Neurocomputing 000 (2017) 1–11

Contents lists available at ScienceDirect

Neurocomputing

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

A complex-valued multichannel speech enhancement learning algorithm for optimal tradeoff between noise reduction and speech distortion☆

Jingxian Tu[a], Youshen Xia[b,*], Songchuan Zhang[c]

[a] Laboratory of Complex System Simulation & Intelligent Computing, School of Information and Electronic Engineering, Wuzhou University, Wuzhou, China
[b] Department of Software Engineering, College of Mathematics and Computer Science, Fuzhou University, Fuzhou, China
[c] Department of Mathematics, Minjiang University, China

ARTICLE INFO

Article history:
Received 14 October 2016
Revised 25 March 2017
Accepted 6 June 2017
Available online xxx

Communicated by R. Capobianco Guido

Keywords:
Multichannel speech enhancement
Noise reduction
Signal distortion
Complex-valued learning algorithm
convergence analysis

ABSTRACT

To minimize speech distortion and residual noise, an optimal tradeoff between noise reduction and speech distortion needs to be considered. An optimal tradeoff method for single channel speech enhancement was presented by solving a real-valued constrained optimization model in a recent literature. This paper proposes a new optimal tradeoff method for multichannel speech enhancement by solving a complex-valued optimization problem subject to a residual noise constraint with the masking threshold of the clean speech. An effective complex-valued multichannel learning algorithm is developed and its convergence analysis is established completely in a complex domain. Experiment results confirm that the proposed multichannel speech enhancement algorithm outperforms several conventional algorithms in terms of both objective measures and subjective measures.

1. Introduction

Speech signals are usually corrupted by background noise in practice. The objective of speech enhancement is to restore the original signal from noisy observations. Over the past several decades, many of speech enhancement algorithms were presented. According to the availability of microphone information, speech enhancement algorithms can be classified as single channel based algorithms and multichannel based speech enhancement algorithms. The single channel speech enhancement algorithms were developed over several decades [1–8]. As the multi-microphone system was introduced, the multichannel speech enhancement algorithms were proposed in a recent decade. In contrast, since the multichannel microphone can utilize much more information to improve the performance of speech enhancement, the multichannel speech enhancement algorithms can exhibit better performance than the single channel speech enhancement algorithms by using received spatial information of signals. This paper focuses on the multichannel speech enhancement algorithms.

The multichannel speech enhancement algorithms mainly include the classic beamforming technique [9], the multichannel Wiener filtering technique [10–12], the linearly constrained minimum variance technique [13], the multichannel subspace technique [10,14–16], the statistical approach-based enhancement technique [17–19], and the robust and optimization-based technique [20]. Most of them have focused on improving the signal-to-noise ratio (SNR) without considering a tradeoff relationship between speech distortion and residual noise. Since noise reduction may result in speech distortion, a balanced tradeoff between noise reduction and speech distortion must be considered such that noise is maximally reduced while minimizing speech distortion.

This paper proposes a complex-valued multichannel speech enhancement algorithm for an optimal tradeoff between noise reduction and speech distortion. One contribution of this paper is that an optimal balanced tradeoff between noise reduction and speech distortion is achieved by solving a complex-valued quadratic optimization problem subject to a residual noise constraint with the masking threshold of the speech. Another contribution of this paper is that an effective complex-valued algorithm is proposed to

☆ This work is supported by the National Natural Science Foundation of China under Grant No. 61179037, and in part by the Natural Science Foundation of Fujian Province of China under Grant No.2017J01769, and in part by Scientific Research Fund of Education Department of Guangxi Zhuang Autonomous Region under Grant No. 2013YB223 and the Construction Fund of Master's Degree Grant Unit under Gui Degree [2013] No. 4.

* Corresponding author.
E-mail addresses: 568205127@qq.com (J. Tu), ysxia@fzu.edu.cn (Y. Xia), zsc_1977@126.com (S. Zhang).

http://dx.doi.org/10.1016/j.neucom.2017.06.018
0925-2312/

solve the proposed optimization problem. Moreover, the convergence of the proposed complex-valued algorithm is obtained in a complex domain. Experiment results show that the proposed multichannel speech enhancement algorithm outperforms several conventional algorithms in terms of both objective measures and subjective measures.

The remainder of this paper is organized as follows. In Section 2, we introduce related work. In Section 3, we introduce a new speech optimization method for multichannel speech enhancement and propose a complex-valued optimization algorithm for solving the associated optimization problem. In Section 4, we give the convergence analysis of the proposed optimization algorithm. In Section 5, we describe a complex-valued multichannel speech enhancement algorithm. In Section 4, we produce experimental results. In Section 5.2, we give the conclusion of this paper.

2. Speech signal model and related work

We are concerned with the speech signal model where a microphone array with N sensors captures a convolved source signal under some noise environment. The received signals are expressed as

$$\begin{aligned} y_n(t) &= g_n(t) * s(t) + v_n(t) \\ &= x_n(t) + v_n(t), \quad n = 1, 2, \ldots, N \end{aligned} \tag{1}$$

where $*$ denotes linear convolution, $s(t)$ is the unknown source signal, $g_n(t)$ represents the acoustic channel impulse response from the source to the nth microphone, $x_n(t) = g_n(t) * s(t)$ is the speech, $v_n(t)$ is the background noise and $y_n(t)$ is the received output signal at the nth microphone. We assume that the background noise and source signal $s(t)$ are mutually uncorrelated [1]. The problem under consideration focus on noise reduction. Without loss of generality, we consider the desired signal as the speech signal at the 1-st microphone. Our goal is to estimate $x_1(t)$ from received signals $\{y_n(t)\}_{n=1}^N$.

We first rewrite Eq. (1) in the short-time Fourier transform (STFT) domain at time frame l and discrete frequency k as

$$\begin{aligned} Y_n(k,l) &= G_n(k,l)S(k,l) + V_n(k,l) \\ &= X_n(k,l) + V_n(k,l), \quad n = 1, 2, \ldots, N \end{aligned} \tag{2}$$

where $Y_n(k, l)$, $G_n(k, l)$, $S(k, l)$, and $V_n(k, l)$ are the short-time Fourier transforms (STFTs) of $y_n(t)$, $g_n(t)$, $s(t)$, and $v_n(t)$, respectively. Furthermore, for discussing convenience, we omit k and l in all subsequent equations. Let $\mathbf{Y} = [Y_1, \ldots, Y_N]^T$, $\mathbf{V} = [V_1, \ldots, V_N]^T$, $\mathbf{X} = [X_1, \ldots, X_N]^T$, and $\mathbf{G} = [G_1, \ldots, G_N]^T$ where the superscript ^{T} denotes the transpose operator. Then Eq. (2) is expressed as in a N-dimensional vector form:

$$\mathbf{Y} = \mathbf{X} + \mathbf{V} = \mathbf{G}S + \mathbf{V} = \bar{\mathbf{G}}X_1 + \mathbf{V}. \tag{3}$$

where $\bar{\mathbf{G}} \triangleq \frac{\mathbf{G}}{G_1}$ is the frequency impulse response ratio vector. In the STFT domain, the conventional multichannel noise reduction is performed by applying a complex weight to the output of each sensor at the frequency and time frame.

Let $\hat{X_1} = \mathbf{W}^H\mathbf{Y}$ be the estimation of X_1 where $\mathbf{W} \in \mathbb{C}^{N\times 1}$ is a complex linear filter and ^{H} denotes the conjugate transpose. In general, the linear filter is designed to minimize the following mean squared error (MSE) of the residual signal

$$MSE = E[r(\mathbf{W})^H r(\mathbf{W})]$$

where E is the expectation operator and the residual signal is defined by

$$\begin{aligned} r(\mathbf{W}) &\triangleq \hat{X_1} - X_1 \\ &= (\mathbf{W} - \mathbf{e}_1)^H\mathbf{X} + \mathbf{W}^H\mathbf{V} \\ &= r_{\mathbf{X}} + r_{\mathbf{V}} \end{aligned} \tag{4}$$

where $\mathbf{e}_1 = [1, 0, \ldots, 0]^T$, $r_{\mathbf{X}} \triangleq (\mathbf{W} - \mathbf{e}_1)^H\mathbf{X}$ represents the signal distortion, and $r_{\mathbf{V}} \triangleq \mathbf{W}^H\mathbf{V}$ represents the residual noise. Since the clean speech signals x_n and noisy signals v_n are mutually independent, the MSE of the residual signal can be expressed as

$$\begin{aligned} &E\{r(\mathbf{W})^H r(\mathbf{W})\} \\ &= (\mathbf{W} - \mathbf{e}_1)^H E\{\mathbf{X}\mathbf{X}^H\}(\mathbf{W} - \mathbf{e}_1) \\ &\quad + \mathbf{W}^H E\{\mathbf{V}\mathbf{V}^H\}\mathbf{W} \\ &= (\mathbf{W} - \mathbf{e}_1)^H\mathbf{\Phi}_{\mathbf{X}}(\mathbf{W} - \mathbf{e}_1) + \mathbf{W}^H\mathbf{\Phi}_{\mathbf{V}}\mathbf{W} \end{aligned} \tag{5}$$

where $\mathbf{\Phi}_{\mathbf{X}} \triangleq E\{\mathbf{X}\mathbf{X}^H\}$ is the covariance matrix of speech, $\mathbf{\Phi}_{\mathbf{V}} \triangleq E\{\mathbf{V}\mathbf{V}^H\}$ is the covariance matrix of noise, and $(\mathbf{W} - \mathbf{e}_1)^H\mathbf{\Phi}_{\mathbf{X}}(\mathbf{W} - \mathbf{e}_1)$ and $\mathbf{W}^H\mathbf{\Phi}_{\mathbf{V}}\mathbf{W}$ denote the energy of the signal distortion and the energy of the residual noise, respectively.

The conventional multichannel Wiener filter (MWF) approach minimizes the MSE of the residual signal, which leads to the following optimal filter:

$$\mathbf{W}_{MWF} = (\mathbf{\Phi}_{\mathbf{X}} + \mathbf{\Phi}_{\mathbf{V}})^{-1}\mathbf{\Phi}_{\mathbf{X}}\mathbf{e}_1. \tag{6}$$

MWF is the optimal filter in the minimization of MSE. The energy of signal distortion is usually greater than 0 in MWF, this leads to speech distortion. Therefore, MWF does not consider how to obtain the optimal tradeoff between noise reduction and speech distortion. To overcome this defect, several optimization-based tradeoff schemes for minimizing speech distortion and residual noise were presented. By using a balanced tradeoff parameter between the noise reduction and speech distortion, speech distortion weighted MWF (SDW-MWF) approach was proposed by solving the following unconstrained optimization problem [11]:

$$\begin{aligned} &\min f_1(\mathbf{W}) \\ &\quad = (\mathbf{W} - \mathbf{e}_1)^H\mathbf{\Phi}_{\mathbf{X}}(\mathbf{W} - \mathbf{e}_1) + \mu\mathbf{W}^H\mathbf{\Phi}_{\mathbf{V}}\mathbf{W} \end{aligned} \tag{7}$$

where μ is the balanced trade-off parameter. The SDW-MWF solution is then given by

$$\mathbf{W}_{SDW-MWF} = (\mathbf{\Phi}_{\mathbf{X}} + \mu\mathbf{\Phi}_{\mathbf{V}})^{-1}\mathbf{\Phi}_{\mathbf{X}}\mathbf{e}_1. \tag{8}$$

The SDW-MWF is an improvement on the MWF since the balance between noise reduction and speech distortion can be controlled by tuning μ in the SDW-MWF. To avoid setting the balanced tradeoff parameter, a multichannel psychoacoustically motivated (MPM) algorithm was presented in [20]. The MPM speech enhancement approach solves the following complex-valued constrained optimization problem:

$$\begin{aligned} &\min f_2(\mathbf{W}) \\ &\quad = \rho_{X_1} \parallel \bar{\mathbf{G}}^H\mathbf{W} - 1 \parallel^2 + \mathbf{W}^H\mathbf{\Phi}_{\mathbf{V}}\mathbf{W} \\ &s.t \quad \mathbf{W}^H\mathbf{\Phi}_{\mathbf{V}}\mathbf{W} = T(k,l) \end{aligned} \tag{9}$$

where $\rho_{X_1} = E\{X_1X_1^H\}$ is the power spectrum of X_1 and $T(k, l)$ is the masking threshold of the speech at time frame l and discrete frequency k. For discussing convenience, we also omit k and l for $T(k, l)$ in all subsequent equations. It is seen that the MPM approach leads to audible residual noise under certain conditions. In other words, it is not guaranteed that both the speech distortion and residual noise are minimized in the MPM approach.

3. A new optimal tradeoff method for multichannel speech enhancement

For the optimal tradeoff between the noise reduction and speech distortion of multichannel speech enhancement, we propose minimizing the energy cost function of signal distortion satisfying the energy of residual noise below the masking threshold

Please cite this article as: J. Tu et al., A complex-valued multichannel speech enhancement learning algorithm for optimal tradeoff between noise reduction and speech distortion, Neurocomputing (2017), http://dx.doi.org/10.1016/j.neucom.2017.06.018