IEICE TRANS. INF. & SYST., VOL.E89–D, NO.3 MARCH 2006

PAPER Special Section on Statistical Modeling for Speech Processing

Single-Channel Multiple Regression for In-Car Speech Enhancement

Weifeng LI[†a)], **Katsunobu ITOU**[††], *Nonmembers*, **Kazuya TAKEDA**[††], *Member*, *and* **Fumitada ITAKURA**[†††], *Fellow*

SUMMARY We address issues for improving hands-free speech enhancement and speech recognition performance in different car environments using a single distant microphone. This paper describes a new single-channel in-car speech enhancement method that estimates the log spectra of speech at a close-talking microphone based on the nonlinear regression of the log spectra of noisy signal captured by a distant microphone and the estimated noise. The proposed method provides significant overall quality improvements in our subjective evaluation on the regression-enhanced speech, and performed best in most objective measures. Based on our isolated word recognition experiments conducted under 15 real car environments, the proposed adaptive nonlinear regression approach shows an advantage in average relative word error rate (WER) reductions of 50.8% and 13.1%, respectively, compared to original noisy speech and ETSI advanced front-end (ETSI ES 202 050).

key words: *speech enhancement, speech recognition, multi-layer perceptron, mean opinion score, pairwise preference test, environmental adaptation, K-means clustering*

1. Introduction

Speech quality and intelligibility often significantly deteriorate in the presence of background noise, which degrades performance in the subsequent processing, such as speech coding or automatic speech recognition. Consequently, modern communications systems employ some speech enhancement procedures at the preprocessing stage prior to further processing (e.g., speech recognition). Speech enhancement algorithms have therefore been attractive research in the past two decades. Especially, in view of the steady increase for hands-free communication systems in car-driving environments, there is renewed interest in speech enhancement algorithms using a distant microphone.

Among a variety of speech enhancement methods, *spectral subtraction* (SS) [1], [2] and *short-time spectral attenuation* (STSA) based methods [3], [4] are commonly used. Most SS based methods make assumptions about the independence of speech and noise spectra, allowing for simple linear subtraction of the estimated noise spectra. Although scaling factors for emphasis or de-emphasis of the estimated noise have been proposed to reduce "musical tone" artifacts, the specifications of the scaling factors are usually done experimentally. STSA based methods can lead to a nonlinear spectral estimator by introducing a priori SNR; however, they require assumptions about *ad hoc* statistical distributions for speech and noise spectra [5], [6]. Usually both SS and STSA based methods can only handle additive noise.

In previous work, we proposed a new and effective multi-microphone speech enhancement approach based on multiple regression of log spectra [7] that used multiple spatially distributed microphones. Their idea is to approximate the log spectra of a close-talking microphone by effectively combining of the log spectra of distant microphones. The approach made no assumption about the positions of the speaker and noise sources with respect to the microphones, and worked in very small computation amounts. It has been shown to be very effective based on our previous in-car speech recognition experiments [8].

In this paper, we extend the idea to single-microphone cases and propose that the log spectra of clean speech are approximated through the nonlinear regression of the log spectra of the observed noisy speech and the estimated noise. The proposed approach, which can be viewed as generalized log spectral subtraction, has the following properties: 1) It does not need any assumption concerning independence and statistical distribution of speech and noise spectra; 2) It can deal with a wide range of distortions, rather than only additive noise; 3) Regression weights are obtained through statistical optimization. Once the optimal regression weights are obtained in the learning phase, they are utilized to generate the estimated log spectra in the test phase, where clean speech is no longer required.

The main aim of this paper is to describe the proposed method and evaluate its performance on speech enhancement and recognition. Moreover, a two-stage noise spectra estimator is developed for additional improvement of the speech recognition performance. To develop a data-driven in-car recognition system, we also devise an effective algorithm for automatically adapting regression weights for different noise environments. The organization of this paper is as follows: In Sect. 2, we describe the in-car speech corpus used in this paper. In Sect. 3, we present the proposed regression-based speech enhancement algorithm. In Sect. 4, we present subjective and objective evaluation ex-

Manuscript received June 30, 2005.
Manuscript revised September 22, 2005.

[†]The author is with the Department of Information Electronics, Graduate School of Engineering, Nagoya University, Nagoya-shi, 464–8603 Japan.
[††]The authors are with the Department of Media Science, Graduate School of Information Science, Nagoya University, Nagoya-shi, 464–8603 Japan.
[†††]The author is with the Faculty of Science and Technology, Meijo University, Nagoya-shi, 468–8502 Japan.
a) E-mail: lee@sp.m.is.nagoya-u.ac.jp

DOI: 10.1093/ietisy/e89–d.3.1032

periments on regression-enhanced speech. We describe our speech recognition experiments using the proposed method in Sect. 5 and present the improvements in Sect. 6. Finally, conclusions are drawn in Sect. 7.

2. In-Car Speech Data and Speech Analysis

The speech data used are from CIAIR in-car speech corpus [9]. Speech captured by a microphone at the visor position is used in the following experiments. Speech collected at a close-talking microphone (with a headset) is used for reference speech. Test data includes Japanese 50 word sets under 15 driving conditions (three driving environments × five in-car states = 15 driving conditions, as listed in Table 1). For each driving condition, 50 words were uttered by each of 18 speakers. The training data for acoustical modeling comprised a total of 7,000 phonetically balanced sentences, uttered by 202 male speakers and 91 female speakers. 3,600 sentences were collected in the idling-normal condition and 3,400 were collected while driving a data collection vehicle (DCV) on the streets near Nagoya University (city-normal condition).

Speech signals were digitized into 16 bits at a sampling frequency of 16 kHz. For spectral analysis, a 24-channel MFB analysis was performed on 25-millisecond windowed speech with a frame shift of 10 milliseconds. Spectral components lower than 250 Hz were filtered out to compensate for the spectrum of engine noise, which was concentrated in the lower frequency region. Log MFB parameters were then estimated. The estimated log MFB vectors were transformed into 12 mean normalized mel-frequency cepstral coefficients (CMN-MFCC) using Discrete Cosine Transformation (DCT) and mean normalization, after which time derivatives (Δ CMN-MFCC) were calculated. These analyses were realized by using HTK toolkits.

3. Regression-Based Speech Enhancement

Let $s(i)$, $n(i)$, and $x(i)$ respectively denote the reference clean speech (referred to as speech at a close-talking microphone in this paper), noise, and observed noisy signals. By applying a window function and analysis using short-time discrete Fourier transform (DFT), in the time-frequency domain we have $S(k,l)$, $N(k,l)$, and $X(k,l)$, where k and l denote frequency bin and frame indexes, respectively. After the log operation of the amplitude, we obtain $S^{(L)}(k,l)$, $X^{(L)}(k,l)$, and $N^{(L)}(k,l)$:

Table 1 15 driving conditions (3 driving environments × 5 in-car states).

driving environment	idling city driving expressway driving
in-car state	normal CD player on air-conditioner (AC) on at low level air-conditioner (AC) on at high level window (near driver) open

$$S^{(L)}(k,l) = \log|S(k,l)|,$$

$$X^{(L)}(k,l) = \log|X(k,l)|,$$

$$N^{(L)}(k,l) = \log|N(k,l)|.$$

The idea of regression-based speech enhancement is to approximate $S^{(L)}(k,l)$ by combining $X^{(L)}(k,l)$ and $N^{(L)}(k,l)$, as shown in Fig. 1. Let $\hat{S}^{(L)}(k,l)$ denote the estimated version obtained from the inputs of $X^{(L)}(k,l)$ and $N^{(L)}(k,l)$. We can obtain $\hat{S}^{(L)}(k,l)$ by employing a *multi-layer perceptron* (MLP) regression method, where a network with one hidden layer composed of eight neurons is used. (The number of neurons are determined experimentally.)

$$\hat{S}^{(L)}(k,l) = b_k + \sum_{p=1}^{8}\Big(w_{k,p}\tanh(f(X^{(L)}(k,l), N^{(L)}(k,l))\Big),$$

where $\tanh(\cdot)$ is the tangent hyperbolic activation function and

$$f(X^{(L)}(k,l), N^{(L)}(k,l)) = b_{k,p} + w^x_{k,p}X^{(L)}(k,l) + w^n_{k,p}N^{(L)}(k,l).$$

Here p is the index of the hidden neurons. The parameters (regression weights) $\Theta = \{b_k, w_{k,p}, w^x_{k,p}, w^n_{k,p}, b_{k,p}\}$ are found by minimizing the mean squared error (MSE):

$$\mathcal{E}(k) = \sum_{l=1}^{J}[S^{(L)}(k,l) - \hat{S}^{(L)}(k,l)]^2, \quad (1)$$

through the back-propagation algorithm [10]. Here, J denotes the number of training examples (frames). Once $\hat{S}^{(L)}(k,l)$ is obtained for each frequency bin, enhanced speech can be generated by taking the exponential operation and performing short-time inverse discrete Fourier transform (IDFT) with the combination of the phase of the observed noisy speech.

The proposed approach is cast into single-channel methodology because once the optimal regression parameters are obtained by regression learning, they can be utilized to generate $\hat{S}^{(L)}(k,l)$ in the test phase, where the speech of the close-talking microphone is no longer required. Multiple regression means that regression is performed for each frequency bin. The use of minimum mean squared error in the log spectral domain is motivated by the fact that log spectral measure is more related to the subjective quality of

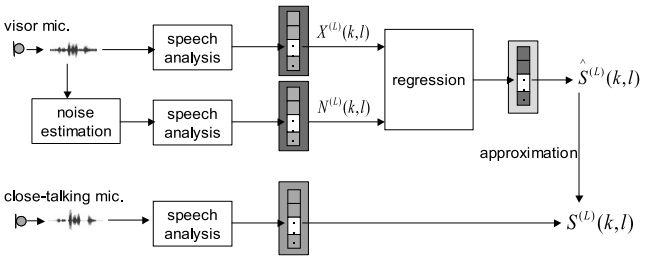


Fig. 1 Concept of regression-based speech enhancement.

speech [11] and that some better results have been reported with log distortion measures [12]† [13]. Although both the proposed regression-based method and *log-spectra amplitude* (LSA) estimator [6] employ minimum mean squared errors (MMSE) cost function in the log domain, the former makes no assumptions regarding the distributions of speech and noise spectra. The proposed method differs from [13] in that it does not need to estimate the mean and variance of the log spectra of clean speech, which is nontrivial because only noisy speech is available. Moreover, the proposed method employs more general regression models and is frame-based (without delay).

4. Speech Enhancement Performance

4.1 Experimental Data

The test speech was based on 50 isolated word sets under seven real driving conditions listed in Table 2. Figure 2 shows a block diagram of the regression-based speech enhancement system for a particular driving condition. For each driving condition, the data uttered by 12 speakers were used for learning the regression weights, and the remaining 300 words from different six speakers (three male and three female) were used for open testing.

For comparison, a *parametric formulation of the generalized spectral subtraction* (PF-GSS) [14] and a *log-spectra amplitude* (LSA) estimator [6] were also applied. For PF-GSS, the version with constraint, which was suggested by the authors, was used. An *a priori* SNR was calculated by the well-known "decision-directed" approach [4]. An *improved minima controlled recursive averaging* (IMCRA) method [15] was used to estimate noise for all the enhanced methods. We selected PF-GSS and LSA because they can provide good noise reduction and reduce the annoying "musical tone" artifacts of enhancement schemes based on conventional spectral subtraction while maintaining relatively low computational complexity. Four types of speech (or algorithms) must be evaluated:

1. original: observed noisy speech with no processing;
2. PF-GSS: speech enhanced using the PF-GSS method;
3. LSA: speech enhanced using the LSA method;
4. regression: speech enhanced using the proposed regression method.

Table 2 Seven driving conditions for speech enhancement evaluation.

driving environment	in-car state
city driving	normal
city driving	CD player on
city driving	air-conditioner on at high level
city driving	window open
idling	normal
expressway driving	normal
expressway driving	window open

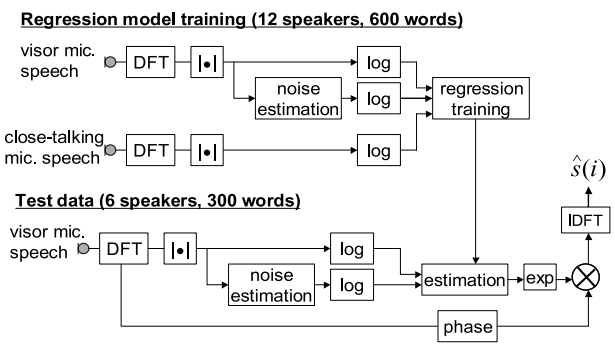


Fig. 2 Diagram of regression-based speech enhancement.

4.2 Subjective Evaluations

For each driving condition, five speech samples were randomly selected from the 300 test signals. The characteristics of enhanced speech signals differ according to driving conditions and algorithms. Therefore, the total number of speech samples was five samples × seven driving conditions × four algorithms = 140.

Twelve test listeners or subjects (eight male and four female students aging from 19 to 28 years) participated in the evaluations of the original and enhanced speech. They had no prior experience in psycho-acoustic measurements and no history of hearing problems. They were seated in a soundproof booth. Signal presentation was controlled by computer. Signals were fed to listeners via a Sony-dynamic stereo headphone (MDR-CD900ST). Presentation level was individually adjusted so that perception was "loud but still comfortable" to guarantee that most signal parts were audible to the listener.

One reliable and easily implemented subjective measure is *Mean Opinion Score* (MOS). In this method, human listeners rate test speech on a five-grade scale. Since MOS introduces listener judgement bias, Hansen and Pellom suggested incorporating a subjective *Pairwise Preference Test* (PPT) [16]. In PPT, a series of pairwise randomized processed signals are presented, and listeners simply select the one they prefer. An advantage of PPT over MOS is its ease for subjects and the elimination of judgement bias [17].

We performed both MOS and PPT on overall quality. For MOS, listeners rated the speech signals on a five-grade test based on Absolute Category Rating (ACR), as shown in Table 3. The four kinds of speech signals, which were randomly arranged, were presented as one measurement block. To adjust the rating differences, listeners evaluated speech signals corrupted by different noise levels and processing artifacts at the beginning of the subjective quality assessment. For PPT, the four algorithms described in the last subsection were compared. The six comparisons were presented as one block and randomly arranged in each of these blocks. Listeners were asked to state a preference for one of the two presented algorithms.

†In [12], Porter and Boll found that for speech recognition, minimizing the mean squared errors in the log $|DFT|$ is superior to using all other DFT functions and to spectral magnitude subtraction.

Table 3 Attributes of five-point scale.

grade value	quality description
5	excellent
4	good
3	fair
2	poor
1	bad

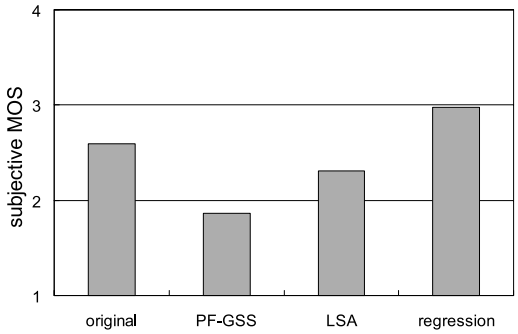


Fig. 3 Subjective MOS (averaged over seven driving conditions).

Figure 3 shows the subjective MOS results for the four algorithms averaged over the seven driving conditions. It is found that the subjective MOS of PF-GSS and LSA are lower than the original observed noisy speech. This indicates that although a significant amount of noise reduction was obtained (which can be found in Table 5), PF-GSS and LSA enhancement methods seem to decrease overall speech quality rather than to increase it because of a loss or distortion of speech components introduced. This is in line with the results of most publications (e.g., [17], [18]) on single-microphone speech enhancement schemes. Compared to PF-GSS, LSA obtained higher MOS for the less "musical tone" artifacts introduced, while the regression-based enhancement method yielded higher subjective MOS.

The PPT results are shown in Table 4. The numbers in each row, which were calculated as vote percentages, denote the preference rates of one algorithm to another algorithms. As same as MOS measure, the PF-GSS and LSA methods are not preferred over the original observed speech. Compared to PF-GSS, LSA gives higher preference scores. The regression-based enhancement method achieves significantly higher preference rates than all other algorithms, which clearly demonstrates the superiority of the proposed method.

4.3 Objective Evaluations

Since subjective measures are time-consuming and costly, objective measures, inspired by signal processing techniques, provide an efficient and economical alternative. We also performed objective evaluations of the four algorithms. The objective evaluation platform proposed by Hansen and Pellom [16] was employed, which includes the following measures: *Itakura-Saito Distortion* (ISD), *Log-Likelihood Ratio* (LLR), *Log-Area-Ratio* (LAR), *Segmental SNR* (SegSNR), and *Weighed Spectral Slope* (WSS). The WSS measure is based on an auditory model in which 36 overlapping filters of progressively larger bandwidth are used to estimate the smoothed short-time speech spectrum [16], [17]. The measure calculates a weighted difference between the spectra slopes in each band.

Table 4 Preference rates between algorithms.

	original	PF-GSS	LSA	regression
original	0	75.48%	51.67%	31.43%
PF-GSS	24.52%	0	23.10%	10.24%
LSA	48.33%	76.90%	0	25.00%
regression	68.57%	89.76%	75.00%	0

Table 5 Results of objective evaluations (averaged over seven driving conditions).

	orignal	PF-GSS	LSA	regression
ISD	1.47	0.95	1.19	0.91
LLR	0.43	0.44	0.45	0.27
LAR	4.71	4.61	4.70	3.42
SegSNR	−8.02	−6.55	−5.49	−5.54
WSS	52.55	71.58	66.50	47.57

Speech collected by a close-talking microphone (with a headset) was referred to as reference speech. To calculate each of these measures, signals were segmented in frames of 25 ms with a window shift of 10 ms. Because the mean quality measure is typically biased by a few frames in the tails of the quality measure distortion, taking the median of the frame-level is more meaningful [16]. Therefore, finding the median was used in our experiments.

Table 5 summarizes the objective evaluation measures for the four algorithms. The proposed regression-based speech enhancement method performs best in the ISD, LLR, LAR, and WSS measures except SegSNR, further evidence for its superiority. PF-GSS and LSA enhancement methods provide quality improvements over the original noisy speech in the ISD, LAR, and SegSNR measure, but not in the LLR and WSS measures. It is found that the rank order of the WSS measure is consistent with subjective MOS measure, as shown in Fig. 3.

5. Speech Recognition Experiments

In this Section and Sect. 6, we focus on improving the performance of in-car speech recognition using regression methods. Test data are extended to 50 word sets under all of the 15 real car driving conditions, as listed in Table 1. 1,000-state triphone Hidden Markov Models (HMM) with 32 Gaussian mixtures per state were used for acoustical modeling. They were trained over a total of 7,000 phonetically balanced sentences collected at the visor microphone (3,600 in the idling-normal condition, and 3,400 while driving on the streets near Nagoya university (city-normal condition)). The feature vector is a 25-dimensional vector (12 CMN-MFCC + 12 Δ CMN-MFCC +Δ log energy).

The above regression algorithms are implemented in each frequency bin mainly because they allow re-synthesis

of estimated speech, which is crucial for speech enhancement. However, for speech recognition one may directly obtain log mel-filter bank (MFB) outputs, i.e., each log MFB output of clean speech is estimated using the nonlinear regression method described in Sect. 3. A diagram of in-car regression-based speech recognition for a particular driving condition is given in Fig. 4. Once the estimated log MFB output is obtained for each mel-filter bank, the estimated log MFB vectors are transformed into mean normalized mel-frequency cepstral coefficients (CMN-MFCC) for recognition.

For comparison, we also performed recognition experiments using a linear regression method and ETSI advanced front-end [19]. In the linear regression method, no hidden layer (neurons) was used. The acoustical model used for ETSI advanced front-end experiments was trained over the training data processed with ETSI advanced front-end. The recognition performance averaged over the 15 driving conditions is given in Fig. 5. It is found that all the enhancement methods outperform the original noisy speech. LSA gives higher recognition accuracy than PF-GSS. ETSI advanced front-end very marginally outperforms LSA. Although linear regression is less effective than the conventional enhancement methods, nonlinear regression achieves the best recognition performance, outperforming ETSI advanced front-end by about 1.8%. Therefore, the nonlinear regression method is used in the following experiments.

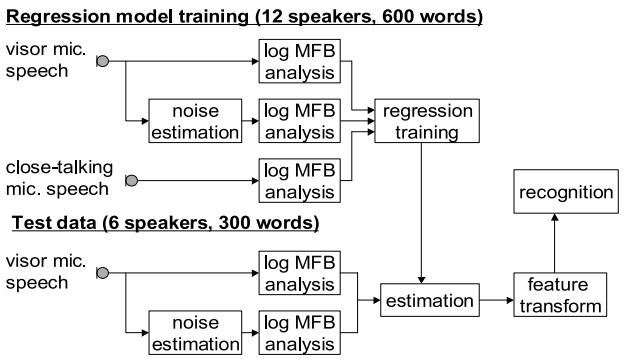


Fig. 4 Diagram of regression-based speech recognition.

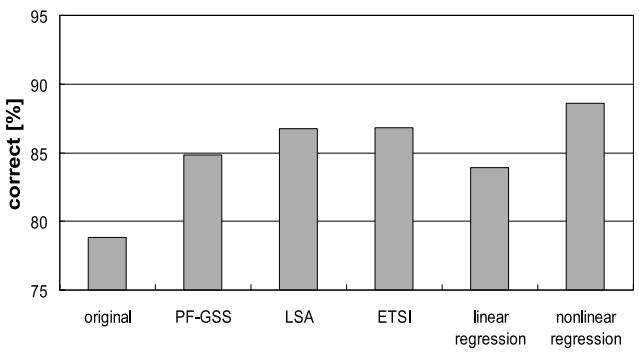


Fig. 5 Recognition performance of different speech enhancement methods (averaged over 15 driving conditions).

6. Improvements of In-Car Speech Recognition

6.1 Incorporation of Two-stage Noise Spectra Estimation

Noise spectra estimation plays an important role in speech enhancement systems. In our studies, better estimation of noise spectra is expected to improve the estimation of the log spectra of clean speech and result in higher recognition accuracy. In this subsection, a *maximum a posterior* (MAP) noise estimator is developed to combine the conventional noise estimation algorithms, i.e., after conventional noise estimation, an MAP noise amplitude estimator is employed, as shown in Fig. 6. This idea is motivated by conventional STSA speech enhancement algorithms such as [4], [20]. However, MAP estimation is not utilized to enhance the *speech* but rather to enhance the *noise*.

In the proposed estimator, we assume $x(i) = s(i) + n(i)$. By using short-time discrete Fourier transform (DFT), in the time-frequency domain we have

$$X(k,l) = S(k,l) + N(k,l),$$

where

$$X(k,l) = R(k,l)\exp\{j\varphi_x(k,l)\},$$

$$S(k,l) = A(k,l)\exp\{j\varphi_s(k,l)\},$$

$$N(k,l) = B(k,l)\exp\{j\varphi_n(k,l)\},$$

with frequency bin index k and frame index l, both of which we drop in this subsection for compactness.

The MAP noise amplitude estimator is given by

$$\hat{B} = \arg\max_{B} \; p(R|B)p(B), \tag{2}$$

where $p(\cdot)$ denotes a probability density function (pdf). Let us assume complex Gaussian models for noise and speech spectral components with variances $\lambda_n = E\{|N|^2\}$ and $\lambda_s = E\{|S|^2\}$, respectively, where $E\{\cdot\}$ denotes the expectation operator, and the variances of their real and imaginary parts are $\lambda_n/2$ and $\lambda_s/2$, respectively. We then have a Rician likelihood $p(R|B)$ and a Rayleigh prior $p(B)$ as

$$p(B) = \frac{2B}{\lambda_n}\exp\left(-\frac{B^2}{\lambda_n}\right); \tag{3}$$

$$p(R|B) = \frac{2R}{\lambda_s}\exp\left(-\frac{B^2+R^2}{\lambda_s}\right)I_0\left(\frac{2RB}{\lambda_s}\right), \tag{4}$$

where $I_0(\cdot)$ is a 0-order modified Bessel function of the first kind. Following [21], the 0-order modified Bessel function

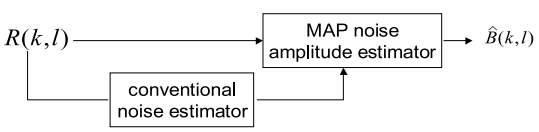


Fig. 6 Concept of two-stage noise estimator.

of the first kind can be approximated as $I_0(z) \approx e^z/\sqrt{2\pi z}$. For obtaining the noise amplitude estimator, the requirement that the gradient of $\log[p(R|B)p(B)]$ with respect to B vanishes yields

$$2\left(\frac{1}{\lambda_n}+\frac{1}{\lambda_s}\right)B-\frac{2R}{\lambda_s}-\frac{1}{2B}=0. \quad (5)$$

By solving the above equation, we can obtain

$$\hat{B}=\left(\frac{1}{2(1+\xi)}+\sqrt{\frac{1}{4(1+\xi)^2}+\frac{\xi}{4\gamma(1+\xi)}}\right)\cdot R,$$

where *a priori* and *a posteriori* SNRs are defined as $\xi = \lambda_s/\lambda_n$ and $\gamma = R^2/\lambda_n$, respectively, Here λ_n is obtained using a conventional noise estimator, and *a priori* SNR is calculated by the well-known "decision-directed" approach [4].

In our previous work [22], a two-stage noise spectra estimator (IMCRA+MAP) had been shown to yield lower estimation errors than conventional IMCRA estimator for different noise types (such as white, pink, car, and so on). It was also shown that when an IMCRA+MAP estimator was integrated into a speech enhancement system, higher segmental SNR and recognition accuracy were obtained. In our in-car speech experiments, the recognition performance of nonlinear regression in Fig. 5 was further improved by about 1.7% through the incorporation of the two-stage noise estimator. Therefore, the two-stage noise estimator was used in the following studies.

6.2 Environmental Adaptation

The regression-based recognition experiments described above require prior information on driving conditions. To develop a data-driven in-car recognition system, regression weights should be changed adaptively for different driving conditions. In this subsection, we develop a method that discriminates in-car environments by using the features of noise signals. The basic procedure is as follows: 1) Cluster the noise signals, i.e., short-time non-speech segments preceding utterances, into several groups. 2) For each noise group, train the optimal regression weights using the speech segments. 3) For unknown input speech, find a corresponding noise group through the non-speech segments, and perform the estimation with the optimal weights for the noise cluster. A diagram describing the environmental adaptation system is shown in Fig. 7. In our experiments, non-speech signals (preceding the utterance by 200 ms, i.e., 20 frames) were viewed as noise signals.

Clustering the noise signals can be viewed as a kind of computational auditory scene analysis (CASA) [23]. It is a nontrivial task in our experiments since the difference between driving conditions is not so significant. An important step is feature selection. In our studies, Mel-frequency cepstral coefficients (MFCC) were selected because of their good discriminating ability, even for audio classification [24], [25]. The MFCC features were extracted frame by frame, their means in one noisy signal computed,

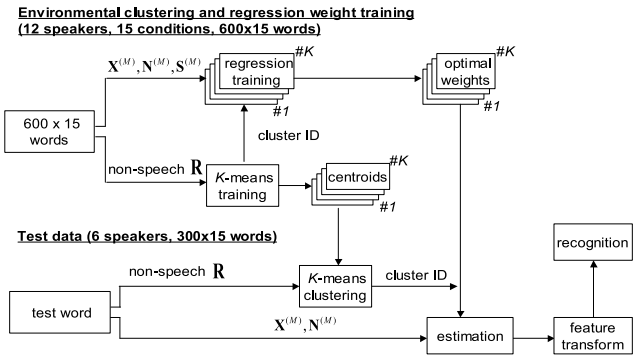


Fig. 7 Diagram of environmental clustering and regression-based speech recognition. $\mathbf{X}^{(M)}$, $\mathbf{N}^{(M)}$, and $\mathbf{S}^{(M)}$ denote the log MFB outputs obtained from observed noisy speech, estimated noise, and reference clean speech, respectively. $\mathbf{R}$ denotes the vector representation of driving environment using Eq. (6).

and then concatenated into a feature vector to represent the driving environment:

$$\mathbf{R} = [\overline{C_1}, \ldots, \overline{C_{12}}, \overline{E}], \quad (6)$$

where C_i and E denote i-order MFCC and log energy, respectively. The upper bar denotes the mean values of the features. Since the variances among $\overline{C_1}, \ldots, \overline{C_{12}}$ and $\overline{E}$ are different, all of the elements in $\mathbf{R}$ are normalized so that their mean and variance across all of the noise signals are 0 and 1.0, respectively. Prototypes of noise clusters are obtained by applying the *K-means-clustering* algorithm to the feature vectors extracted from the training set of noise signals. In our experiments, the data uttered by 12 speakers were used to cluster the noise conditions, and the data uttered by another six speakers were used for testing, as shown in Fig. 7.

Figure 8 shows the word recognition accuracies for different numbers of clusters using adaptive nonlinear regression methods. "original" and "ETSI" are cited for comparison. As seen from this figure, even the performance of 1 cluster (i.e., adaptation using universal regression weights) significantly outperforms the original noisy speech and can perform as well as ETSI advanced front-end, demonstrating the robustness of the proposed regression methods. For regression based adaptations, as the number of clusters increases up to four, the recognition accuracies consistently increase due to the availability of more noise information, while too many clusters (e.g., eight or more) yield a degradation of the recognition performance. Finally, compared to the original noisy speech and ETSI advanced front-end, we obtained relative word error rate (WER) reductions of 50.8% and 13.1%, respectively, by using the proposed regression method based on four clusters.

6.3 Discussion

In order to examine the relationship between the amount of data for learning the regression weights and the recognition performance, we preformed the following experiments (using nonlinear regression methods and with driving conditions known). The training data comprising the first one

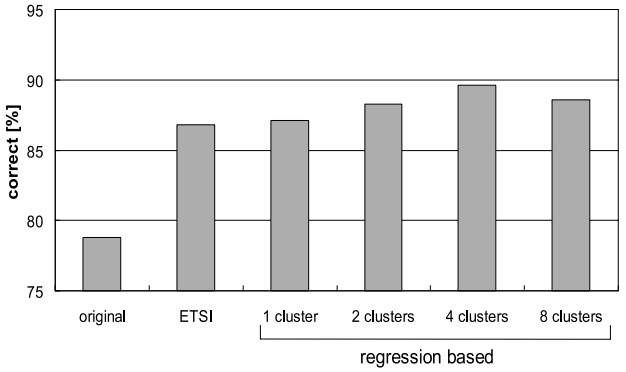


Fig. 8 Recognition performance for different cluster using adaptive regression methods (averaged over 15 driving conditions).

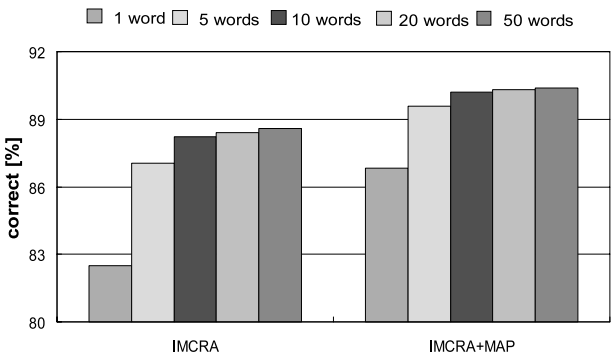


Fig. 9 Recognition performance for different amount of training data for learning the regression weights (averaged over 15 driving conditions).

word, five words, 10 words, 20 words, and all 50 words from each of 12 speakers were used for learning the regression model respectively. The recognition results are shown in Fig. 9. "IMCRA" and "IMCRA+MAP" denote the different noise spectra estimator used. From this figure, it is shown that the two-stage noise spectra estimator (IMCRA+MAP) shows its advantages over IMCRA estimator. With the amount of training data increasing from "1 word" to "5 words", the recognition performance increases significantly. However, when the amount of training data exceeds "10 words", the recognition performance is not sensitive to the amount of training data. It deserves to mention that using the training data only "1 word" with two-stage noise estimator can perform as well as ETSI advanced front-end.

We also performed the experiments to examine the recognition performance when the test words are different from the training words for the regression model. For each driving condition, the first 25 words from each of 12 speakers were used for learning the regression model, and the test data include the remaining 25 words from each of another 6 speakers. With driving conditions known and "IMCRA" estimator, the recognition performance is 88.5%, which is almost as high as those using the 25 words from back for both training and testing (88.8%).

7. Conclusions

A regression-based speech enhancement method was proposed, that approximates the log spectral of clean speech with the inputs of the log spectra of noisy speech and estimated noise. The proposed method employs statistical optimization and makes no assumptions about the independence or the distributions of the speech and noise spectra. The proposed method provided consistent improvements in our subjective evaluation of regression-enhanced speech and also performed best in most of the objective measures. The results of our studies on isolated word recognition under 15 real car driving conditions show that the proposed method outperforms conventional single-channel speech enhancement algorithms. Other methods for speech enhancement may be combined with the proposed method to obtain improved recognition accuracy in noisy environments. This method is expected to enhance recognition accuracy in very noisy situations and to be applicable to a large number of real-life environments.

Acknowledgement

This work has been partially supported by grant-in-aid #A(1)15200014 and MEXT leading project.

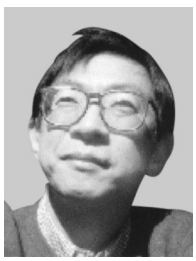
References

[1] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acoustics, Speech, and Signal Processing, vol.ASSP-27, no.2, pp.113–120, 1979.
[2] J. Deller, J. Proakis, and J.H.L. Hansen, Discrete-Time Processing of Speech Signals, Macmillan, New York, 1993.
[3] O. Cappe and J. Laroche, "Evaluation of short-time spectral attenuation techniques for the restoration of music recordings," IEEE Trans. Speech Audio Process., vol.3, no.1, pp.84–93, 1995.
[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude," IEEE Trans. Acoustics, Speech, and Signal Processing, vol.ASSP-32, no.6, pp.1109–1121, 1984.
[5] R. Martin, "Speech enhancement using MMSE short time spectral estimation with gamma distributed speech priors," Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, pp.253–256, 2002.
[6] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," IEEE Trans. Acoustics, Speech, and Signal Processing, vol.ASSP-33, no.2, pp.443–445, 1985.
[7] W. Li, T. Shinde, H. Fujimura, C. Miyajima, T. Nishino, K. Itou, K. Takeda, and F. Itakura, "Multiple regression of log spectra for in-car speech recognition using multiple distributed microphones," IEICE Trans. Inf. & Syst., vol.E88-D, no.3, pp.384–390, March 2005.
[8] W. Li, K. Itou, K. Takeda, and F. Itakura, "Optimizing regression for in-car speech recognition using multiple distributed microphones," Proc. International Conference on Spoken Language Processing, pp.2689–2692, 2004.
[9] N. Kawaguchi, S. Matsubara, H. Iwa, S. Kajita, K. Takeda, F. Itakura, and Y. Inagaki, "Construction of speech corpus in moving car environment," Proc. International Conference on Spoken Language Processing, pp.362–365, 2000.

[10] S. Haykin, Neural Networks—A Comprehensive Foundation, Prentice-Hall, 1999.
[11] S.R. Quackenbush, T.P. Barnwell, and M.A. Clements, Objective Measures of Speech Quality, Prentice-Hall, 1988.
[12] J.E. Porter and S.F. Boll, "Optimal estimators for spectral restoration of noisy speech," Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, pp.18.A.2.1–18.A.2.4, 1984.
[13] F. Xie and D.V. Compernolle, "Speech enhancement by spectral magnitude estimation—A unifying approach," Speech Commun., vol.19, pp.89–104, 1996.
[14] B.L. Sim, Y.C. Tong, J.S. Chang, and C.T. Tan, "A parametric formulation of the generalized spectral subtraction method," IEEE Trans. Speech Audio Process., vol.6, no.4, pp.328–337, 1998.
[15] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," IEEE Trans. Speech Audio Process., vol.11, no.5, pp.466–475, 2003.
[16] J.H.L. Hansen and B.L. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," Proc. International Conference on Spoken Language Processing, pp.2819–2822, 1998.
[17] M. Marzinzik, Noise reduction schemes for digital hearing aids and their use for the hearing impaired, Ph.D. Thesis, University of Oldenburg, 2000.
[18] T. Yamada, M. Kumakura, and N. Kitawaki, "Relation between subjective/objective quality of noise reduction algorithms and speech recognition performance," IEICE Technical Report, SP2004-119, 2004.
[19] "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithm," ETSI ES 202 050 v1.1.1, 2002.
[20] P.J. Wolfe and S.J. Godsill, "Effective alternatives to the ephraim and malah suppression rule for audio speech enhancement," EURASIP Journal on Applied Signal Processing, vol.2003, no.10, pp.1043–1051, 2003.
[21] R.J. McAulay and M.L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," IEEE. Trans. Acoust. Speech Signal Process., vol.ASSP-28, no.2, pp.137–145, 1980.
[22] W. Li, K. Itou, K. Takeda, and F. Itakura, "Two-stage noise spectra estimation and regression based in-car speech recognition using single distant microphone," Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, pp.I-533–536, 2005.
[23] D.F. Rosenthal and H.G. Okuno (Eds.), Computational Auditory Scene Analysis, Lawrence Erlbaum, Mahwah, NJ, 1998.
[24] M.J. Carey, E.S. Parris, and H.L. Thomas, "A comparison of features for speech, music discrimination," Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, pp.149–152, 1999.
[25] V. Peltonen, J. Tuomi, A. Klapuri, J. Huopaniemi, and T. Sorsa, "Computational auditory scene recognition," Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, pp.1941–1944, 2002.

Weifeng Li received the B.E. degree in Mechanical Electronics at Tianjin University , China, in 1997. In 2003, he received the M.E. degree in Information Electronics at Nagoya University, Japan. Currently, he is a Ph.D. candidate at the graduate school of engineering in Nagoya University, Japan. His research interests are in the area of speech signal processing and robust speech recognition.

Katsunobu Itou received the B.E., M.E. and Ph.D. degrees in computer science from Tokyo Institute of Technology in 1988, 1990 and 1993 respectively. From 2003, he has been an associate professor at Graduate School of Information Science of the Nagoya University. His research interest is spoken language processing. He is a member of the IPSJ and ASJ.



Kazuya Takeda received the B.S. degree, the M.S. degree, and the Dr. of Engineering degree from Nagoya University, in 1983, 1985, and 1994 respectively. In 1986, he joined ATR (Advanced Telecommunication Research Laboratories), where he involved in the two major projects of speech database construction and speech synthesis system development. In 1989, he moved to KDD R & D Laboratories and participated in a project for constructing voice-activated telephone extension system. He has joined Graduate School of Nagoya University in 1995. Since 2003, he is a professor at Graduate School of Information Science at Nagoya University. He is a member of the IEEE and the ASJ.

Fumitada Itakura was born in Toyokawa near to Nagoya, in 1940. He earned undergraduate and graduate degrees at Nagoya University. In 1968, he joined NTT's Electrical Communication Laboratory in Musashino, Tokyo. He completed his Ph.D. in speech processing in 1972, writing his dissertation on "Speech Analysis and Synthesis System based on a Statistical Method." He worked on isolated word recognition in the Acoustics Research Department of Bell Labs under James Flanagan from 1973 to 1975. Between 1975 and 1981, he researched problems in speech analysis and synthesis based on the Line Spectrum Pair [LSP] method. In 1981, he was appointed as Chief of the Speech and Acoustics Research Section at NTT. He left this position in 1984 to take a professorship in communications theory and signal processing at Nagoya University. After 20 years of teaching and research at Nagoya University, he retired from Nagoya University and joined Meijo University in Nagoya. His major contributions include theoretical advances involving the application of stationary stochastic process, linear prediction, and maximum likelihood classification to speech recognition. He patented the PARCOR vocoder in 1969 the LSP in 1977. His awards include the IEEE ASSP Senior Award, 1975, an award from Japan's Ministry of Science and Technology, 1977, the 1986 Morris N. Liebmann Award (with B.S. Atal), the 1997 IEEE Signal Processing Society Award, and the IEEE third millennium medal. He is a fellow of the IEEE, a fellow of the Institute of Electronics and Communication Engineers of Japan, and a member of the Acoustical Society of Japan.