Signal Processing 106 (2015) 266–281

Contents lists available at ScienceDirect

Signal Processing

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

Joint source separation and dereverberation using constrained spectral divergence optimization

Karan Nathwani *, Rajesh M. Hegde

Department of Electrical Engineering, Indian Institute of Technology, Kanpur, India

ARTICLE INFO

Article history:
Received 29 April 2014
Received in revised form
29 July 2014
Accepted 3 August 2014
Available online 12 August 2014

Keywords:
Group delay function
Source separation and dereverberation
Divergence minimization
Subband envelope

ABSTRACT

A novel method of joint source separation and dereverberation that minimizes the divergence between the observed and true spectral subband envelopes is discussed in this paper. This divergence minimization is carried out within the non-negative matrix factorization (NMF) framework by imposing certain non-negative constraints on the subband envelopes. Additionally, the joint source separation and dereverberation framework described herein utilizes the spectral subband envelope obtained from group delay spectral magnitude (GDSM). In order to obtain the spectral subband envelope from the GDSM, the equivalence of the magnitude and the group delay spectrum via the weighted cepstrum is used. Since the subband envelope of the group delay spectral magnitude is robust and has a high spectral resolution, less error is noted in the NMF decomposition. Late reverberation components present in the separated signals are then removed using a modified spectral subtraction technique. The quality of separated and dereverberated speech signal is evaluated using several objective and subjective criteria. Experiments on distant speech recognition are then conducted at various direct-to-reverberant ratios (DRR) on the GRID corpus. Experimental results indicate significant improvements over existing methods in the literature.

1. Introduction

The objective of any source separation method is to recover the original signals from a composite signal. The problem of separation becomes more difficult when signals are mixed under a reverberant environment. Reverberation occurs when the distance between the speaker and the microphone is large enough to create multiple paths for the speech signal to arrive at the microphone. The reverberation results in degradation in intelligibility of the speech signal and the speech recognition performance.

Several algorithms have been developed for single channel speech dereverberation. The temporal averaging method is proposed in Xizhong and Guang [55] to estimate the room acoustic impulse response (AIR). This is done by complex cepstrum which utilizes an adaptive segmentation technique. The inverse filter solution is then obtained after pre-estimation of RIR. In Furuya et al. [18], the blind estimation of inverse filters required to obtain the dereverberated signal is explained. The inverse filters in Furuya et al. [18] are estimated by computing the correlation matrix between input signals, instead of room impulse response. A two stage algorithm for a single microphone has been proposed in Wu and Wang [54], where an inverse filter is estimated to reduce the coloration effects during the first stage. The spectral subtraction is then applied as a post-processing step to minimize the influence of long-term reverberation. In Tomar [53], the maximum kurtosis of the speech residual is proposed for blind dereverberation of speech signal. A non-negative matrix factorization

* Corresponding author. Tel.: +91 8765155012.
E-mail address: nathwani@iitk.ac.in (K. Nathwani).

http://dx.doi.org/10.1016/j.sigpro.2014.08.009
0165-1684/

(NMF) method which utilizes gammatone subband magnitude domain dereverberation is proposed in Kumar et al. [31]. In Kumar et al. [31], the Fourier transform spectral magnitude is used in an NMF framework for automatic speech recognition (ASR) applications. In Bees et al. [5], the dereverberation is carried out by using cepstrum to determine the acoustic impulse response and then used for inverse filtering to obtain the estimate of clean speech. The truncation error present in Bees et al. [5] is removed in Xizhong and Guang [55], but still inverse filtering is required. Authors in Kameoka et al. [29] present a blind dereverberation method designed to recover the subband envelope of an original speech signal from its reverberant version. The problem is formulated as a blind deconvolution problem with non-negative constraints, regularized by the sparse nature of speech spectrograms. In Nakatani et al. [39], a harmonicity-based dereverberation method is discussed to reduce the amount of reverberation in the signal picked up by a single microphone. A variant of spectral subtraction described in Kinoshita et al. [30] utilizes multi-step forward linear prediction for speech dereverberation. It precisely estimates and suppresses the late reverberations, which result in enhancing the ASR performance. All these methods deal withspeech dereverberation problem in a single source environment.

A considerable work has also been done to address the source separation problem under an anechoic environment. In an instantaneous frequency method [23], the objective is to extract target component of speech mixed with interfering speech and to improve the recognition accuracy that is obtained using the recovered speech signal. The instantaneous frequency is used to reveal the underlying harmonic structures of a complex auditory scene. In latent variable decomposition [45], each magnitude spectral vector of speech signal is represented as the outcome of a discrete random process. The latent dirichlet decomposition method [44] is a generalization of latent variable decomposition that models distribution process as a mixture of multinomial distributions. In this model, the mixture weights of the component multinomial vary from one analysis window to the other. Non-negative matrix factorization [48,33,49] is also an effective method in the context of mixed speaker separation by decomposing the STFT magnitude matrix [21]. A convolutive version of NMF is described in Smaragdis [52] that utilizes temporal variations into the account for source separation. The single channel separation of speech and music is discussed in Litvin et al. [34] by utilizing discrete energy separation algorithm (DESA). Apart from the single channel, multichannel underdetermined blind source separation in an anechoic environment is discussed in Bofill and Zibulevsky [8] and Niknazar et al. [41]. In Bertrand and Moonen [7], a non-negative BSS in a noise free environment using multiplicative updates and subspace projection is presented.

In general, the problem of source separation and dereverberation is looked at separately and solutions have been proposed for each of them individually as can be noted from the aforementioned discussion. However, the efforts have also been made in addressing the joint source separation and dereverberation problem. The joint optimization method for blind source separation (BSS) and dereverberation for multi-channel is discussed in Yoshioka et al. [60] by optimizing the parameters for the prediction matrices and for the separation matrices. A BSS framework in a noisy and reverberant environment based on a matrix formulation is proposed in Aichner et al. [1]. The method in Aichner et al. [1] allows simultaneous exploitation of nonwhiteness and nonstationarity of the source signals using second-order statistics. In Xu et al. [56], the joint block Toeplitzation and block-inner diagonalization (JBTBID) of a set of correlation matrices of the observed vector sequence is obtained for convolutive BSS. In Yoshioka et al. [59], the conditional separation and dereverberation method (CSD) for simultaneously achieving blind source separation and dereverberation of sound mixtures is discussed. A tractable BSS framework is explained in Arberet et al. [4] for estimating and combining spectral source models from noisy source estimates. In Rotili et al. [47], a general broadband approach to BSS for convolutive mixtures based on second-order statistics is discussed. The optimum inverse filtering algorithm based on the Bezouts theorem is used in the dereverberation stage. This is computationally more efficient and allows the inversion of long impulse responses in real-time applications. An integrated method for joint multichannel blind dereverberation and separation of convolutive audio mixtures is discussed in Yoshioka et al. [58]. All the above methods follow the tandem approach to solve the separation and reverberation problem for multichannel scenario. Additionally, the above joint blind source separation and dereverberation methods require multichannel input. This assumption has been relaxed in this work by considering the single channel case.

The contributions of the paper are as follows. The paper proposes a new model for joint blind source separation and dereverberation for the single channel under a multisource environment. In this work, the different impulse response is considered for different location of the speakers. Additionally, the proposed method uses subband envelope of the mixed speaker signal computed from group delay spectral magnitude (GDSM) [57,38] within the NMF framework. Due to the high resolution property of group delay function [57,24,3,10,9], this method reduces the error in the decomposition of observed subband envelope (OSE) sequence of the mixed signal into its constituent convolutional components.

In this work, the spectral divergence between observed subband envelope and true subband envelope (TSE) is minimized within the NMF framework. The convolutional components satisfying the non-negative constraint are then updated in an iterative manner. Once the subband envelope updates are obtained for each speaker, the spectral magnitude is then obtained by computing square root operation on the corresponding subband envelopes. Due to a fixed number of iterations in an NMF processing, some amounts of late reverberation and residual noise are still present in the updates of separated spectral magnitude. Hence, the remaining late reverberation and noise components are removed by post-processing methods. The experiments on source separation and speech