Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 72 (2015) 96–108

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Fast distributed multichannel speech enhancement using novel frequency domain estimators of magnitude-squared spectrum ☆

Jingxian Tu [a], Youshen Xia [b,*]

[a] *Center for Discrete Mathematics and Theoretical Computer Science, Fuzhou University, 360116, China*
[b] *College of Mathematics and Computer Science, Fuzhou University, Fuzhou 360116, China*

Received 28 October 2014; received in revised form 4 May 2015; accepted 11 May 2015
Available online 19 May 2015

Abstract

This paper proposes two novel frequency domain estimators for fast distributed multichannel speech enhancement in background of white and colored noise. The proposed two frequency domain estimators are maximum a posterior (MAP) and minimum mean square error (MMSE) estimators, respectively. They significantly generalize two single channel optimal frequency domain estimators of magnitude-squared spectrum. Compared with the optimal multichannel frequency domain estimator generalizing the single channel short-time spectral amplitude and log-spectral amplitude estimators, the proposed two frequency domain estimators have a very low computational cost. Computed results show that the proposed two estimators reduce both colored background noise and speech distortion. Furthermore, the proposed two multichannel algorithms have a much faster computational speed than conventional multichannel algorithms for distributed multichannel speech enhancement.

Keywords: Multichannel speech enhancement; Amplitude and phase estimator; Minimum mean square error; Maximum a posterior

1. Introduction

Speech enhancement has been studied because of its many applications, such as voice communication, speech recognition, voiced-control systems, and the transmitted speech signals, where received speech signals are usually corrupted by white and colored noise. Over the past several decades, many of speech enhancement algorithms were presented. Speech enhancement algorithms can be classified as single channel algorithms and multichannel speech enhancement algorithms (Loizou, 2007). Single channel speech enhancement algorithms mainly include the spectral subtraction method (Boll, 1979), the wiener filtering method (Abd et al., 2008 and references therein), the subspace method (Ephraim and Van Trees, 1995; Hu and Loizou, 2002; Wei and Xia, 2013), the Kalman filtering method (Paliwal and Basu, 1987; Gibson et al., 1991; Gabrea, 2005; Bobillet et al., 2007), and the statistical approach (Lotter and Vary, 2003; Martin, 2005; Plourde and Champagne, 2011; McCallum and Guillemin, 2013). In the subspace method, the noisy signal space is first separated into two orthogonal subspaces: the noisy subspace and the signal subspace, and then signal enhancement is used to remove the noise subspace and to estimate the clean speech signal from the noisy speech subspace. In the Kalman filtering method, the speech signal is modeled as autoregressive (AR) process and the speech signal is then recovered from Kalman filter. In recent decades, multichannel speech enhancement techniques have been

☆ This work is supported by the National Natural Science Foundation of China, China under Grant Nos. 61179037 and 61473330.

* Corresponding author. Tel.: +86 591 22865171.
E-mail addresses: 568205127@qq.com (J. Tu), ysxia@fzu.edu.cn (Y. Xia).

http://dx.doi.org/10.1016/j.specom.2015.05.004
0167-6393/© 2015 Elsevier B.V. All rights reserved.

abstractly studied as the multi-microphone system was introduced. The multichannel microphone can utilize much more information to improve the performance of speech enhancement. Moreover, by using spatial information of signals, the multi-microphone system can exhibit better performance than the single-microphone system (Benesty et al., 2005; Vary and Martin, 2006; Xia, 2012). The multichannel speech enhancement techniques include mainly the standard beamforming technique (Benesty et al., 2007; Cox et al., 1987), the multi-channel Wiener filtering technique (Huang et al., 2008 and references therein), the multichannel subspace technique (Jabloun and Champagne, 2001; Doclo and Moonen, 2002; Rombouts and Moonen, 2003; Doclo and Moonen, 2005), the space-temporal prediction (STP)-based multichannel subspace technique (Chen et al., 2008), and the statistical approach-based multichannel technique (Hendriks et al., 2009).

Recently, Trawicki and Johnson proposed two optimal frequency domain estimators for distributed multichannel speech enhancement (Trawicki and Johnson, 2012). The two optimal estimators generalize the single channel short-time spectral amplitude (STSA) and log-spectral amplitude (LSA) estimators by Ephraim and Malah (1984) and Ephraim and Malah (1985), respectively. A closed form solution of the spectral phase was given, however, there exist two computational costs for implementation. One is to compute two bessel functions which are expressed by the Taylor series expansion. Another is to compute the closed-form solution. They result in a slow speed.

For fast distributed multichannel speech enhancement in background of colored noise, we propose two novel frequency domain estimators, called maximum a posterior (MAP) and minimum mean square error (MMSE) estimators, by significantly generalizing two single channel optimal frequency domain estimators of magnitude-squared spectrum (Lu and Loizou, 2011). Compared with the optimal multichannel frequency domain estimator (Trawicki and Johnson, 2012), the proposed two frequency domain estimators have a low computational cost for implementation. Simulation results show that the proposed two multichannel algorithms can achieve faster and higher noise reduction than conventional multichannel algorithms.

This paper is organized as follows. Section 2 describes multichannel models and existing frequency domain estimators. Section 3 introduces two conventional single channel estimators and proposes two optimal multichannel frequency domain estimators. Section 4 presents performance evaluation, and Section 5 gives the conclusion.

2. System model and estimator

We are concerned with a distributed microphone system which can accurately time align the M noisy observations (Trawicki and Johnson, 2012). The time domain multichannel microphone model is described as

$$y_i(t) = c_i s(t) + n_i(t), \quad i = 1, 2, \ldots, M \tag{1}$$

where M is the number of channels, $y_i(t)$ and $n_i(t)$ are the noisy speech and noise in time t and channel i, $s(t)$ is the true source signal, and $c_i \in [0, 1]$ are time invariant attenuation factors. In a special case that $M = 1$ and $c_1 = 1$, the multichannel model(1) becomes a single channel model. Taking the short-time Fourier transform (STFT) of (1), the model can be expressed in the frequency domain as

$$Y_i(w_k) = c_i S(w_k) + N_i(w_k) \tag{2}$$

where w_k represents the frequency in frequency bin k for each microphone i and $Y_i(w_k), S(w_k)$ and $N_i(w_k)$ are the STFT of the noisy speech signal, clean speech signal, and noise signal, respectively. The above equation can be expressed in spectral amplitude and spectral phase as

$$Y_i(l,k)e^{j\theta_Y^{(i)}(l,k)} = c_i S(l,k)e^{j\theta_S(l,k)} + N_i(l,k) \tag{3}$$

where $\{Y_i(l,k), S(l,k)\}$ denote the magnitudes and $\left\{\theta_Y^{(i)}(l,k), \theta_S(l,k)\right\}$ denote the phases, respectively at frame l, frequency bin k, and channel i of the noisy speech and clean speech. To simplify the notation, (3) is rewritten without the explicit dependencies as

$$Y_i e^{j\theta_Y^{(i)}} = c_i S e^{j\theta_S} + N_i$$

The goal is to determine the best estimate of the spectral amplitude S and spectral phase θ_S, based on known Y_i and $\theta_Y^{(i)}$.

Ephraim and Malah (1984) proposed one single channel speech enhancement using a minimum mean-square error short-time spectral amplitude (MMSE-STSA) estimator, given by

$$\begin{aligned}\widehat{S}_{STSA} &= E\{S|Y_1(w)\} = \int_0^{\infty}\int_0^{2\pi} Sp(S, \theta_S|Y_1(w))d\theta_S\, dS \\ &= \Gamma(1.5)\left(\frac{\sigma_S^2}{1+\xi_1}\right)^{0.5} e^{\left(-\frac{u_1}{2}\right)}\left[(1+u_1)I_0\left(\frac{u_1}{2}\right) + u_1 I_1\left(\frac{u_1}{2}\right)\right]\end{aligned} \tag{4}$$

where

$$\sigma_S^2 = E\{S^2\}, \quad u_1 = \frac{\xi_1\gamma_1}{1+\xi_1} \tag{5}$$

and ξ_1 and γ_1 denote the a priori and a posteriori SNRs, respectively defined as

$$\xi_1 = \frac{\sigma_S^2}{\sigma_{N_1}^2}, \quad \gamma_1 = \frac{Y_1^2}{\sigma_{N_1}^2} \quad \left(\sigma_{N_1}^2 = E\{N_1^2\}\right). \tag{6}$$

Also, Ephraim and Malah (1985) proposed another single channel speech enhancement using a minimum mean-square error log-spectral amplitude (MMSE-LSA) estimator, given by

$$\begin{aligned}\widehat{S}_{LSA} &= exp\{E[lnS|Y_1(w)]\} \\ &= \frac{\xi_1 Y_1}{1+\xi_1} exp\left\{\frac{1}{2}\int_{u_1}^{+\infty}\frac{e^{-t}}{t}dt\right\}\end{aligned} \tag{7}$$