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 $S_{\min}=\min$