Journal of Systems Architecture 64 (2016) 76–85

Contents lists available at ScienceDirect

Journal of Systems Architecture

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

On using multivariate polynomial regression model with spectral difference for statistical model-based speech enhancement

Soojeong Lee, Joon-Hyuk Chang*

School of Electronic Engineering, Hanyang University, 222 Wangsimni-ro, Seongdong, Seoul 133-791, Republic of Korea

ARTICLE INFO

Article history:
Received 22 July 2015
Revised 13 October 2015
Accepted 31 October 2015
Available online 10 November 2015

Keywords:
Speech enhancement
Polynomial regression
Spectral difference
Decision-directed method
Noise power estimation

ABSTRACT

In this paper, we propose a statistical model-based speech enhancement technique using a multivariate polynomial regression (MPR) based on spectral difference scheme. In the analyzing step, three principal parameters, the weighting parameter in the decision-directed (DD) method, the long-term smoothing parameter for the noise estimation, and the control parameter of the minimum gain value are estimated as optimal operating points technique by using to the spectral difference under various noise conditions. These optimal operating points, which are specific according to different spectral differences, are estimated based on the composite measure, which is a relevant criterion in terms of speech quality. Thus, we apply the MPR technique by incorporating the spectral differences as independent variables in order to estimate the optimal operating points. The MPR technique offers an effective scheme to represent complex nonlinear input-output relationship between the optimal operating points and spectral differences so that operating points can be determined according to various noise conditions in the off-line step. In the on-line speech enhancement step, different parameters are chosen on a frame-by-frame basis through the regression according to the spectral difference. The performance of the proposed method is evaluated using objective and subjective speech quality measures in various noise environments. Our experimental results show that the proposed algorithm yields better performances than conventional algorithms.

1. Introduction

Speech enhancement is crucial in various speech communication systems including robust speech recognition, mobile communication, and hearing aids due to ambient acoustic noise [1–8]. However, the performance of a speech enhancement system deteriorates in non-stationary and low signal-to-noise ratio (SNR) noise environments. Thus, speech enhancement techniques should consider accurate noise power estimation as well as SNR estimation in various non-stationary noise environments [2]. For example, in the soft decision approach, the long-term smoothed power spectrum of the background noise depending on the probability of speech absence is adopted [9]. The speech absence probability (SAP) is derived from a likelihood ratio test (LRT) and used for gain modification. The SAP based on the statistical model of speech is generally calculated with the help of *a priori* SNR, which is estimated using a non-linear recursive procedure, called the “decision-directed (DD)” approach [2]. The *a priori* SNR determined by the DD rule takes into account the current short-time frame, with a fixed weight $(1-\alpha)$ and the processing output in the previous frame, with weight α [2,9]. Note that the parameter α should be carefully chosen since it substantially controls over the trade-off between the degree of smoothing in the *a priori* SNR at noisy parts and the acceptable level of transient distortion in the signal. In contrast to the conventional DD technique, which has a fixed weight factor, the adaptive weight factor determined by the deviation of the *a posteriori* SNR was proposed in [10]. However, this estimator does not consider a variety of noise conditions depending on noise variation.

On the other hand, the aforementioned soft decision approach has been applied to the noise power estimation module using the SAP in the long-term smoothed power spectrum of the background noise [9,11–13]. In [14], the noise power estimate was updated during periods of speech absence and speech presence. Considering the estimation of noise power, most noise power estimation algorithms are packaged with parameters that can be tuned for their performance. Note that fixed value parameters are not always optimal under each noise type since the developer should choose an operating point that yields a reasonable performance in various noise environments. In fact, Krishnamurthy and Hansen proposed an environmental sniffing framework to offer an accurate estimate of the noise update for a given environment [15]. Also, a voice activity detector (VAD) employing a support vector machine (SVM)-based noise classifier was proposed by Sangwan et al. [16] to set the best operating point when

* Corresponding author. Tel.: +82 2 2220 0355; fax: +82 2 2281 9912.
E-mail address: jchang@hanyang.ac.kr (J.-H. Chang).

http://dx.doi.org/10.1016/j.sysarc.2015.10.007
1383-7621/© 2015 Elsevier B.V. All rights reserved.

tuning parameters of the VAD. Recently, the idea of utilizing acoustic environment classification for statistical model-based speech enhancement was developed in [17]. From a practical standpoint, statistical model-based speech enhancement is considered to be a target platform in which the SAP is derived based on the LRT using the DD method in order to estimate the *a priori* SNR and is used to modify the spectral gain and update the noise power [17]. However, this approach requires a training process based on acoustic environment classification, which is time-consuming and computationally expensive. In addition, there may be an instance when it is not easy to achieve an exact acoustic environment classification result. Hence, there is a need to develop a simple but efficient methodology to provide an optimal operating points in noise environments.

In this paper, we propose a novel statistical model-based speech enhancement approach using a multivariate polynomial regression (MPR) model [18–20] which employs a spectral difference-based optimal parameter [21,22]. In the analyzing stage of the noisy signal, the principal parameters of the statistical model-based speech enhancement algorithm, such as the weighting parameter in the DD method, the long-term smoothing parameter for the noise estimation, and the control parameter of the minimum gain value are uniquely determined as optimal operating points according to the ambient noise characteristic to ensure the superior performance in a given noise environment. These optimal operating points, which are specific to different spectral patterns according to the ambient noise, are estimated and prepared based on the composite measures which are regarded as a relevant objective method to measure subjective quality processed by enhancement algorithms [22]. Based on this, we apply an approach to estimate the optimal points using the MPR model [18] at which the MPR offers the best fitting a nonlinear relationship between the mean and variance of the spectral difference as independent variables and the corresponding conditional mean of the optimal points as a dependent variable. It is noted that the spectral difference as the independent variables can be used to measure the degree of noises' nonstationarity [21]. In our training step, the MPR coefficients are estimated, which represents the optimal-fitting surface between the independent and dependent variables, hence, the estimated optimal points using the MPR model fits better the optimal points than the conventional algorithms. This idea yields in an advantage against the previous method [21] in that the spectral difference is used for the MPR by incorporating the sigmoid-type technique as the mapping function.

In the test speech enhancement step, the optimal parameters are efficiently estimated by the regression coefficients for speech enhancement according to the spectral difference of the noisy signal, which is determined on frame-by-frame basis. This observation exhibits that these points can be simply applied without acoustic noise classification, i.e., the Gaussian mixture model [17]. The performances of the proposed method are evaluated by objective and subjective speech quality measures in various noise environments. The experimental results showed that the proposed method exhibits better performances when compared with conventional algorithms.

The remainder of this paper is organized as follows. The soft decision-based speech enhancement approach is reviewed in Section 2, while a speech enhancement method using the MPR technique with the spectral difference is proposed in Section 3. In Section 4, we show the results. Finally, conclusion remarks are given in Section 5.

2. Review of soft decision based-speech enhancement

Let $x(n)$ and $d(n)$ denote clean speech and uncorrelated additive noise signals, respectively. The observed noisy speech signal $y(n)$ is the sum of a clean speech signal $x(n)$ and noise $d(n)$, where n is a discrete-time index. By taking a discrete Fourier transform (DFT), we then have the following expression:

$$Y_k(t) = X_k(t) + D_k(t), \tag{1}$$

where $k(=1,2,\ldots,K)$ is the frequency bin and t is the frame index, respectively. Given two hypotheses, H_0 and H_1 which indicate speech absence and presence, respectively, it is assumed that

$$\begin{aligned} &H_0 : \text{speech absent} : Y_k(t) = D_k(t) \\ &H_1 : \text{speech present} : Y_k(t) = X_k(t) + D_k(t). \end{aligned} \tag{2}$$

Assuming that the clean speech $X_k(t)$ and additive noise $D_k(t)$ are statistically independent and noisy spectral components are characterized by zero-mean complex Gaussian distributions, the probability density functions (PDF's) conditioned on the two hypotheses of H_0 and H_1 are given by

$$p(Y_k(t)|H_0) = \frac{1}{\pi\lambda_{d,k}(t)} \exp\left\{-\frac{|Y_k(t)|^2}{\lambda_{d,k}(t)}\right\}, \tag{3}$$

$$p(Y_k(t)|H_1) = \frac{1}{\pi(\lambda_{x,k}(t)+\lambda_{d,k}(t))} \exp\left\{-\frac{|Y_k(t)|^2}{\lambda_{x,k}(t)+\lambda_{d,k}(t)}\right\}, \tag{4}$$

where $\lambda_{x,\,k}(t)$ and $\lambda_{d,\,k}(t)$ denote the variances of the clean speech and noise for the kth spectral component at the tth frame, respectively [9].

For soft decision, the global SAP (GSAP) $p(H_0|Y(t))$ conditioned on the current observations is derived such that

$$\begin{aligned} p(H_0|Y(t)) &= \frac{p(Y(t)|H_0)p(H_0)}{p(Y(t)|H_0)p(H_0) + p(Y(t)|H_1)p(H_1)} \\ &= \frac{1}{1 + \frac{p(H_1)}{p(H_0)} \prod_{k=0}^{K-1} \Lambda(Y_k(t))}, \end{aligned} \tag{5}$$

where $P(H_0) = (1 - P(H_1))$ is the *a priori* probability of speech absence. By substituting (3) and (4) into (5), the likelihood ratio $\Lambda(Y_k(t))$ at the kth frequency can be obtained as [9]:

$$\Lambda(Y_k(t)) = \frac{p(Y_k(t)|H_1)}{p(Y_k(t)|H_0)} = \frac{1}{1+\xi_k(t)} \exp\left\{\frac{\gamma_k(t)\xi_k(t)}{1+\xi_k(t)}\right\}, \tag{6}$$

where the *a posteriori* SNR $\gamma_k(t)$ and the *a priori* SNR $\xi_k(t)$ are defined by

$$\gamma_k(t) \equiv \frac{|Y_k(t)|^2}{\lambda_{d,k}(t)}, \tag{7}$$

$$\xi_k(t) \equiv \frac{\lambda_{x,k}(t)}{\lambda_{d,k}(t)}. \tag{8}$$

Also, if $\hat{\xi}_k(t)$ and $\hat{\gamma}_k(t)$ are the estimates for $\xi_k(t)$ and $\gamma_k(t)$, respectively, $\hat{\xi}_k(t)$ could be estimated using the well-known DD approach [2] as follows:

$$\hat{\xi}_k(t) \equiv \alpha_\xi \frac{|\hat{X}_k(t-1)|^2}{\hat{\lambda}_{d,k}(t-1)} + (1-\alpha_\xi)F[\hat{\gamma}_k(t) - 1] \tag{9}$$

where $\hat{X}_k(t-1)$ represents the estimated clean speech spectrum in the previous frame and $F[x] = x$ if $x \geq 0$; otherwise $F[x] = 0$. Here, $\alpha_\xi (0 \leq \alpha_\xi \leq 1)$ is a weighting factor that controls the trade-off between the noise reduction and the transient signal distortion by being chosen very close to 1 (i.e., $\alpha_\xi = 0.99$). Also, $\hat{\gamma}_k(t)$ is directly obtained from the ratio of the input power $|Y_k(t)|^2$ and the estimate of $\lambda_{d,\,k}(t)$.

Since the estimation of the noise power spectrum is a major component in speech enhancement. The soft decision method is substantially adopted for obtaining a long-term smoothed noise power spectrum of the background noise as the estimate for $\lambda_{d,\,k}(t)$ as follows [9]:

$$\hat{\lambda}_{d,k}(t+1) = \zeta_d \hat{\lambda}_{d,k}(t) + (1-\zeta_d)E[|D_k(t)|^2|Y_k(t)], \tag{10}$$