Subjective and Objective Quality Assessment of Regression-enhanced Speech in Real Car Environments

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
lee@sp.m.is.nagoya-u.ac.jp

Abstract

In this paper, we propose a nonlinear regression method for speech enhancement, whose idea approximates the log spectra of clean speech with the inputs of the log spectra of noisy speech and estimated noise. We compared both subjective and objective assessments on regression-enhanced speech to those obtained through spectral subtraction (SS) and short-time spectral amplitude (STSA) methods. Our subjective evaluation experiments, which included Mean Opinion Score (MOS) and Pairwise Preference Test (PPT), show that the proposed regression-based speech enhancement method provides consistent improvements in overall quality in all seven driving conditions. The proposed method also performs the best in most objective measures.

1. Introduction

Most speech enhancement methods are based on *spectral subtraction* (SS) [1] and *short-time spectral amplitude* (STSA) analysis [2] [3]. Although modifications to SS have been proposed to reduce "musical tone" artifacts, most make assumptions about the independence of the speech and noise spectra, which is not true in many cases. A STSA-based method makes assumptions about the distributions of the speech and noise spectra and requires the estimation of *a priori* SNR. In a previous work, we proposed a regression-based enhancement method for in-car speech recognition [4]. In the proposed method, log mel-filter bank (MFB) outputs of clean speech are approximated by using those of the estimated noise and the original noisy speech. It employs statistical optimization and avoids assumptions about the independence and distributions of speech and noise spectra. In this work, the log spectra of clean speech are approximated through a regression method for each frequency bin that generates enhanced speech signals. In this paper, we present our subjective and objective studies on the regression-enhanced speech in comparison to those obtained through SS and STSA enhancement methods.

The two categories employed in the assessment of speech quality are subjective and objective measures. Subjective measures usually focus on speech intelligibility and overall quality. One reliable and easily implemented subjective measure is *Mean Opinion Score* (MOS). In this method, human listeners rate the test speech on a five-grade scale. Since MOS introduces listener judgement bias, Hansen and Pellom suggested incorporating a subjective *Pairwise Preference Test* (PPT) [5]. 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 the subjects and the elimination of judgement bias [6]. Although subjective measures are more accurate and robust, they are time-consuming and costly. On the contrary, objective measures, inspired by signal processing techniques, provide an efficient and economical alternative. Traditionally objective measures have been used to evaluate speech quality in the areas of speech coding or synthesis. Recently, some pioneers have developed a few system protocols to apply objective speech quality assessment to enhanced speech analysis.

While most literatures [5][6][7] perform subjective/objective evaluation of the enhancement methods using the simulated noisy data, i.e., by artificially adding noise to clean speech, in our studies evaluations are carried out using realistic in-car stereo data under seven driving environments.

The organization of this paper is as follows: In Section 2, we present our proposed regression-based speech enhancement algorithm. In Section 3, the experimental data used for evaluation is described. In Sections 4 and 5 we present subjective and objective evaluation experiments, respectively. Finally, Section 6 summarizes this paper.

2. Regression-based speech enhancement

Let $s(i)$, $n(i)$, and $x(i)$ respectively denote the reference clean speech, noise, and observed signals [1]. By applying a window function and analysis using short-time Fourier transform (STFT), in the time-frequency domain we have $S(k,l)$, $\hat{N}(k,l)$, and $X(k,l)$, where k and l denote frequency bin and frame indexes, and the hat above N denotes the estimated version. After determining the log operation of the amplitude, we obtain $S^{(L)}(k,l)$, $X^{(L)}(k,l)$, and $\hat{N}^{(L)}(k,l)$:

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

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

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

Let $\hat{S}^{(L)}(k,l)$ denote the estimated version obtained from the inputs of $S^{(L)}(k,l)$ and $\hat{N}^{(L)}(k,l)$ by employing a *multi-layer*

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

[1] Note that it is unnecessary to assume $x(i) = s(i) + n(i)$. A wide range of distortions, including non-stationary distortion, joint additive and convolutional distortion, and even nonlinear distortion can be handled.

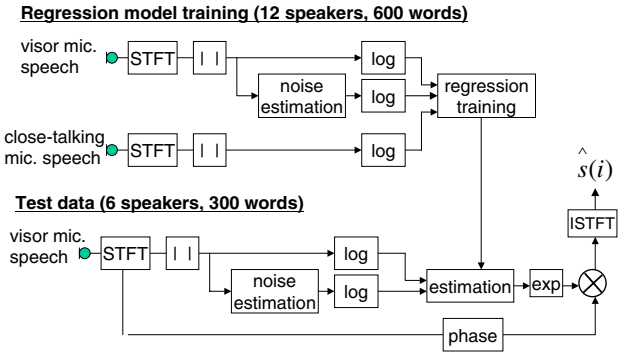

Figure 1: Diagram of regression-based speech enhancement.

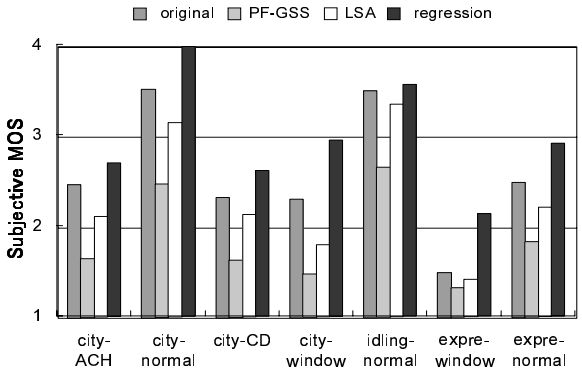

Figure 2: Subjective MOS for each driving condition.

perceptron (MLP) regression method, where a network with one hidden layer composed of eight neurons is used:

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

where

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

and $\tanh(\cdot)$ is the tangent hyperbolic activation function. The parameters $\Theta = \{b_k, w_{k,p}, w^x_{k,p}, w^n_{k,p}, b_{k,p}\}$ are found by minimizing the mean squared error:

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

through the back-propagation algorithm [8]. Here, N denotes the number of training examples (frames). Once $S^{(L)}(k,l)$ is obtained for each frequency bin, enhanced speech can be generated by taking the exponential operation and performing inverse short-time Fourier transform (ISTFT) with the combination of the phase of the observed noisy speech.

Although both the proposed regression-based method and *log-spectra amplitude* (LSA) estimator [3] employ the MMSE cost function in the log domain, the former makes no assumptions regarding the distributions of the speech and noise spectra . Note that noise spectra are estimated in the DFT domain and then transformed into log domain as input parameters.

3. Experimental data

The speech data used is 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. Speech signals are digitized into 16 bits at a sampling frequency of 16 kHz.

The test speech was based on 50 isolated word sets under seven real driving conditions, as listed in Table 1. Figure 1 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 was used for learning the regression weights, and the remaining words from six speakers (three male and three female) were used for open testing.

For comparison, a *parametric formulation of the generalized spectral subtraction* (PF-GSS) [10] and a *log-spectra amplitude* (LSA) estimator [3] were also applied. For PF-GSS, the version with constraint, which was suggested by the authors, was used. The *a priori* SNR was calculated by the well-known "decision-directed" approach [2]. An *Improved Minima Controlled Recursive Averaging* (IMCRA) method [11] was used to estimate the noise for all the enhanced methods. We selected PF-GSS and LSA because they can provide good noise reduction and overcome 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.

4. Subjective evaluation

For each driving condition, five speech samples were randomly selected from the 300 test signals, as shown in Figure 1. The characteristics of enhanced speech signals differ according to the 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 evaluation of the original and enhanced speech. They had no prior

Table 1: Seven driving conditions

driving environment	in-car state
city driving	Air-Conditioner (AC) at High level (ACH)
city driving	normal
city driving	CD player on
city driving	driver window open
idling	normal
expressway driving	driver window open
expressway driving	normal

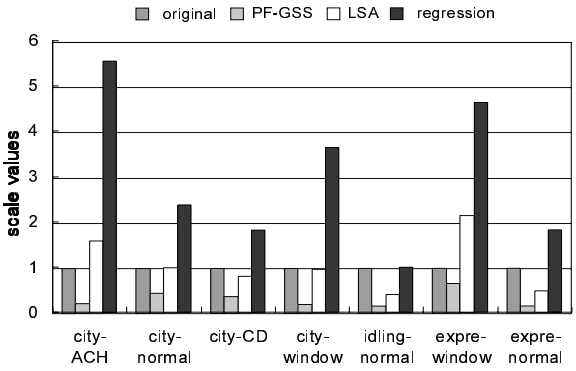

Figure 3: BTL scale value for each driving condtion.

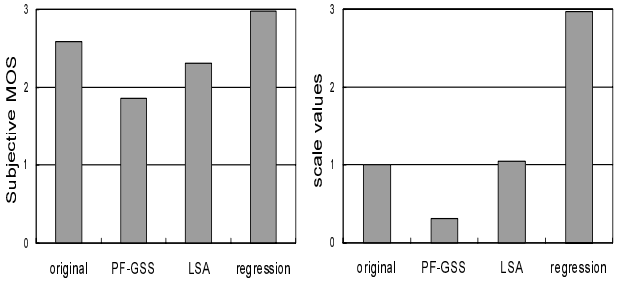

Figure 4: Overall subjective performance averaged over seven driving conditions (Left: MOS; Right: PPT).

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 for the listener.

We performed both *Mean Opinion Score* (MOS) and *Pairwise Preference Test* (PPT) on overall quality. For MOS, listeners rated the speech signals on a five-grade test based on the Absolute Category Rating (ACR). 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 section were compared. The six comparisons were presented as one measurement block. The comparisons were randomly arranged in each of these blocks. Listeners were asked to state a preference for one of the two presented algorithms. For analysis of the paired comparison data, the *Bradley-Terry-Luce* (BTL) model [12] was used. Fitting a BTL model to the paired comparison data results in a scale value for each algorithm. A matlab function to estimate choice model parameters from paired comparison data, well developed by Wichelmaier and Schmid [13], was used to analyze our paired comparison data. In our experiments, the original noisy speech (i.e., no processing) was assigned a value of one. Hence, scale values larger than one denote that an algorithm is judged better than no processing.

Figures 2 and 3 show the results of subjective MOS and PPT for each driving condition, respectively. Figure 4 summarizes the MOS and PPT measures for the four algorithms averaged over the seven driving conditions. From Figure 2, we can see that the subjective MOS is different across driving conditions and MOS is lower when signal-to-noise ratio (SNR) is low (e.g., expressway-window open). The subjective MOS of PF-GSS and LSA are lower than those of the original observed speech signals (with no processing). This indicates that although a significant amount of noise reduction was obtained, PF-GSS and LSA enhancement methods seem to decrease overall speech quality rather than to increase it, meanwhile leading to a loss or distortion of speech components. This is in line with the results of most publications (e.g., [6] [14]) on single microphone speech enhancement schemes. Compared to PF-GSS, LSA obtained higher MOS for the less "musical tone" artifacts introduced. On the other hand, it is found that the regression-based enhancement method yielded higher subjective MOS consistently across all seven driving conditions. Moreover, in some driving conditions, such as expressway-window open where SNR is very low, the regression-based enhancement method significantly outperformed the three algorithms. This clearly demonstrates the superiority of the proposed method.

From the PPT results in Figure 3, it is found that the regression-based enhancement method still achieves the highest scale values across driving conditions, This is even significant for city-ACH, city-window open, and expressway-window open conditions, where SNRs are low. The PF-GSS speech enhancement method is not preferred over the original observed speech in any of the seven driving conditions, perhaps due to the rather annoying "musical tone" artifacts introduced. Compared to PF-GSS, LSA gives higher scale values. However, the preference of LSA over the original speech apparently depends on background noise (driving conditions). In some driving conditions, such as city-ACH and expressway-window open where SNR is low, LSA is preferred over the original speech while not in other driving conditions. This may be explained by the tradeoff of noise reduction and speech distortion.

5. Objective Evaluation

We also performed objective evaluations of the four algorithms. The objective evaluation platform proposed by Hansen and Pellom [5] 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 speech collected by a close-talking microphone (with a headset) is referred to as reference speech. To calculate each of these measures, signals are sampled with 16 kHz and 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 [5]. Therefore, finding the median was used in our experiments.

Besides these measures, we also evaluated the four algorithms via the *Signal-to-Deviation Ratio* (SDR) measure and speech recognition experiments. SDR is defined as

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

where $S^{(L)}(m,l)$ and $\hat{S}^{(L)}(m,l)$ denote the reference log mel-

Table 2: Results of objective evluation (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
SDR	18.28	20.61	20.84	22.10
WER	20.80	14.84	12.51	11.57

filter bank (MFB) element from the close-talking microphone and the estimated log MFB element, respectively. m and l denote the filter bank and frame indexes, respectively. For MFB analysis, a 24 channel mel-filter bank (MFB) 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 engine noise are concentrated in the low-frequency region. 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 experiments consist of 12 CMN-MFCCs + 12 $\triangle$ CMN-MFCCs + $\triangle$ log energy. 1,000-state triphone Hidden Markov Models (HMM) with 32 Gaussian mixtures per state, trained over 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 on the streets near Nagoya university (city-normal condition)), were used for acoustical modeling.

Table 2 summarizes the objective evaluation measures for the four algorithms. Evaluation values are averaged over the seven driving conditions, as shown in Table 1. We see that the proposed regression-based speech enhancement method consistently performs best in the ISD, LLR, LAR, WSS, SDR, and WER measures (except SegSNR), further evidence for the superiority of the proposed method. Except for the LLR and WSS measures, PF-GSS and LSA enhancement methods provide quality improvements over the original noisy speech. However, rank order is not consistent across the measures. While the rank order of the WSS measure looks more similar to the subjective MOS measure as shown in Figure 4 (left part), the rank order of SDR and WER measures are actually more alike. From this table, it is found that, compared to PF-GSS and LSA, the proposed enhancement method obtains a relative word error rate (WER) reduction of 15.58% and 4.52%, respectively.

6. Conclusions

A regression-based speech enhancement method has been proposed. The proposed method employs statistical optimization and makes no assumptions about the independence or the distributions of the speech and noise spectra. In our subjective evaluation experiments (both MOS and PPT) conducted under seven driving conditions, the proposed method provided consistent improvements in overall quality. The proposed method also performed best in most of the objective measures.

7. References

[1] Boll S. F., "Suppression of acoustic noise in speech using spectral subtraction", IEEE Trans. Acoustics, Speech and Signal Processing, ASSP-27(2):113-120, 1979.

[2] Ephraim Y. and Malah D., "Speech enhancement using a minimum mean-square error short-time spectral amplitude", IEEE Trans. Acoustics, Speech and Signal Processing, ASSP-32(6):1109-1121, 1984.

[3] Ephraim Y. and Malah D., "Speech enhancement using a minimum mean-sqaure error log-spectral amplitude estimator", IEEE Trans. Acoustics, Speech and Signal Processing, ASSP-33(2):443-445, 1985.

[4] Li W., Itou K., Takeda K., and Itakura F., "Two-stage noise spectra estimation and regression based in-car speech recognition using single distant microphone", in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, pp. I-533-536, 2005.

[5] Hansen J. H.L., and Pellom B. L., "An effective quality evaluation protocol for speech enhancement algorithms", in Proc. International Conference on Spoken Language Processing, pp. 2819-2822, 1998.

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

[7] Kato M., Serizawa M., and Toki N., "Noise suppression with high speech quality based on weighted noise estimation for 3G Handsets", NEC Research & Development, 44(4):340-347, 2003.

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

[9] Kawaguchi N., Matsubara S. Iwa H., Kajita S., Takeda K., Itakura F., and Inagaki Y. "Construction of speech corpus in moving car environment", in Proc. International Conference on Spoken Language Processing, pp. 362-365, 2000.

[10] Sim B. L., Tong Y. C., Chang J. S., and Tan C. T., "A parametric formulation of the generalized spectral subtraction method", IEEE Trans. Speech and Audio Processing, 6(4):328-337, 1998.

[11] Cohen I., "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging", IEEE Trans. Speech and Audio Processing, 11(5):466-475, 2003.

[12] Colonius, H., "Representation and uniqueness of the Bradley-Terry-Luce model for pair comparisons", British Journal of Mathematical and Statistical Psychology, 33:99-103, 1980.

[13] Wickelmaier F. and Schmid C., "A matlab function to estimate choice model parameters from paired-comparison data", Behavior Reserach Methods, Instruments, & Computers, 36(1):29-40, 2004.

[14] Yamada T., Kumakura M., and Kitawaki N., "Relation between subjective/objective quality of noise reduction algorithms and speech recognition performance", Technical Report of IEICE, SP2004-119:139-144, 2004.