Applied Acoustics 87 (2015) 205–211

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Estimation of speech absence uncertainty based on multiple linear regression analysis for speech enhancement

Jihwan Park [a], Jong-Woong Kim [a], Joon-Hyuk Chang [a,*], Yu Gwang Jin [b], Nam Soo Kim [b]

[a] School of Electronic Engineering, Hanyang University, Seoul 133-791, Republic of Korea
[b] School of Electrical Engineering and INMC, Seoul National University, Seoul 151-742, Republic of Korea

ARTICLE INFO

Article history:
Received 12 November 2013
Received in revised form 24 June 2014
Accepted 25 June 2014

Keywords:
Multiple linear regression analysis
A priori SNR
Speech absence probability

ABSTRACT

We propose a novel approach to improve the performance of speech enhancement systems by using multiple linear regression to improve the technique of estimating the speech presence uncertainty. Conventional speech enhancement techniques use a fixed ratio Q of the *a priori* probability of speech presence and speech absence, or determine the value of Q simply by comparing one particular parameter against a threshold in deriving the speech absence probability (SAP) associated with the speech presence uncertainty. To further improve the performance of the SAP, we attempt to adaptively change Q according to a linear model consisting of the regression coefficients obtained by results from multiple linear regression analysis and two principal parameters: *a priori* SNR and the ratio between the local energy of the noisy speech and its derived minimum since these parameters correlate strongly with the value of Q. Distinct values of Q for each frequency in each frame are consequently assigned in time which leads to improved tracking performance of speech absence uncertainty and thus better performance of the proposed speech enhancement compared to conventional approaches. The superiority of the proposed approach is confirmed through extensive objective and subjective evaluations under various noise conditions.

1. Introduction

Because ambient noise drastically degrades the performance of speech processing systems, emerging applications in this field are demanding increasing performance in terms of ambient noise reduction in adverse environments. For example, the mobile phone system is particularly sensitive to various ambient noise environments involving nonstationary noise and low input signal-to-noise ratio (SNR). Early approaches based on the spectral weighting rule have been developed to achieve speech enhancement. These include Wiener filtering [1], minimum mean square error (MMSE) estimation [2], soft decision estimation [3], and MMSE log-spectral amplitude criteria [5]. These approaches are further developed by using a soft decision scheme in which the speech absence probability (SAP) is derived based on the likelihood ratio test (LRT) and used for gain modification [4,6]. The SAP plays an important role on the performance of speech processing systems. In practice, the spectral gain for noise suppression is modified by the SAP, which is estimated for each frequency bin in each frame on a Fourier transform domain. Furthermore, the soft decision-based schemes have been further improved by [4] called the global soft decision. This method is performed *globally*: speech activity is determined for each frame rather than for each frequency bin, thereby providing a robust estimation of the SAP. In the soft decision-based technique, the ratio Q of the *a priori* probability of speech presence and speech absence is the crucial parameter in deriving the SAP since Q must reflect the average ratio of speech presence and absence from the initial frame until the current frame. However, in most of conventional techniques for estimating the uncertainty of speech presence, the SAP is derived using a fixed Q for all frequency components in every frame. For instance, Q was set to 1 in order to address the worst-case in which speech and noise are equally likely to occur in [1]. Also, Q was chosen as 0.2 based on the listening test as in [2], while the global soft decision method in [4] adopted 0.0625 for the value of Q.

Some previous work has considered ways to estimate and update Q. Malah et al. [6] derived an algorithm to assign distinct values of Q to different frequency bins for each frame by comparing the *a posteriori* SNR with the given threshold. However, the *a posteriori* SNR is sensitive to outliers under the time-varying noise condition. Soon et al. [7] proposed a method to update the *a priori* probability of speech absence by comparing the conditional probabilities of speech presence and speech absence. On the other hand, Cohen [8] proposed the minima-controlled recursive averaging (MCRA) approach, which is known to be the successful noise power

* Corresponding author. Tel.: +82 2 2220 0355; fax: +82 2 2291 0357.
E-mail address: jchang@hanyang.ac.kr (J.-H. Chang).

http://dx.doi.org/10.1016/j.apacoust.2014.06.017
0003-682X/

estimation due to its robustness to the type and intensity of environmental noises. In particular, the presence of speech in subbands is determined by Cohen's parameter (S_r), which is the ratio between the local energy of noisy speech and its derived minimum. This algorithm is known to be computationally efficient, but it is insensitive to temporal variation. Recently, a method to track the *a priori* probability of speech absence was devised in [9] by using S_r at the MCRA method instead of the *a posteriori* SNR in Malah's method. However, these conventional approaches did not address how to incorporate spectral variation, which characterizes the *a priori* speech evolution.

In this paper, we propose a novel approach to control Q based on multiple linear regression analysis by using the *a priori* SNR and the ratio S_r. Practically, the global soft decision-based speech enhancement is considered to be a target platform in which the SAP is derived based on Q as well as the statistical model, in which the *a priori* SNR is estimated and is used to modify the spectral gain and update the noise power. Firstly, through an in-depth linear regression analysis, we investigate the extent to which Q is correlated with the *a priori* SNR and S_r. This is achieved with the help of the Pearson's correlation coefficient test [10,11], which is known to be efficient in estimating the correlation between two variables. Secondly, in an off-line training step, we apply the method of least squares to estimate the linear model's regression coefficients of Q on two parameters: *a priori* SNR and S_r. Finally, in an on-line processing step, Q is adaptively determined and used to control the SAP depending on the values of the *a priori* SNR and S_r to improve the overall performance of the proposed speech enhancement technique over conventional alternatives. We evaluate our proposed algorithm through extensive objective and subjective quality tests, which demonstrate the algorithm's improved performance over conventional methods.

The rest of the paper is organized as follows. Section 2 gives a brief review of the techniques used for speech presence uncertainty estimation, and Section 3 presents the proposed method, which uses multiple linear regression analysis. Section 4 describes the experimental setup and results in detail; Section 5 presents conclusions.

2. Review of speech absence uncertainty estimation techniques

We first briefly review the notion of the soft decision-based method for estimating speech absence uncertainty. It is assumed that a noise signal $d(t)$ is added to a speech signal $x(t)$, with their sum being denoted as the noisy speech signal $y(t)$. By taking the discrete Fourier transform (DFT) of the noisy signal $y(t)$, we then have the following in the time-frequency domain:

$$Y(k,n) = X(k,n) + D(k,n), \tag{1}$$

where $k(=0,1,\ldots,K-1)$ is the frequency bin and n is the frame index. Given two hypotheses, H_0 and H_1 which indicate speech absence and presence, respectively, it is assumed that:

$$\begin{aligned} H_0 &: Y(k,n) = D(k,n), \\ H_1 &: Y(k,n) = X(k,n) + D(k,n). \end{aligned} \tag{2}$$

Based on the complex Gaussian probability distribution assumption of the clean speech and noise spectra, the probability density functions (PDFs) conditioned on the two hypotheses H_0 and H_1 are given by [4]

$$p(Y(k,n)|H_0) = \frac{1}{\pi\lambda_d(k,n)} \exp\left\{-\frac{|Y(k,n)|^2}{\lambda_d(k,n)}\right\}, \tag{3}$$

$$p(Y(k,n)|H_1) = \frac{1}{\pi(\lambda_x(k,n)+\lambda_d(k,n))} \exp\left\{-\frac{|Y(k,n)|^2}{\lambda_x(k,n)+\lambda_d(k,n)}\right\}, \tag{4}$$

where $\lambda_x(k,n)$ and $\lambda_d(k,n)$ denote the variances of the clean speech and noise, respectively. If the spectral component of each frequency bin is assumed to be statistically independent, the SAP $P(H_0|Y(k,n))$, which is conditioned on the current observation, is derived such that [1,4]:

$$\begin{aligned} P(H_0|Y(k,n)) &= \frac{p(Y(k,n)|H_0)P(H_0)}{p(Y(k,n))} \\ &= \frac{p(Y(k,n)|H_0)P(H_0)}{p(Y(k,n)|H_0)P(H_0)+p(Y(k,n)|H_1)P(H_1)} \\ &= \frac{1}{1+\frac{P(H_1)}{P(H_0)}\Lambda(Y(k,n))}, \end{aligned} \tag{5}$$

where $P(H_0) = 1 - P(H_1)$ is the *a priori* probability of speech absence. Substituting (3) and (4) into (5), the likelihood ratio $\Lambda(Y(k,n))$ at the kth frequency is expressed as follows [4]:

$$\Lambda(Y(k,n)) = \frac{p(Y(k,n)|H_1)}{p(Y(k,n)|H_0)} = \frac{1}{1+\xi(k,n)} \exp\left\{\frac{\gamma(k,n)\xi(k,n)}{1+\xi(k,n)}\right\}, \tag{6}$$

where

$$\xi(k,n) \equiv \frac{\lambda_x(k,n)}{\lambda_d(k,n)}, \tag{7}$$

$$\gamma(k,n) \equiv \frac{|Y(k,n)|^2}{\lambda_d(k,n)}, \tag{8}$$

where $\xi(k,n)$ and $\gamma(k,n)$ are called the *a priori* SNR and the *a posteriori* SNR, respectively. Also, $P(H_1)/P(H_0) \triangleq Q$ in (5) is defined as the ratio of the *a priori* probability of speech presence and absence [4]. By using the SAP mentioned above, the spectrum of enhanced speech signal, $\widehat{X}(k,n)$, can be obtained by applying a parametric gain to each spectral component of the noisy speech signal. Here, we employ the minimum mean square error (MMSE) estimator based on SAP as follows:

$$\widehat{X}(k,n) = (1 - P(H_0|Y(k,n)))G_{\mathrm{MMSE}}(\hat{\xi}(k,n), \hat{\gamma}(k,n))Y(k,n), \tag{9}$$

where G_{MMSE} is the gain function of the MMSE estimator given in [2,4]. Also, estimate of the *a priori* SNR $\hat{\xi}(k,n)$ and *a posteriori* SNR $\hat{\gamma}(k,n)$ are obtained by using the decision-directed method [2] with $\alpha_{DD}(=0.99)$ and long-term smoothing with $\zeta_{\lambda_d}(=0.98)$, respectively, as follows [4]:

$$\hat{\xi}(k,n) = \alpha_{DD}\frac{|\widehat{X}(k,n-1)|^2}{\lambda_d(k,n-1)} + (1-\alpha_{DD})U[\gamma(k,n)-1], \tag{10}$$

$$\hat{\gamma}(k,n) = \frac{|Y(k,n)|^2}{\hat{\lambda}_d(k,n)}, \tag{11}$$

where

$$\hat{\lambda}_d(k,n) = \zeta_{\lambda_d}\hat{\lambda}_d(k,n-1) + (1-\zeta_{\lambda_d})|Y(k,n)|^2, \tag{12}$$

when the speech signal is not present, and $U[z] = z$ if $z \geqslant 0$ and $U[z] = 0$ otherwise.

As mentioned above, some approaches assigned a fixed value of Q [1–4], but Q can be differently determined for each frequency bin in each frame in the method of Malah et al. [6] by comparing the *a posteriori* SNR with a given threshold. Also, Q can be adaptively determined by the ratio of the local energy of noisy speech and its derived minimum in [8]. Indeed, this method is inherently based on the MCRA approach, in which the decision rule for the presence of speech is derived as

$$S_r(k,n) \underset{I(k,n)=0}{\overset{I(k,n)=1}{\gtrless}} \delta, \tag{13}$$

where δ is a given threshold and $I(k,n)$ is an indicator function. $S_r(k,n)$ is actually derived by $|Y(k,n)|^2/S_{\min}(k,n)$ in which $S_{\min}=\min$

$\{|Y(k,n-L+1)|^2\,|Y(k,n-L+2)|^2\cdots|Y(k,n)|^2\}$. Using $I(k,n)$, a method to track Q is devised as in the method of Lee et al. [9] as follows:

$$Q(k,n)=\alpha_p Q(k,n-1)+(1-\alpha_p)I(k,n), \tag{14}$$

where $\alpha_p(=0.2)$ is a smoothing parameter.

3. Enhanced speech absence probability based on multiple linear regression technique

As stated in the previous section, previous work has proposed using either a fixed value of Q, or an adaptive Q according to an indicating function derived from the MCRA technique [4,8]. In our approach, we devise a novel method to find Q adaptively based on multiple linear regression analysis (MLRA) which is an outperforming technique for predicting the continuous outputs with two or more variables when given observations can be modeled by a linear function [10,12]. There exist any other modeling techniques such as Gaussian mixture regression, nonlinear regression, and deep neural network [13–15]. However, these are computationally burdensome in estimating even a few coefficients. For this reason, we select the efficient way to predict Q based on the MLRA technique. The main objective of the present study is to show that this new composite estimate for Q correlates better to the true Q which is obtained by [6] with the true noise power than the conventional Q estimators used in other speech enhancement techniques.

At first, we briefly introduce the linear regression analysis, which is an approach for modeling the relationship between a scalar measured variable Y and one or more explanatory variables X. With a single explanatory variable, the technique is called simple linear regression; in this study, however, we focus on multiple regression, which includes more than one explanatory variable. More specifically, we use multiple linear regression, which can be considered as a generalization of the linear regression that considers more than one independent variable and a specific case of general linear models formed by restricting the number of dependent variables to one. We note two parameters such as the *a priori* SNR, ξ [4], and S_r [8] as the independent variables in the MLRA method. For this, an extensive analysis such as the scatter plot was performed in order to verify that these two parameters correlate well with the true Q.

To inspect the correlation between the true Q and two independent variables, ξ and Sr, these parameters are mapped to the interval between 0 and 1 by using the linear feature scaling technique [16] which can reduce sparseness of the feature data. For this, we set the maximum of Q as 100 $((1-P(H_0))/P(H_0)=100)$ and maximum of ξ and Sr set to 30 dB in which we assume that current frame belongs to exactly speech period when the true Q is greater than 100, and ξ and Sr are greater than 30 dB. The normalized variables can be obtained by dividing the feature data by each of the maximum value. For our experiments, we prepared 96 speech files, from the NTT database [17] at which each sample includes the speech material spoken by 16 male and 16 female speakers. To simulate noisy environments, the four different noise types of babble, car, office, and street noise were added to the clean speech data, each at the SNR levels of 5, 10, and 15 dB. To derive ξ and S_r, the speech enhancement techniques in [4,8] with the true noise power were applied to these noisy sentences. Then, we conducted the Pearson's correlation coefficient test [10,11] to confirm that ξ and S_r are substantially correlated with Q. In this test, the Pearsons correlation coefficient test is quantified between the independent variables and normalized true Q by ρ coefficient indicating the degree of the linear dependence between two parameters. The value ρ ranges between 1 and -1 in which high $|\rho|$ implies higher correlation. ρ between ξ and S_r can be calculated using the following equation [10,11]:

$$\rho_\xi=\frac{\sum_i\left(Q-\overline{Q}\right)\left(\xi-\overline{\xi}\right)}{\sqrt{\sum_i\left(Q-\overline{Q}\right)^2}\sqrt{\sum_i\left(\xi-\overline{\xi}\right)^2}}, \tag{15}$$

$$\rho_{S_r}=\frac{\sum_i\left(Q-\overline{Q}\right)\left(S_r-\overline{S_r}\right)}{\sqrt{\sum_i\left(Q-\overline{Q}\right)^2}\sqrt{\sum_i\left(S_r-\overline{S_r}\right)^2}}, \tag{16}$$

where given $i(=1,2,\ldots,N)$ observations, and $\overline{Q},\overline{\xi}$, and $\overline{S_r}$ are the population mean of Q,ξ, and S_r, respectively. Using this technique, the *a priori* SNR is shown to be highly correlated with Q with $\rho_\xi=0.85$ (as in Fig. 1), which is considered to be high [10,11] and S_r is also highly correlated with Q, with $\rho_{S_r}=0.86$ (as in Fig. 2), which shows the selected two parameters can be successfully used as the independent variables for predicting Q. Based on this, supposing that there are N data points $\{Y_i,X_i\}$, where $i=1,2,\ldots,N$, we build the basic model for the simple linear regression as given by

$$Y_i=\beta_0+\beta_1X_i+\epsilon_i, \tag{17}$$

where β_0 and β_1 are the regression coefficients and ϵ_i is the error. Note that this model provides the best fit for the data points where 'best' is understood in terms of the least-squares approach, with the best fit being that which minimizes the sum of the squared residuals between the measured variable and the hypothesis from the linear regression model. Extending this equation into the multiple linear regression model [10,12], our approach for using two independent variables, ξ and S_r, can be expressed as follows:

$$Q_i=\beta_0+\beta_1\xi_i+\beta_2S_{ri}+\epsilon_i, \tag{18}$$

where β_0,β_1, and β_2 are the constant regression coefficients. The estimation of regression coefficients is performed based on the method of least squares using the sum of squares of the error term, E, which is represented by

$$E=\sum_{i=1}^{N}\epsilon_i^2=\sum_{i=1}^{N}\{Q_i-(\beta_0+\beta_1\xi_i+\beta_2S_{ri})\}^2, \tag{19}$$

where N denotes the total number of observations and the regression coefficients are obtained by differentiating E partially with respect to β_0,β_1, and β_2 and setting to zero such that

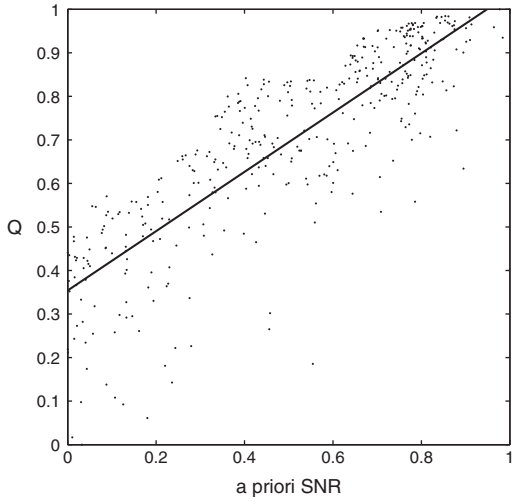

Fig. 1. Visualizing regression surfaces through a scatter plot of Q versus ξ.

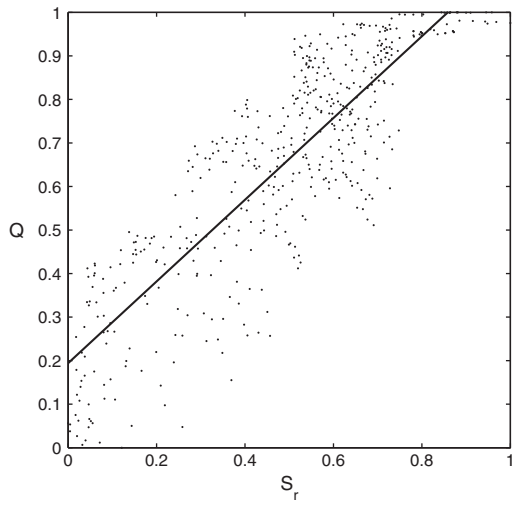

Fig. 2. Visualizing regression surfaces through a scatter plot of Q versus S_r.

$$\frac{\partial E}{\partial \beta_0} = \frac{\partial E}{\partial \beta_1} = \frac{\partial E}{\partial \beta_2} = 0. \tag{20}$$

In time, β_0, β_1, and β_2 are estimated using the solutions of the normal equations [10]. These normal equations can be expressed as a vector–matrix form as follows:

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}^{\mathrm{T}}\mathbf{Y}, \tag{21}$$

where

$$\mathbf{X} = \begin{bmatrix} 1 & \xi_1 & S_{r1} \\ 1 & \xi_2 & S_{r2} \\ \vdots & \vdots & \vdots \\ 1 & \xi_N & S_{rN} \end{bmatrix}, \tag{22}$$

and the vector form of the least square estimator of β and true Q are denoted as $\boldsymbol{\beta} = [\hat{\beta}_0\ \hat{\beta}_1\ \hat{\beta}_2]^{\mathrm{T}}$ and $\mathbf{Y} = [Q_1\ Q_2\ \cdots\ Q_N]^{\mathrm{T}}$, respectively. Since the regression coefficients are obtained from the above process, we thus obtain the estimate of $Q(k,n)$ according to the following regression equation.

$$\widehat{Q}(k,n) = \hat{\beta}_0 + \hat{\beta}_1\hat{\xi}(k,n) + \hat{\beta}_2\widehat{S}_r(k,n), \tag{23}$$

where the estimate of $Q(k,n)$ is adaptively changed according to the estimates of $\xi(k,n)$ [4] and $S_r(k,n)$ [8] which are updated by the estimated noise power. By substituting the estimate of Q to (5), the SAP can be calculated. The use of the multiple regression technique could improve the SAP performance if the estimate of Q more closely resembles the true Q. Indeed, by comparing the estimates of Q by the proposed method with the true Q, it is seen that Q obtained from the proposed algorithm indeed looks more accurate than that derived from the conventional algorithms as shown in Fig. 3. As a result, it is expected precisely estimated Q leads to increase the performance of speech enhancement.

4. Experimental results

Our performance evaluation consisted of two parts. Firstly, we compared the objective speech quality of the proposed algorithm to that of Malah's algorithm [6] and that of Cohen's algorithm

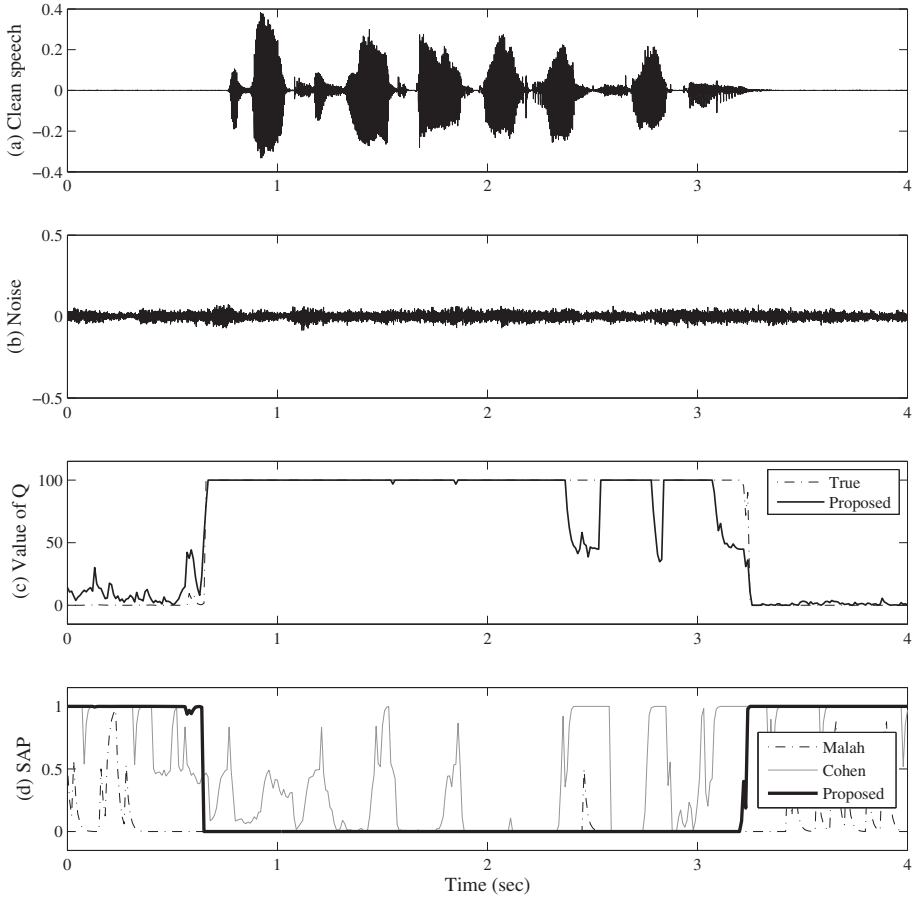

Fig. 3. Performance comparison of the Malah's method, Cohen's method, and the proposed method (babble noise, 10 dB SNR, $k = 2$). (a) Clean speech waveform. (b) Babble noise waveform. (c) Proposed estimate of Q. (d) SAP in short-time frames, Malah's method (dashed) [6], Cohen's method (solid) [8], and the proposed method (bold).

Table 1
Comparison of PESQ and C_{ovl} scores in various noise environments (95% confidence interval).

Environments		PESQ			C_{ovl}		
Noise	SNR (dB)	Malah [6]	Cohen [8]	Proposed	Malah [6]	Cohen[8]	Proposed
Babble	5	2.25 ± 0.01	2.27 ± 0.01	**2.30 ± 0.01**	2.46 ± 0.01	2.50 ± 0.01	**2.53 ± 0.01**
	10	2.55 ± 0.01	2.56 ± 0.01	**2.60 ± 0.01**	2.88 ± 0.01	2.91 ± 0.01	**2.93 ± 0.01**
	15	2.91 ± 0.01	2.93 ± 0.01	**2.97 ± 0.01**	3.42 ± 0.01	3.44 ± 0.01	**3.47 ± 0.01**
Car	5	2.52 ± 0.01	2.44 ± 0.01	**2.54 ± 0.01**	2.73 ± 0.01	2.68 ± 0.01	**2.75 ± 0.01**
	10	2.82 ± 0.01	2.75 ± 0.01	**2.84 ± 0.01**	3.13 ± 0.01	3.08 ± 0.01	**3.15 ± 0.01**
	15	3.05 ± 0.00	2.99 ± 0.00	**3.06 ± 0.00**	3.48 ± 0.01	3.44 ± 0.01	**3.50 ± 0.01**
Office	5	2.27 ± 0.01	2.21 ± 0.01	**2.27 ± 0.01**	2.54 ± 0.01	2.53 ± 0.01	**2.57 ± 0.01**
	10	2.54 ± 0.01	2.52 ± 0.01	**2.58 ± 0.01**	2.94 ± 0.02	2.93 ± 0.02	**2.96 ± 0.02**
	15	3.03 ± 0.01	2.99 ± 0.01	**3.05 ± 0.01**	3.58 ± 0.01	3.55 ± 0.01	**3.59 ± 0.01**
Street	5	2.68 ± 0.01	2.71 ± 0.01	**2.77 ± 0.01**	3.00 ± 0.01	3.03 ± 0.01	**3.07 ± 0.01**
	10	2.99 ± 0.01	3.01 ± 0.01	**3.05 ± 0.01**	3.43 ± 0.01	3.44 ± 0.01	**3.48 ± 0.01**
	15	3.30 ± 0.01	3.30 ± 0.01	**3.35 ± 0.01**	3.85 ± 0.01	3.84 ± 0.01	**3.88 ± 0.01**
Destroyer-operation	5	2.24 ± 0.01	2.22 ± 0.01	**2.25 ± 0.01**	2.46 ± 0.01	2.46 ± 0.01	**2.48 ± 0.01**
	10	2.58 ± 0.01	2.56 ± 0.01	**2.59 ± 0.01**	2.91 ± 0.01	2.90 ± 0.01	**2.92 ± 0.01**
	15	2.90 ± 0.01	2.87 ± 0.01	**2.91 ± 0.01**	3.28 ± 0.01	3.27 ± 0.01	**3.29 ± 0.01**
Factory	5	2.40 ± 0.01	2.40 ± 0.01	**2.44 ± 0.01**	2.67 ± 0.01	2.68 ± 0.01	**2.69 ± 0.01**
	10	2.73 ± 0.01	2.74 ± 0.01	**2.77 ± 0.01**	3.13 ± 0.01	3.13 ± 0.01	**3.15 ± 0.01**
	15	3.08 ± 0.01	3.08 ± 0.01	**3.16 ± 0.01**	3.61 ± 0.01	3.60 ± 0.01	**3.62 ± 0.01**

Table 2
Comparison of MOS scores in various noise environments (95% confidence interval).

Noise	Method	SNR (dB)		
		5	10	15
Babble	Malah [6]	2.08 ± 0.03	2.71 ± 0.02	2.97 ± 0.02
	Cohen [8]	2.12 ± 0.03	2.74 ± 0.03	3.03 ± 0.03
	Proposed	**2.16 ± 0.03**	**2.81 ± 0.03**	**3.11 ± 0.03**
Car	Malah [6]	2.74 ± 0.02	3.26 ± 0.02	3.79 ± 0.01
	Cohen [8]	2.66 ± 0.02	3.24 ± 0.02	3.78 ± 0.01
	Proposed	**2.83 ± 0.02**	**3.32 ± 0.02**	**3.86 ± 0.01**
Office	Malah [6]	2.41 ± 0.02	2.72 ± 0.02	3.45 ± 0.02
	Cohen [8]	2.34 ± 0.03	2.62 ± 0.02	3.38 ± 0.03
	Proposed	**2.48 ± 0.03**	**2.81 ± 0.02**	**3.54 ± 0.03**
Street	Malah [6]	2.71 ± 0.02	3.20± 0.02	3.66 ± 0.02
	Cohen [8]	2.81 ± 0.02	3.25 ± 0.02	3.73 ± 0.03
	Proposed	**2.87 ± 0.02**	**3.45 ± 0.02**	**3.93 ± 0.03**
Destroyer-operation	Malah [6]	2.02 ± 0.02	2.50 ± 0.02	2.93 ± 0.02
	Cohen [8]	2.11 ± 0.02	2.53 ± 0.02	3.01 ± 0.02
	Proposed	**2.17 ± 0.02**	**2.59 ± 0.02**	**3.08 ± 0.03**
Factory	Malah [6]	2.21 ± 0.03	2.68 ± 0.03	3.17 ± 0.02
	Cohen [8]	2.29 ± 0.03	2.80 ± 0.03	3.36 ± 0.03
	Proposed	**2.37 ± 0.03**	**2.89 ± 0.02**	**3.42 ± 0.03**

[8]. We then performed a subjective quality test and a study of speech spectrograms.

Unlike the training database, we used the speech data taken from the TIMIT database [18] for checking the sensitivity of the training data to the overall algorithm, which included 1344 phrases spoken by 112 male and 56 female speakers, with each phrase consisting of two different meaningful utterances. These files having various length from 2 s to 14 s, were collected at a sampling rate of 8 kHz, and had frames 10 ms long. To the clean speech waveforms, we added four types of noise from the NOISEX-92 database [19]: babble, car, office, and street noises, each with SNRs of 5, 10, and 15 dB. Also, we included two noises such as destroyer-operation and factory, which were not used in the training for ensuring the robust performance of the proposed algorithm. The regression coefficients applied in the proposed method were those obtained from the actual number of data $N(=40,000)$ as in (23): $\widehat{\beta}_0 = 0.248$, $\widehat{\beta}_1 = 0.479$, and $\widehat{\beta}_2 = 0.382$. Also, the experiment was implemented using the experimentally optimized parameter values from the methods of Malah [6] and Cohen [8]: $\gamma_{TH} = 0.8$, $\delta = 5$, and $L = 100$. These regression coefficients and these two parameters extracted from the noisy speech inputs were used to derive the estimate of Q in (23), allowing us to improve the previous methods for estimating SAP.

Among our evaluation methods, the first was two well-known objective tests: the ITU-T P.862 perceptual evaluation of speech quality (PESQ) [20] and the second was the composite measure (C_{ovl}) proposed by Hu and Loizou [11]. Note that C_{ovl} is known to have a strong correlation with the overall perceptual speech quality. The proposed method outperformed the conventional two algorithms in every testing condition as Table 1 summarizes the results of the PESQ and C_{ovl} tests. The performance gain was not large, but was quite consistent, showing that our proposed approach reliably improves the performance regardless of the given conditions. In particular, it is noted that the consistent

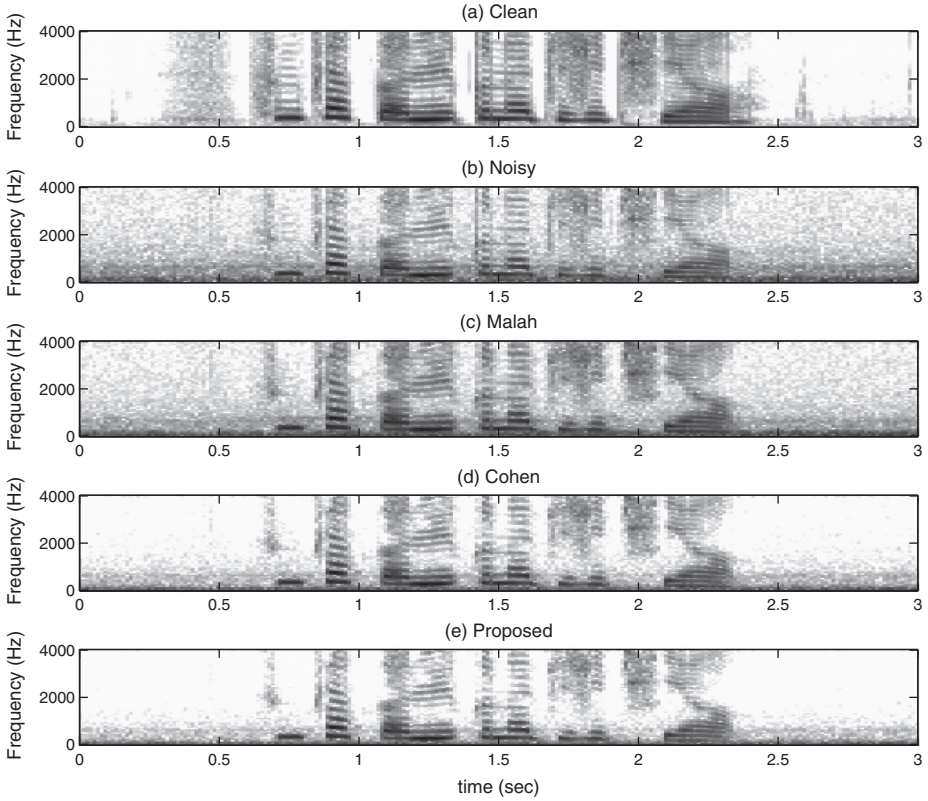

Fig. 4. Comparison of the spectrograms under the car noise (SNR = 5 dB). (a) Spectrogram of clean speech. (b) Spectrogram of noisy speech. (c) Spectrogram of Malah's algorithm [6]. (d) Spectrogram of Cohen's algorithm [8]. (e) Spectrogram of the proposed algorithm.

performance improvement was observed for the open set of noises such as destroyer-operation and factory, which confirms the robustness of our approach.

To validate the performance of the proposed algorithm, we conducted mean opinion score (MOS) tests on a number of the aforementioned noisy speech samples. Ten listeners performed listening tests and each listener gave for each test sentence a score from one to five: 5 (Excellent), 4 (Good), 3 (Fair), 2 (Poor) and 1 (Bad). All the scores were then averaged to yield the final MOS results. This score represents the listeners' global appreciation of each method's residual noise and speech distortion. As a result as in Table 2, the proposed approach scored comparably to or outperformed the conventional methods in terms of overall subjective quality under the various noise environments. This is because the proposed method reduces the distortion of speech through improved SAP, while also reducing background noise. The subjective listening test confirms the superiority of the proposed technique over the other algorithms studied.

By an evidence like an example of the spectrogram for the case of the car noise as in Fig. 4, the proposed algorithm reduces residual noise well while preserving the speech spectra. This shows that speech enhanced with the proposed method sounds more pleasant referring to the subjective MOS results in Table 2 and residual noise is perceptually more comfortable (Fig. 4c and d).

5. Conclusions

We proposed a novel approach to improve the performance of the SAP using the values of Q obtained from the MLRA technique based on the strong correlation of the selected parameters (ξ and S_r) to the value of Q. This approach enables the SAP to be estimated more accurately under various noisy conditions by adaptively changing Q based on the multiple linear regression analysis technique. In both objective and subjective quality tests over various noise types and SNR levels, the proposed method yielded performance superior to that of conventional methods.

Acknowledgements

This work was supported by NRF Grant funded by the Korean Government (MEST) (2012R1A2A2A01004895) and this research was supported by the MSIP, Korea, under the ITRC support program (NIPA-2014-H0301-14-1019) supervised by the NIPA.

References

[1] McAulary RJ, Malpass ML. Speech enhancement using a soft-decision noise suppression filter. IEEE Trans Acoust, Speech, Signal Process 1980;28(2):137–45.
[2] Ephraim Y, Malah D. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. IEEE Trans Acoust, Speech, Signal Process 1984;32(6):1109–21.
[3] Scalart P, Filho JV. Speech enhancement based on a priori signal to noise estimation. Proc IEEE Int Conf Acoust Speech Signal Process 1996:629–32.
[4] Kim NS, Chang J-H. Spectral enhancement based on global soft decision. IEEE Signal Process Lett 2000;7(5):108–10.
[5] Ephraim Y, Malah D. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Trans Acoust, Speech, Signal Process 1985;33(2):443–5.
[6] Malah D, Cox RV, Accardi AJ. Tracking speech-presence uncertainty to improve speech enhancement in nonstationary noise environments. Proc IEEE Int Conf Acoust Speech Signal Process 1999:789–92.
[7] Soon I, Koh S, Yeo C. Improved noise suppression filter using self-adaptive estimator of probability of speech absence. Signal Process 1999;75(2):151–9.
[8] Cohen I, Berdugo B. Noise estimation by minima controlled recursive averaging for robust speech enhancement. IEEE Signal Process Lett 2002;9(1):12–5.
[9] Lee W, Song J-H, Chang J-H. Minima-controlled speech presence uncertainty tracking method for speech enhancement. Signal Process 2011;91(1):155–61.
[10] Draper NR, Smith H. Applied regression analysis. Wiley; 1981.
[11] Hu Y, Loizou P. Evaluation of objective quality measures for speech enhancement. IEEE Trans Audio, Speech, Lang Process 2008;16(1):229–38.

[12] Lin Z-C, Wu W-J. Multiple linear regression analysis of the overlay accuracy model. IEEE Trans Semicond Manuf 1999;12(2):229–37.
[13] Lee S, Chang J-H, Nam SW, Lim C, Rajan S, Dajani HR, et al. Oscillometric blood pressure estimation based on maximum amplitude algorithm employing Gaussian mixture regression. IEEE Trans Instrum Meas 2013;62(12):3387–9.
[14] Hahne JM, Jiang N, Rehbaum H, Farina D, Meinecke FC, Müller K-R, et al. Linear and nonlinear regression techniques for simultaneous and proportional myoelectric control. IEEE Trans Neural Syst Rehab Eng 2014;22(2):269–79.
[15] Xu Y, Dai L-R, Lee C-H. An experimental study on speech enhancement based on deep neural networks. IEEE Signal Process Lett 2014;21(1):65–8.
[16] Aksoy S, Haralick RM. Feature normalization and likelihood-based similarity measures for image retrieval. Pattern Recogn Lett 2001;22(5):563–82.
[17] Chang J-H, Kim NS, Mitra SK. Voice activity detection based on multiple statistical models. IEEE Trans Signal Process 2006;56(6):1965–76.
[18] Garofolo JS, Lamel LF, Fisher WM, Fiscus JG, Pallett DS, Dahlgren NL, et al. TIMIT acoustic–phonetic continuous speech corpus. In: Linguistic data consortium. Philadelphia (PA, USA); 1993 [Corpus LDC93S1].
[19] Varga A, Steeneken HJM. Assessment for automatic speech recognition: II. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems. Speech Commun 1993;12(3):247–51.
[20] ITU-T P.862, Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs. 2001.