TWO-STAGE NOISE SPECTRA ESTIMATION AND REGRESSION BASED IN-CAR SPEECH RECOGNITION USING SINGLE DISTANT MICROPHONE

Weifeng Li[♯], *Katunobu Itou*[†], *Kazuya Takeda*[†] *and Fumitada Itakura*[‡]

Graduate School of Engineering[♯], Graduate School of Information Science[†], Nagoya University
Faculty of Science and Technology[‡], Meijo University
Nagoya, 464–8603 Japan

ABSTRACT

In this paper, we present a two-stage noise spectra estimation approach. After the first-stage noise estimation using the improved minima controlled recursive averaging (IMCRA) method, the second-stage noise estimation is performed by employing a maximum a posteriori (MAP) noise amplitude estimator. We also develop a regression-based speech enhance system by approximating the clean speech with the estimated noise and original noisy speech. Evaluation experiments show that the proposed two-stage noise estimation method results in lower estimation error for all test noise types. Compared to original noisy speech, the proposed regression-based approach obtains an average relative word error rate (WER) reduction of 65% in our isolated word recognition experiments conducted in 12 real car environments.

1. INTRODUCTION

Noise spectra estimation plays a fundamental role in speech enhancement and speech recognition. Conventional noise estimation methods, which are based on the explicit detection of voice activity, can be difficult in the case of varying background noise or if the signal-to-noise (SNR) is low. In [1], a number of methods which do not need any explicit voice activity detectors (VADs), such as energy clustering, Hirsch histograms, low energy envelope tracking, and so on, are excellently summarized. With picking a quantile value rather than the minima value, quantile based method [2] can be viewed as a generalization of the minimum statistics (MS) approach [3]. More recently, Cohen proposed an improved minima controlled recursive averaging (IMCRA) approach [4] which involved the use of minimum statistics and speech presence probability. On the other hand, once the estimated noise spectra are obtained, one can employ an enhancement filter to estimate the spectral amplitude (or component) of a speech signal in the second stage, by assuming an *ad hoc* statistical model for speech and noise [5] [6]. In this paper, we estimate the spectral amplitude (or component) of the noise signal in a similar manner to that used in speech spectral estimation in the second stage. Therefore, a two-stage noise spectra estimation is developed. In light of the statistical information for short-time spectral amplitude (or component), the second-stage noise estimation can be expected to yield a further improvement of estimation performance. In this paper, specifically, we develop a second-stage *maximum a posteriori* (MAP) noise amplitude estimator based on first-stage IMCRA noise estimation. However, the methods used in the first stage and second stage are not limited, and can be extended to other types of first-stage and second-stage noise estimators. The finally estimated noise spectra can be further integrated into a speech enhance system.

Among a variety of speech enhancement methods, *spectral subtraction* (SS) [7] based methods and *short-time spectral estimation* (STSE) based methods [5] [8] [6] are commonly applied. Most of SS based methods make assumptions about the uncorrelation of the speech and noise spectra, while the STSE based methods requre the assumptions about an *ad hoc* statistical model for speech and noise. On the other hand, some feature mapping have been implemented through look up tables [9], curve fitting [10] and neural networks [11] [12] [13]. In the neural netwworks based feature mapping methods, the assumptions embedded in the SS and SETE methods can be released. The approach described in this paper uses neural networks to approximate the log spectral of clean speech with the inputs of the log spectra of the noisy speech and estimated noise. While other neural network based enhancement or compensation methods are implemented in time domain [11] or in cepstrum domain [12], the proposed method is a minimum mean square error (MMSE) estimator in the log spectral domain, since MMSE criterion in the log domain is more consistent with the human auditory system and distance metrics used in speech recognition system [13]. The proposed method differs from [13] in that we employ more general regression model with less input parameters. While the previous works are usually evaluated on the simulated noisy data, i.e., by artificially adding the noise to the clean speech, the proposed approach is evaluated using realistic in-car stereo data in 12 car environments.

The organization of this paper is as follows: In Section 2, we present the proposed algorithms including a noise amplitude estimator and the regression method. In Section 3, we evaluate the proposed two-stage noise estimation method. In Section 4, the regression-based in-car speech recognition experiments are described. In Section 5, we summarize this paper.

2. ALGORITHMS

2.1. MAP noise amplitude estimator

We assume that the noisy signal $x(i)$ is given by $s(i)+n(i)$, where $s(i)$ is the clean speech signal which is assumed to be independent of the additive noise $n(i)$. By using short-time Discrete Fourier transform (DFT), in the time-frequency domain we have

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

This work is partially supported by a Grant-in-Aid for Scientific Research (A) (15200014).

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 the frequency bin index k and the frame index l. We will drop both the frequency bin index k and the frame index l in this subsection, for compactness.

The MAP noise amplitude estimator is given by

$$\hat{B} = \text{argmax}\, p(R|B)p(B), \tag{1}$$

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

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

$$p(R|B) = \frac{2R}{\lambda_s}\exp(-\frac{B^2+R^2}{\lambda_s})I_0(\frac{2RB}{\lambda_s}), \tag{3}$$

where $I_0(z) = \frac{1}{2\pi}\int_0^{2\pi}\exp(z\cos\theta)d\theta$ is the 0-order modified Bessel function of first kind. The 0-order modified Bessel function of 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

$$2(\frac{1}{\lambda_n}+\frac{1}{\lambda_s})A - \frac{2R}{\lambda_s} - \frac{1}{2B} = 0. \tag{4}$$

Therefore, the gain function for the noise amplitude estimator can be obtained as

$$G_N = \frac{\hat{B}}{R} = \frac{1}{2(1+\xi)} + \sqrt{\left(\frac{1}{2(1+\xi)}\right)^2 + \frac{1}{4\gamma(1+\frac{1}{\xi})}}, \tag{5}$$

where the *a priori* and *a posteriori* SNRs are defined as $\xi = \lambda_s/\lambda_n$ and $\gamma = R^2/\lambda_n$ respectively [5].

2.2. Regression based enhancement

Let $s(i)$, $n(i)$ and $x(i)$ denote the reference clean speech, noise and the observed signals. (Note that it is not necessary to assume $x(i) = s(i)+n(i)$. A wide range of distortions, including nonstationary distortion, joint additive and convolutional distortion, and even nonlinear distortion can be handled.) By the application of a window function and analyzed using short-time Fourier trainsform, in the time-frequency domain we have $S(k,l)$, $\hat{N}(k,l)$ and $X(k,l)$, where k and l denote the frequency bin index and the frame index, and the hat above N denote the estimated version. After the mel-filter-bank (MFB) analysis and the log operation, we obtain $S^{(L)}(m,l)$, $X^{(L)}(m,l)$ and $\hat{N}^{(L)}(m,l)$, i.e.,

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

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

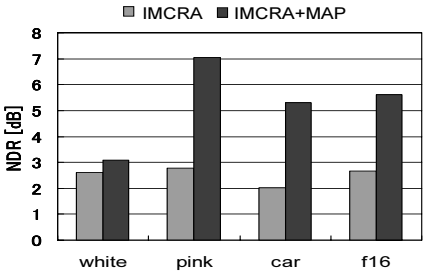

Fig. 1. Averaged NDR values for the IMCRA and the two-stage IMCRA+MAP noise estimators.

$$\hat{N}^{(L)}(m,l) = \log\sum_k r_k^m|\hat{N}(k,l)|,$$

where r_k^m denotes the weights of the mth filter bank. Let $\hat{S}^{(L)}(m,l)$ denote the estimated log MFB ouput of the mth filter bank at frame l, and it can be obtained from the inputs of $S^{(L)}(m,l)$ and $\hat{N}^{(L)}(m,l)$ by employing multi-layer perceptron (MLP) regression method, where the network with one hidden layer composed of 8 neurons is used, i.e.,

$$\hat{S}^{(L)}(m,l) = b_m + \sum_{p=1}^{8}\left(w_{m,p}\tanh\left(b_{m,p} + w_{m,p}^x X^{(L)}(m,l) + w_{m,p}^n\hat{N}^{(L)}(m,l)\right)\right) \tag{6}$$

where $\tanh(\cdot)$ is the tangent hyperbolic activation function. The parameters $\Theta = \{b_m, w_{m,p}, w_{m,p}^x, w_{m,p}^n, b_{m,p}\}$ are found by minimizing the mean squared error:

$$\mathcal{E}(m) = \sum_{l=1}^{L}[S^{(L)}(m,l) - \hat{S}^{(L)}(m,l)]^2, \tag{7}$$

through the back-propagation algorithm [14]. Here, L denotes the number of training examples.

Although both the proposed regression-based method and *log-spectra amplitude* (LSA) estimator [8] employ the MMSE cost function in the log domain, the former makes no assumptions regarding the distributions of the spectra of speech and noise. Note that the spectra of noise are estimated in the DFT domain, and then are transformed into log MFB domain as input parameters. The regression on log MFB outputs is to take into account the correlation among the neighboring frequency bins, and it results in small computation amounts, which is suitable for speech recognition.

3. EVALUATION OF NOISE ESTIMATION

The noise signals used in our evaluation are taken from the Noisex92. They include white noise, pink noise, car noise and F16 cockpit noise. The speech signals include 100 Japanese phonetically balanced sentences (10 sentences for each of 5 female speakers and 5 male speakers), which are recorded using a close-talking microphone when the car is stopped with the engine running (a part of CIAIR in-car speech corpus [15]). The speech signals are degraded by various types of noise with SNRs in the range [-5, 15] dB. Speech signals are digitized into 16 bits at a sampling frequency of 16 kHz. The spectral analysis is implemented with hamming window of 32 ms (512 samples) and a shift of 16 ms.

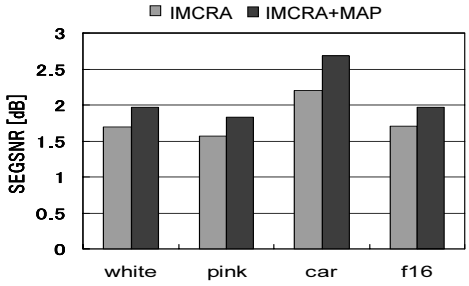

Fig. 2. Averaged segmental SNR improvement for the enhanced speech using the IMCRA and the two-stage IMCRA+MAP noise estimators.

To compute the gain function in (5), λ_n is obtained by the IMCRA method [4]. A priori SNR is calculated by the well-known "decision-directed" approach [5]. We compare the noise spectral estimation performance using the noise-to-deviation ratio (NDR), which is defined as

$$\text{NDR [dB]} = 10\log_{10}\frac{\sum_l\sum_k[\lambda_n(k,l)]^2}{\sum_l\sum_k[\lambda_n(k,l)-\hat{\lambda}_n(k,l)]^2}, \qquad (8)$$

where λ_n and $\hat{\lambda}_n$ denote the reference noise power spectral and the noise power spectral as estimated by the tested method, and L is the number of frames in the analyzed signal. Fig. 1 presents the results of NDR values averaged over [-5, 15] dB by the IMCRA and the proposed IMCRA+MAP estimators for various noise types. It shows that the latter estimator obtains significantly higher NDR values.

We also examine the performance of the proposed estimation method when integrated into a speech enhancement system. We applied a MAP speech amplitude estimator [16] for speech enhancement, in which the gain function can be obtained in a similar manner to the MAP noise amplitude and is given as

$$G_S = \frac{\hat{A}(k,l)}{R(k,l)} = \frac{1}{2(1+\frac{1}{\xi})} + \sqrt{\left(\frac{1}{2(1+\frac{1}{\xi})}\right)^2 + \frac{1}{4\gamma(1+\frac{1}{\xi})}}. \qquad (9)$$

Note that the difference between Equation (5) and Equation (9). We measure the resulting enhanced speech using segmental SNR defined as

$$\text{SegSNR [dB]} = \frac{10}{L}\sum_{l=1}^{L}\log_{10}\frac{\sum_j[s(l,j)]^2}{\sum_j[s(l,j)-\hat{s}(l,j)]^2} \qquad (10)$$

where s and $\hat{s}$ denote the reference clean speech and enhanced speech respectively. L is the number of frames in one utterance. Fig. 2 summarizes the results of the segmental SNR improvement for various noise types (averaged over [-5, 15] dB for each type). The enhanced speech obtained by using the proposed IMCRA+MAP noise estimators consistently yields a higher improvement in the segmental SNR for all noise types.

4. IN-CAR SPEECH RECOGNITION EXPERIMENTS

The speech data used is from CIAIR in-car speech corpus [15]. The speech at a close-talking microphone (recorded by wearing a headset) is referred to as clean speech. The speech captured by a microphone at the visor position is used for recognition experiments. Speech signals are digitized into 16 bits at a sampling frequency of 16 kHz. For spectral analysis, 24-channel mel-filterbank (MFB) analysis is performed on 25 millisecond-long windowed speech, with a frame shift of 10 milliseconds. Spectral components lower than 250 Hz are filtered out because the spectra of the engine noise are concentrated in the low-frequency region. Then log MFB parameters are estimated. The estimated log MFB vectors are transformed into CMN-MFCC vectors using Discrete Cosine Transformation (DCT), and then the time derivatives are calculated. The final feature vectors used in the recognition system consist of 12 CMN-MFCCs + 12 △ CMN-MFCCs + △ log energy.

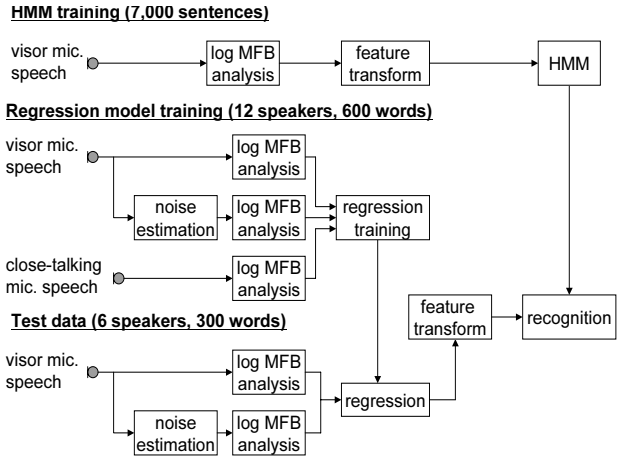

Fig. 3. Diagram of regression-based speech recognition.

We performed isolated word recognition experiments on the 50 word sets under 12 real car driving conditions (3 driving environments × 4 in-car states as listed in TABLE 1). Fig. 3 shows a block diagram of the regression-based speech recognition system. For each driving condition, the data uttered by 12 speakers was used for learning the regression weights and the remaining words uttered by 6 speakers (3 male and 3 female) were used for open testing. 1,000-state triphone Hidden Markov Modes (HMMs) with 32 Gaussian mixtures per state, trained with a total of 7,000 phonetically balanced sentences collected at the visor microphone (3,600 were collected in the idling-normal condition and 3,400 were collected while driving the DCV on the streets near Nagoya university (city-normal condition)), were used for acoustical models. We also applied the MAP speech amplitude estimator (Equation (9)) for comparison.

Fig. 4 shows the performance for recognizing different speech (averaged over 12 driving conditions). Compared with original speech, the enhanced speech using (9) provides a significant improvement compared to the original speech. The proposed regres-

Table 1. 12 driving conditions

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

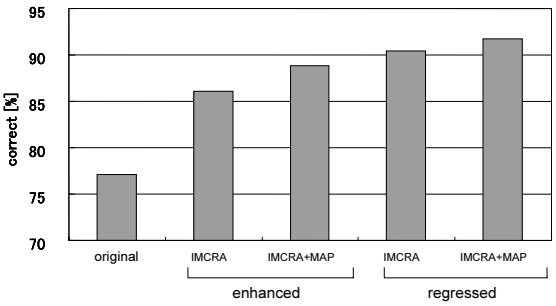

Fig. 4. Averaged word recognition performance for different speech.

sion method yields furthermore higher recogntion accuracy. Using the two-stage IMCRA+MAP noise estimator provides a improvement in recognition accuracy compared to original IMCRA noise estimator. The regression method with IMCRA+MAP noise estimator performs best and achieves an accuracy of 91.7%, which obtains an average relative word error rate (WER) reduction of 65%, compared to original noisy speech.

The effectiveness of the approximation by regression is verified from the viewpoint of the signal-to-deviation ratio (SDR), which is given by

$$\text{SDR [dB]} = 10\log_{10}\frac{\sum_l\sum_m[S^{(L)}(m,l)]^2}{\sum_l\sum_m[S^{(L)}(m,l)-\hat{S}^{(L)}(m,l)]^2}, \tag{11}$$

where $S^{(L)}(m,l)$ and $\hat{S}^{(L)}(m,l)$ denote the reference log MFB element from the close-talking microphone and the estimated log MFB element respectively. L denotes the number of frames during one utterances. The SDR values are averaged over the number of utterances. Table 2 shows the SDR values obtained using different methods. SDR values are further improved considerably by using IMCRA+MAP noise estimators, compared with the improvement achieved using IMCRA estimators. The regression method using the IMCRA+MAP noise estimator yields the highest SDR, which results in an improvement of approximately 4 dB compared with that of the original speech. These results clearly demonstrate the effectiveness of the regression method.

Table 2. Averaged SDR values for different speech.

original	enhanced		regressed	
	IMCRA	IMCRA+MAP	IMCRA	IMCRA+MAP
18.46	19.30	21.05	22.07	22.34

5. SUMMARY

In this paper, a two-stage noise spectra estimation approach and a regression-based speech enhancement approach are proposed. The second-stage enhancement-filter-like noise estimation is performed after the first-stage conventional noise estimation. In the proposed regression-based speech enhance system, the log spectra of the clean speech are approximated by using those of the estimated noise and the original noisy speech. Lower estimation errors are obtained by using the proposed two-stage noise estimation method. Use of the regression-based method results in a significant improvement in recognition accuracy.

6. REFERENCES

[1] C. Ris and S. Dupont, "Assessing local noise level estimation methods: application to noise robust ASR," *Speech Communication*, vol. 34, no. 1-2, pp. 141–158, 2001.

[2] V. Stahl; A. Fischer and R. Bippus, "Quantile based noise estimation for spectral subtraction and wiener filtering," in *Proc. IEEE ICASSP*, 2000, pp. 1875–1878.

[3] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, 2001.

[4] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech and Audio Processing*, vol. 11, no. 5, pp. 476–475, 2003.

[5] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[6] R. Martin, "Speech enhancement using MMSE short time spectral estimation with gamma distributed speech priors," in *Proc. IEEE ICASSP*, 2002, pp. 253–256.

[7] S. Boll, "A spectral subtraction algorithm for suppression of acoustic noise in speech," in *Proc. IEEE ICASSP*, 1979, pp. 200–203.

[8] 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. 32, no. 2, pp. 443–445, 1985.

[9] J.E. Porter and S.F. Boll, "Optimal estimators for spectral restoration of noisy speech," in *Proc. IEEE ICASSP*, 1984, pp. 18.A.2.1–18.A.2.4.

[10] F. Xie and D. V. Compernolle, "Speech enhancement by nonlinear spectral estimation — a unifying approach," in *Proc. EUROSPEECH*, 1993, pp. 617–620.

[11] S. Tamura and A. Waibel, "Noise reduction using connectionist models," in *Proc. IEEE ICASSP*, 1988, pp. 553–556.

[12] Helge B. D. Sorensen, "A cepstral noise reduction multilayer neural network," in *Proc. IEEE ICASSP*, 1991, pp. 993–996.

[13] F. Xie and D. V. Compernolle, "A family of mlp based nonlinear spectral estimators for noise reduction," in *Proc. IEEE ICASSP*, 1994, pp. 53–56.

[14] S. Haykin, *Neural Networks – A Comprehensive Foundation*, Prentice Hall, 1999.

[15] N. Kawaguchi; S. Matsubara; H. Iwa; S. Kajita; K. Takeda; F. Itakura and Y. Inagaki, "Construction of speech corpus in moving car environment," in *Proc. IEEE ICSLP*, 2000, pp. 362–365.

[16] P. J. Wolfe and S. J. Godsill, "Efficient alternatives to the ephraim and malah suppression rule for audio signal enhancement," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 10, pp. 1043–1051, 2003.