

Weifeng Li, Kazuya Takeda, Fumitada Itakura

Center for Integrated Acoustic Information Research
Nagoya University

Abstract

In this paper, we address issues in improving hands-free speech recognition performance in different car environments using multiple spatially distributed microphones. In previous work, we proposed multiple regression of the log-spectra (MRLS) for estimating the log-spectra of speech at a close-talking microphone. In this paper, the idea is extended to nonlinear regressions. Isolated word recognition experiments under real car environments show that, compared to the nearest distant microphone, recognition accuracies could be improved by about 40% for very noisy driving conditions by using the optimizing regression method. The proposed approach outperforms linear regression methods and also outperforms adaptive beamformer by 8% and 3% respectively in terms of averaged recognition accuracies.

1. Introduction

Noise-robustness is today one of the most challenging and important problems in automatic speech recognition (ASR). State-of-the-art speech recognition systems are known to perform reasonably well by using a close-talking microphone worn near the mouth of the speaker. In applications where the use of a close-talking microphone is neither desirable nor practical, the use of a distant microphone is required. However, as the distance between speaker and microphone increases, the recorded signal becomes more susceptible to distortion from background noise which severely degrades the performance of ASR systems. This problem can be greatly alleviated by the use of multiple microphones to capture the speech signals. Among the various approaches using multiple microphones, microphone arrays are commonly used for data collection and speech enhancement. In real car environments, traditional beamforming is difficult to apply because both speaker position and noise location are not fixed [1]. Recently an adaptive beamforming-based approach (Generalized Sidelobe Canceller (GSC) [2] for example) has become attractive for its dynamic parameter adjustment. However, a persistent problem in microphone arrays has been the poor low frequency directivity for practical array dimensions [3]. Also, problems such as "signal leakage" [4] inherent in GSC degrade its quality.

As a new multi-microphone approach to improve the in-car speech recognition performance, multiple regression of log spectra (MRLS) was proposed in [5], in which multiple spatially distributed microphones are used. The basic idea of MRLS is to approximate the log mel-filter bank (log MFB) outputs of a close-talking microphone by a linear combination with the log spectra of the distant microphones, and it can be regarded as an extension of spectral subtraction. The regression method was shown to be effective in real car environments because it can statistically optimize the parameters in terms of minimum log spectral distance using a large in-car speech corpus.

In this paper, we extend the linear regression to nonlinear regressions, such as multi-layer perceptron (MLP) and support vector machines (SVM), to approximate the speech signals of the close-talking microphone because we believe that a better approximation will result in the improvement of recognition performance. We also investigate the regression approach on the cepstrum domain beside the log-spectra. The effectiveness of the proposed nonlinear approaches are demonstrated in the improvement of both signal-to-deviation ratio (SDR) and word recognition accuracies.

The organization of this paper is as follows: In Section 2, regression methods used for combining the data from the distant microphones are presented. In Section 3, we describe the result of experimental evaluations using in-car driver speech under real environments. Comparison with the adaptive beamforming is also presented. Section 4 summarizes this paper.

2. Regression Methods

Let \( \mathbf{x}_i \) denote the feature vector (e.g. log-spectrum or cepstrum) for the \( i \)th distant microphone and \( x^k_i(n) \) denote the corresponding \( k \)th element for frame \( n \). Let \( \mathbf{y} \) denote the feature vector collected from the close-talking microphone and \( y^k(n) \) denote the \( k \)th element for frame \( n \). Let \( \hat{\mathbf{y}} \) denote the estimated feature vector obtained from the feature vectors of five distant microphones. Each element of \( \hat{\mathbf{y}} \) is approximated independently.
In the linear regression (LR), \( \hat{y}^k(n) \) is given by
\[
\hat{y}^k(n) = \sum_{i=1}^{5} w_i x_i^k(n) + b. \tag{1}
\]

\( w_i \)'s are the real valued elements of the 5-dimensional vector \( w \) and \( b \) is the bias, which are obtained by minimizing mean squared error over the training examples.

In support vector machine regression (SR), we introduce the vector
\[
x^k(n) = [x_1^k(n), x_2^k(n), x_3^k(n), x_4^k(n), x_5^k(n)],
\]
and the \( k \)th element of the feature vector is estimated by
\[
\hat{y}^k(n) = \sum_{m=1}^{N} (\alpha(m) - \alpha^*(m)) [K(x^k(n), x^k(m))] + b. \tag{2}
\]
where \( \alpha(m) \) and \( \alpha^*(m) \) are the Lagrange multipliers for the \( m \)th support vectors \( x^k(m) \). The Lagrange multipliers (\( \alpha, \alpha^* \)) and the support vectors are found by solving the dual optimization problem in support vector learning [6]. \( K(\cdot) \) denotes the kernel function for which we use the Gaussian kernel in the experiments.

For multi-layer perceptron regression (MR), the network with one hidden layer composed of 8 neurons is used. The \( k \)th element of the feature vector is estimated by
\[
\hat{y}^k(n) = \sum_{p=1}^{8} \left( w_p \tanh \left( b_p + \sum_{i=1}^{5} w_{ip} x_i^k(n) \right) \right) + b. \tag{3}
\]
where \( \tanh(\cdot) \) is the tangent hyperbolic activation function. The parameters \( \{ w_p, w_{ip}, b_p, b \} \) are found by minimizing the mean squared error over the training examples using a gradient descent algorithm [7].

3. In-car Recognition Experiments

3.1. Experimental setup

A data collection vehicle (DCV) has been specially designed for developing the in-car speech corpus at Center for Integrated Acoustic Information Research (CIAIR), Nagoya University, Nagoya, Japan. Figure 1 shows the side view and the top view of the arrangement of microphones in the data collection vehicle. The driver wears a headset with a close-talking microphone (1 in Figure 1) placed in it. Five distant microphones (3 to 7) are placed around the driver. A four-element linear microphone array (9 to 12) with an inter-element spacing of 5 cm is located at the visor position. The speech signals captured by the close-talking microphone are used as the reference in the regression analysis. Speech signals are digitized into 16 bits at a sampling frequency of 16 kHz.

For spectral analysis, 24 channel mel-filter-bank (MFB) analysis is performed by applying a triangular window on the FFT spectrum of 25 millisecond-long windowed speech, with a frame shift of 10 milliseconds. Spectral components lower than 250 Hz are filtered out. Then log MFB parameters are estimated. Finally 12 mean normalized mel-frequency cepstral coefficients (CMN-MFCC) are obtained through Discrete Cosine Transformation (DCT) on the log MFB parameters and by subtracting the cepstral means.

We performed isolated word recognition experiments on the 50 word sets under 15 driving conditions (3 driving environments \( \times 5 \) in-car states = 15 driving conditions as listed in Table 1). For each driving condition, 50 words are uttered by each of 18 speakers. Among them the data uttered by 12 speakers (6 male and 6 female) is used for learning the regression models and the remaining words uttered by 6 speakers (3 male and 3 female) are used for recognizing. 1,000-state triphone Hidden Markov Models (HMMs) with 32 Gaussian mixtures per state trained over a total of 7,000 phonetically balanced sentences (uttered by 202 male speaker and 91 female speaker) is used for acoustical modeling. (3,600 of them are collected in the idling-normal condition and 3,400 of them are collected while driving the DCV on the streets near the Nagoya university (city-normal condition).)

![Figure 1: Side view and top view of the arrangement of multiple spatially distributed microphones and the linear array in the data collection vehicle.](image)
We performed linear regression [5] (Equation (1)) on the 24 log MFB outputs and the 12 CMN-MFCCs for both training data and test data. The regression is also performed on the log energy parameter. The estimated feature vectors used for recognition system consist of 12 CMN-MFCCs + 12 Δ CMN-MFCCs + Δ log energy. In the case of regression on log MFB outputs, the estimated log MFB vectors are then converted into CMN-MFCC vectors using DCT, and then perform the derivatives. In this way, we obtain two regression-based HMMs, which are used to recognize the test data obtained in the same way (“LRB-LRB” and “LRC-LRC” correspondingly). For comparison, we also cite the “CTK” acoustical model trained using the close-talking microphone speech and “DST” model trained using the speech from the nearest microphone (microphone 6 in Figure 1). The recognition performance averaged over the 15 driving conditions is given in Figure 3. We also recognize the estimated data by using the “CTK” model and “DST” model.

From Fig 3, it is found that good recognition performance can be obtained by using the close-talking microphone for good quality speech signals. The regression methods outperform the nearest distant microphone. This shows that the approximation of the speech signals from the close-talking microphone is reasonable and practical to improve the recognition accuracy. The regression on log MFB shows the advantage over that on CMN-MFCC. The “DST-LRB” performs best and outperforms the “DST” by about 10% averagely. So the “DST” model and the regression on log MFB are used in the following.

### 3.3. Nonlinear regression for in-car speech recognition compared to adaptive beamforming

Next we perform the condition-dependent nonlinear regressions to obtain estimated test data. The SVM regression method (Equation (2)) and MLP regression method (Equation (3)) are considered. The parameters \((C, \sigma, \epsilon)\) in SVM are specified as \((10, 5, 0.5)\). The learning rate and the number of iterations are set as 0.001 and 1000 respectively. The regression performance is evaluated by the signal-to-deviation ratio (SDR), which is defined as

\[
\text{SDR} [\text{dB}] = 10 \log_{10} \frac{\sum_{n=1}^{N} \| \hat{y}(n) \|^2}{\sum_{n=1}^{N} \| y(n) - \hat{y}(n) \|^2} \quad (4)
\]

where \(y(n)\) is the reference feature vector from close-talking microphone for frame \(n\). \(N\) denotes the number of frames during one utterances. The SDR is averaged over the number of the utterances. Figure 4 shows the SDR values obtained using the three regression methods.

The adaptive microphone arrays approach is attractive for speech enhancement and speech recognition (e.g. [8]). For comparison, we apply the Generalized Side-lobe C canceller (GSC) [2] to our in-car speech recognition. Four linearly spaced microphones (9 to 12 in Figure 1) are used. The architecture of the used GSC is shown in Figure 5. The three FIR filters are adapted sample-by-sample.
MLP regression is superior to adaptive beamforming. In view of the computation cost, the proposed improvement can be expected by using the regression method. For the case when the window near the driver is open or when the CD player is on, if another microphone is available near the window/CD-player to obtain more information about the wind/music interference, further improvement can be expected by using the regression method. In view of the computation cost, the proposed MLP regression is superior to adaptive beamforming.

3.4. Discussions

Our experiments show that adaptive beamforming is very effective when the window near the driver is open and when the CD player is on. This suggests it could track the target speech signals effectively from other directions. The proposed nonlinear regression shows advantages when the air-conditioner is on. Especially when the air-conditioner is on high level, the proposed method outperforms the nearest distant microphone and adaptive beamforming by about 40% and more than 10% respectively. For the case when the window near the driver is open or when the CD player is on, if another microphone is available near the window/CD-player to obtain more information about the wind/music interference, further improvement can be expected by using the regression method. In view of the computation cost, the proposed MLP regression is superior to adaptive beamforming.

4. Summary

In this work, we have proposed nonlinear regression of the log-spectra for in-car speech recognition by using multiple distant microphones. The results of our studies have shown that the proposed method can obtain good approximation to the speech of a close-talking microphone. The effectiveness is also demonstrated in the improvement of the word recognition accuracies in 15 driving conditions. Other methods for speech enhancement may be combined with the proposed method to obtain improved accuracy in recognition of speech in noisy environments.

5. References


