Digital Signal Processing 26 (2014) 169–182

Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

Noise variance estimation based on dual-channel phase difference for speech enhancement

Seon Man Kim [a], Hong Kook Kim [b,*]

[a] Institute of Sound and Vibration Research, University of Southampton, Southampton SO17 1BJ, UK
[b] School of Information and Communications, Gwangju Institute of Science and Technology, 1 Oryong-dong, Buk-gu, Gwangju 500-712, Republic of Korea

ARTICLE INFO

Article history:
Available online 28 November 2013

Keywords:
Dual-microphone speech enhancement
Signal-to-noise ratio (SNR)
Direction-of-arrival based SNR estimation
Noise variance estimation
Phase difference

ABSTRACT

In this paper, we propose a method for estimating a signal-to-noise ratio (SNR) in order to improve the performance of a dual-microphone speech enhancement algorithm. The proposed method is able to reliably estimate both *a priori* and *a posteriori* SNRs by exploring a direction-of-arrival (DOA)-based local SNR that is defined by using spatial cues obtained from dual-microphone signals. The estimated *a priori* and *a posteriori* SNRs are then incorporated into a Wiener filter. Consequently, it is shown from an objective perceptual evaluation of speech quality (PESQ) comparison and a subjective listening test that a speech enhancement algorithm employing the proposed SNR estimate outperforms those using conventional single- or dual-microphone speech enhancement algorithms such as the Wiener filter, beamformer, or phase error-based filter under different noise conditions ranging from 0 to 20 dB.

1. Introduction

The goal of speech enhancement is to suppress additive background noise components while maintaining the quality and intelligibility of speech [1]. This task is usually accomplished by preserving the characteristics of speech using the short-term spectral amplitude (STSA) by means of a minimum mean square error (MMSE-STSA), an MMSE-log spectral amplitude (MMSE-LSA), or a Wiener filter [1], which results in a spectral gain attenuator for speech enhancement. Actually, the spectral gain attenuator is a function of the *a priori* and *a posteriori* signal-to-noise ratios (SNRs). Thus, the accurate estimation of *a priori* and *a posteriori* SNRs is crucial for speech enhancement in noisy environments. It has been reported that a decision-directed (DD) approach provided a simple but effective estimate of the *a priori* SNR with a reasonable computational cost [1–3], where the *a priori* SNR estimate was obtained by smoothing the *a priori* SNR estimate of the previous frame and the *a posteriori* SNR estimate of the current frame. Thus, the accuracy estimate of the *a priori* SNR estimation in the DD-based approach strongly relied on that of the *a posteriori* SNR estimate that could be directly obtained as the ratio between current noisy speech power and the estimated noise variance. Consequently, inaccurate noise variance estimates may affect the accurate estimation of *a priori* and *a posteriori* SNRs, which could distort estimated clean speech in severely adverse noise environments [3,4].

In order to provide better speech enhancement performance in adverse noise environments, multi-microphone speech signals can be used instead of a single-microphone speech signal [5–11]. A multi-microphone speech enhancement algorithm attempts to utilize additional direction-of-arrival (DOA) information, as opposed to only temporal information for a single-microphone algorithm. The DOA is strongly linked to the phase difference between multi-channel signals [5–7]; thus, one of the important issues is how to effectively utilize spatial cues such as phase differences for target speech estimation within noisy speech.

As one of the multi-microphone approaches, beamforming utilizes a directional sensitivity, which is referred to as the spatial directivity pattern (SDP), and it attempts to attenuate spatially unwanted noises arising from non-target directions [12–14]. For example, a super-directive beamformer (SDB), which is one of the typical fixed beamformers, has been designed for enhancing noisy speech in a diffuse sound field [5,13,15,16]. In particular, the average word accuracy of a speech recognition system using the speech enhanced by the SDB was increased by approximately 20% [15]. On the other hand, adaptive beamformers have been designed to provide higher interference suppression capability than fixed beamformers. In particular, the generalized sidelobe canceler (GSC) [5,14] is widely used because of its simple structure and easy implementation. However, speech enhancement performance using beamforming was constrained by the number of microphones [5]. Thus, the performance using beamforming based on dual-microphone signals may not be satisfactory compared with those of masking-based methods [16].

In order to improve the performance of speech enhancement, a multi-channel speech enhancement algorithm was proposed to

* Corresponding author. Fax: +82 62 715 2204.
E-mail address: hongkook@gist.ac.kr (H.K. Kim).

1051-2004/$ – see front matter
http://dx.doi.org/10.1016/j.dsp.2013.11.012

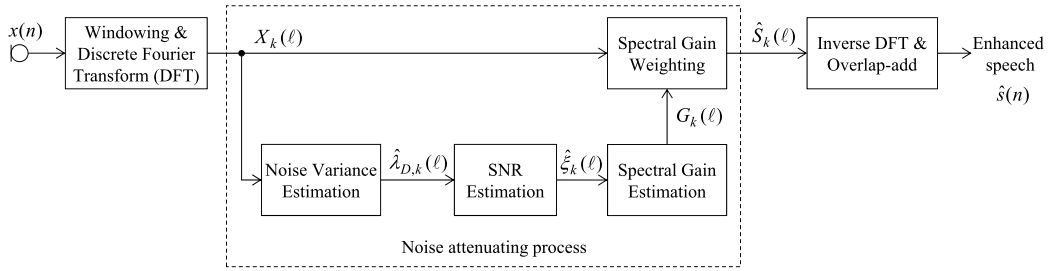

Fig. 1. Block diagram of a conventional single-microphone speech enhancement algorithm.

have a structure of a beamformer followed by a time-varying postfilter [5,17]. The time-varying postfilter was designed by utilizing the power ratio between the input and output signals of the beamformer, and it was applied to the beamformer output signal to further reduce noise components. Moreover, the phase-error based filter (PEF) was proposed in [16], where a soft-mask filter was estimated based on phase errors attributed to dual-microphone phase differences. It has been shown that the PEF provided higher digit recognition accuracy than the dual-microphone SDB [16] with or without a postfilter [17]. In addition, the PEF was effective in non-target directional noise suppression, providing better speech enhancement performance than a beamformer-based filter [16]. Recently, a statistically pre-trained model on a target speaker was employed to improve the reliability of the *a posteriori* SNR in a speech enhancement method based on a single-microphone Wiener filter [3]. Furthermore, this approach was also utilized to further improve speech quality by combining the multi-microphone spatial filter and the MMSE-LSA estimator [10]. However, this approach required a training phase for reliable estimation. An angular spectrum based masking (ASBM) method was proposed in order to provide a better sound power estimate in the target direction as well as a better residual noise power by using a diffuse noise model in each time–frequency bin [18]. As mentioned above, spatial cues play a crucial role in robustly estimating the target speech against noise conditions.

Therefore, in order to better utilize the spatial cues, we propose a method of reliably estimating both *a priori* and *a posteriori* SNRs from spatial cues. To this end, the proposed method first estimates a DOA-based local SNR using phase differences between dual-microphone signals. The estimated DOA-based local SNR, which is defined as the power ratio between the target-directional enhanced and rejected signal, is then used to obtain a spectral gain for local noise variance estimation. A global noise variance is also estimated by recursive long-term smoothing of the power spectra of background noise under uncertainty of target speech presence. Thus, the proposed noise variance estimate is obtained by a weighted sum of the estimated local and global noise variances, where a weighting parameter is empirically determined to maximize speech enhancement performance. After that, the estimated noise variance is used to estimate the *a posteriori* and *a priori* SNRs in the DD approach. Finally, the *a posteriori* and *a priori* SNR estimates are incorporated into a Wiener filter in order to obtain a spectral gain attenuator.

The remainder of this paper is organized as follows. Following this introduction, Section 2 reviews the *a priori* and *a posteriori* SNR estimation method for single- and multi-microphone speech enhancement. After that, Section 3 proposes a dual-microphone speech enhancement algorithm employing the proposed DOA-based SNR estimation method. Next, Section 4 evaluates the performance of the proposed dual-microphone speech enhancement algorithm in terms of the perceptual evaluation of speech quality (PESQ) and a subjective preference and compares it with those of conventional algorithms such as a single-microphone Wiener filter [1], SDB [5], GSC [14], PEF [16], and ASBM [18]. Finally, Section 5 summarizes this paper.

2. SNR estimation in speech enhancement

2.1. Single-microphone SNR estimation

In this subsection, we briefly review a single-microphone speech enhancement algorithm, focusing on a noise variance estimation procedure for SNR estimation. Fig. 1 shows a block diagram of a conventional speech enhancement algorithm based on STSA [1–4]. As shown in the figure, noise variance estimation is first performed, and then an SNR is estimated using the estimated noise variance. Next, a spectral gain attenuator is obtained based on the estimated SNR, which is subsequently applied to an input noisy speech signal. The detailed explanation is as follows.

Assuming that a target speech, $s(n)$, is deteriorated with additive noise, $d(n)$, the noisy speech, $x(n)$, is related to $s(n)$ and $d(n)$ in the frequency domain as

$$X_k(\ell) = S_k(\ell) + D_k(\ell) \tag{1}$$

where $X_k(\ell)$, $S_k(\ell)$, and $D_k(\ell)$ are the complex valued spectral components of $x(n)$, $s(n)$, and $d(n)$, respectively, at the k-th frequency bin ($k = 0, 1, \ldots, K-1$) and the ℓ-th frame. A speech enhancement algorithm attempts to estimate the target speech spectrum as a form of $\hat{S}_k(\ell) = G_k(\ell)X_k(\ell)$, where $G_k(\ell)$ is a spectral gain attenuator estimated from the *a priori* SNR $\xi_k(\ell)$ or both the *a priori* SNR and the *a posteriori* SNR $\gamma_k(\ell)$. For example, $G_k(\ell) = \xi_k(\ell)/(\xi_k(\ell)+1)$ for a Wiener filter, and $G_k(\ell) = \frac{\sqrt{\pi}}{2}\cdot\sqrt{\frac{\gamma_k(\ell)\xi_k(\ell)}{1+\xi_k(\ell)}}/\gamma_k(\ell)\cdot M(\frac{\gamma_k(\ell)\xi_k(\ell)}{1+\xi_k(\ell)})$ for MMSE-STSA, where $M(\kappa) = \exp(-\frac{\kappa}{2})\cdot[(1+\kappa)\cdot I_0(\frac{\kappa}{2})+\kappa\cdot I_1(\frac{\kappa}{2})]$ with the modified Bessel function of zero and first order $I_0(\cdot)$ and $I_1(\cdot)$.

In [1,19], $\xi_k(\ell)$ and $\gamma_k(\ell)$ are represented as

$$\xi_k(\ell) = \lambda_{S,k}(\ell)/\lambda_{D,k}(\ell), \tag{2}$$

$$\gamma_k(\ell) = \left|X_k(\ell)\right|^2/\lambda_{D,k}(\ell) \tag{3}$$

where $\lambda_{S,k}(\ell) = E[|S_k(\ell)|^2]$ and $\lambda_{D,k}(\ell) = E[|D_k(\ell)|^2]$ are the variances of $S_k(\ell)$ and $D_k(\ell)$, respectively. As described in (2) and (3), $\lambda_{D,k}(\ell)$ should be known in order to estimate $\xi_k(\ell)$ and $\gamma_k(\ell)$. It is usual to obtain an estimate of $\lambda_{D,k}(\ell)$, $\hat{\lambda}_{D,k}(\ell)$, based on a hypothesis testing approach [2,4]. In other words, two hypotheses are constructed to test the speech absence probability (SAP), such as H_0: $X_k(\ell) = D_k(\ell)$ and H_1: $X_k(\ell) = S_k(\ell) + D_k(\ell)$. Then, a likelihood ratio, $\Lambda_k(\ell)$, is computed by assuming that $S_k(\ell)$ and $D_k(\ell)$ follow zero-mean complex Gaussian distributions [2,4], such as

$$\begin{aligned}\Lambda_k(\ell-1) &= \frac{p(X_k(\ell-1)|H_1)}{p(X_k(\ell-1)|H_0)} \\ &= \frac{1}{1+\hat{\xi}_k(\ell-1)}\exp\left(\frac{\hat{\gamma}_k(\ell-1)\cdot\hat{\xi}_k(\ell-1)}{1+\hat{\xi}_k(\ell-1)}\right)\end{aligned} \tag{4}$$