COMPARED STUDIES ON SINGLE-CHANEL DENOISING SCHEMES FOR IN-CAR SPEECH ENHANCEMENT

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

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. We compare the speech enhancement performance of proposed method to those of *spectral subtraction* (SS) and *short-time spectral attenuation* (STSA) based methods. The proposed method provides significant overall quality improvements in our subjective evaluation on the regression-enhanced speech. 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 54.2% and 16.5%, respectively, compared to original noisy speech and ETSI advanced front-end.

1. INTRODUCTION

Among a variety of speech enhancement methods, *spectral subtraction* (SS) [1] and *short-time spectral attenuation* (STSA) based methods [2] [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 deemphasis 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 [3] [4]. 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 [5] 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 [6].

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. In Section 2, we present the proposed regression-based speech enhancement algorithm. In Section 3, we present subjective evaluation experiments on regression-enhanced speech. We describe our speech recognition experiments using the proposed method in Section 4 and conclusions are drawn in Section 5.

2. 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 $N^{(L)}(k,l)$:

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

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:

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

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

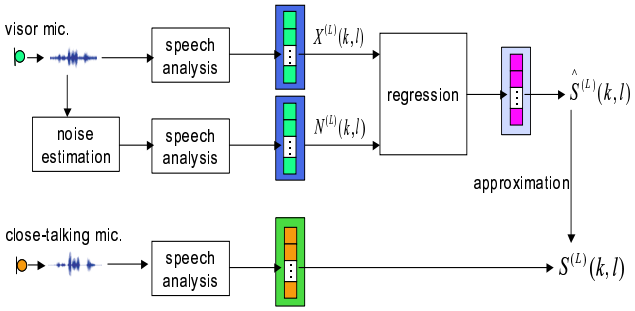

Fig. 1. Concept of regression-based speech enhancement.

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

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

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):

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

through the back-propagation algorithm [7]. 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 speech [8] and that some better results have been reported with log distortion measures [9] [10]. Although both the proposed regression-based method and *log-spectra amplitude* (LSA) estimator [4] 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 [10] 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).

3. SPEECH ENHANCEMENT PERFORMANCE

3.1. Experimental data

The speech data used are from CIAIR in-car speech corpus. 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. The test

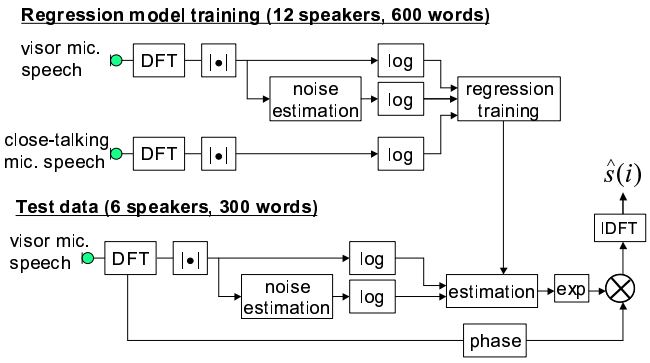

Fig. 2. Diagram of regression-based speech enhancement.

speech was based on 50 isolated word sets under seven real driving conditions listed in Table 1. Fig. 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) [11] and a *log-spectra amplitude* (LSA) estimator [4] 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. An *improved minima controlled recursive averaging* (IMCRA) method [12] 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.

3.2. Subjective evaluations

For each driving condition, five speech samples were randomly selected from the 300 test signals. The characteristics of enhanced

Table 1. 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

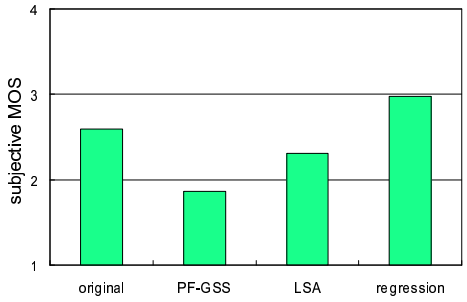

Fig. 3. Subjective MOS (averaged over seven driving conditions).

speech signals differ according to driving conditions and algorithms. Therefore, the total number of speech samples was five samples $\times$ seven driving conditions $\times$ 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 psychoacoustic 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) [13]. 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 [14].

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). 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.

Fig. 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 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., [14]) 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 2. 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.

Table 2. 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

4. SPEECH RECOGNITION EXPERIMENTS

We performed in-car speech recognition experiments using regression methods. Test data are extended to 50 word sets under all of the 15 real car driving conditions, as listed in Table 3. 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 Section 2. A diagram of in-car regression-based speech recognition for a particular driving condition is given in Figure 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 [15]. 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

Table 3. 15 driving conditions (3 driving environments $\times$ 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

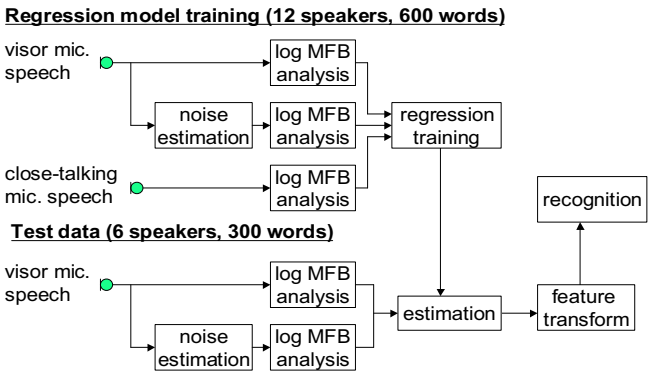

Fig. 4. Diagram of regression-based speech recognition.

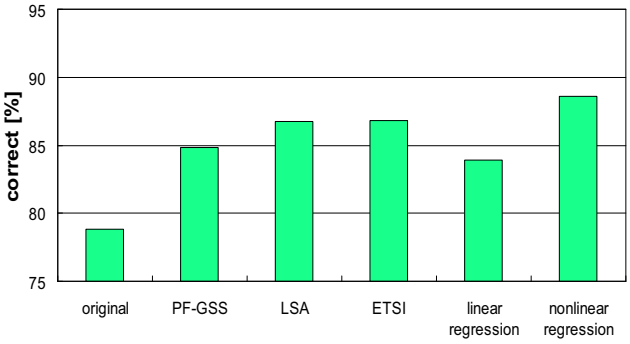

Fig. 5. Recognition performance of different speech enhancement methods (averaged over 15 driving conditions).

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%.

5. 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. 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.

6. 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] O. Cappe and J. Laroche, “Evaluation of short-time spectral attenuation techniques for the restoration of music recordings,” IEEE Trans. Speech and Audio Processing, vol.3, no.1, 1995.

[3] 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.

[4] 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.

[5] 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. on Information & Systems, E88-D, no.3, pp.384-390, 2005.

[6] 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.

[7] S. Haykin, Neural Networks - A Comprehensive Foundation, Prentice-Hall, 1999.

[8] S.R. Quackenbush, T.P. Barnwell, and M.A. Clements, Objective Measures of Speech Quality, Prentice-Hall, 1988.

[9] 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.

[10] F. Xie and D.V. Compernolle, “Speech enhancement by spectral magnitude estimation - A unifying approach,” Speech Communication, vol.19, pp.89-104, 1996.

[11] 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 and Audio Processing, vol.6, no.4, pp.328-337, 1998.

[12] I. Cohen, “Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging,” IEEE Trans. Speech and Audio Processing, vol.11, no.5, pp.466-475, 2003.

[13] 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.

[14] M. Marzinzik, Noise reduction schemes for digital hearing aids and their use for the hearing impaired, Ph.D. thesis, University of Oldenburg, 2000.

[15] “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.