Speech Communication 82 (2016) 38–52

Contents lists available at ScienceDirect

Speech Communication

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

Supervised single-channel speech enhancement using ratio mask with joint dictionary learning☆

Long Zhang[a,b], Guangzhao Bao[a,b], Jing Zhang[a,b], Zhongfu Ye[a,b,c,*]

[a] Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei 230027, China
[b] National Engineering Laboratory for Speech and Language Information Processing, University of Science and Technology of China, Hefei 230027, China
[c] State Key Laboratory of Mathematical Engineering and Advanced Computing, Wuxi 214125, China

ARTICLE INFO

Article history:
Available online 7 June 2016

Keywords:
Single-channel speech enhancement
Ratio mask
Joint dictionary learning
Joint sparse coding
Ideal binary mask
Soft mask

ABSTRACT

A novel structure which combines the advantages of ratio mask (RM) and joint dictionary learning (JDL) is proposed for single-channel speech enhancement in this paper. The novel speech enhancement structure makes full use of the training data and overcomes some shortcomings of generative dictionary learning (GDL) algorithm. RMs of speech and interferer are introduced to provide the discriminative information both in the training stage and enhancement stage of the novel structure. In the training stage, the signals and their corresponding ideal RMs (IRMs) are used to learn the signal and IRM dictionaries jointly by K-SVD algorithm. In the enhancement stage, the mixture signal and mixture RM are sparsely represented over the composite dictionaries composed of the learned signal and IRM dictionaries to formulate a joint sparse coding (JSC) problem. Then, the estimated RMs (ERMs) of speech and interferer in the mixture are calculated to develop two soft mask (SM) filters. The proposed SM filters incorporate ideal binary mask technique and Wiener-type filter to make full use of the discriminative information provided by the ERMs. They are used to both strengthen the speech and suppress the interferer in the mixture. The proposed algorithms have shown their abilities to improve both speech intelligibility and quality. Experimental evaluations verify the proposed algorithms obtain comparable performances to a deep neural network (DNN) based mask estimator with lower computation and perform better than other tested algorithms.

1. Introduction

In the real world, clean speech signals are often degraded by interferers. Enhancing the speech degraded by non-stationary real-world interferers is important for many signal processing applications, including hearing aids, mobile communications, and preprocessing for speech recognition (Loizou, 2007). The goal of single-channel speech enhancement is to reconstruct the underlying clean speech from a single-channel additive mixture signal composed of the clean speech and interferer components. Conventional speech enhancement algorithms can be divided into three categories (Loizou, 2007): spectral subtraction (SS) approaches (Boll, 1979; Kamath and Loizou, 2002; Lu and Loizou, 2008), statistical-model-based approaches (Lim and Oppenheim, 1978; Plapous et al., 2006; Ephraim and Malah, 1984; Ephraim and Malah, 1985; Ephraim, 1992; Cohen and Berdugo, 2001) and subspace approaches (Ephraim and Hany, 1995; Hu and Loizou, 2003; Sun et al., 2008). SS approaches estimate the clean speech spectra by subtracting an estimate of the interferer spectra from the observed mixture. However, this kind of methods assumes that the interferers are stationary and will introduce some artifacts referred to as musical noise. Statistical-model-based approaches include the Wiener algorithms, the famous minimum-mean square error (MMSE) short-time spectral amplitude estimator (MMSE-STSA) (Ephraim and Malah, 1984) and its variant MMSE log-spectral amplitude estimator (MMSE-LSA) (Ephraim and Malah, 1985), the optimally-modified log-spectral amplitude (OMLSA) (Cohen and Berdugo, 2001) estimator and others. The Wiener algorithms estimate the clean speech spectra by applying Wiener filter on either the complex domain or the magnitude domain. MMSE-STSA and its variants always estimate the clean speech's magnitude based on Gaussian distribution of the speech and noise spectra and update the variance of speech spectra using decision-directed (DD) estimator. Some other speech spectra distributions are also used in statistical models, such as Gamma and Laplace (Loizou, 2007; Ephraim, 1992). The optimal spectral gain function of OMLSA, which

☆ This work is supported by Open Project Program of the State Key Laboratory of Mathematical Engineering and Advanced Computing (No. 2015A15).

* Corresponding author.
E-mail addresses: lonzhang@mail.ustc.edu.cn (L. Zhang), gzbao@mail.ustc.edu.cn (G. Bao), zhj336@mail.ustc.edu.cn (J. Zhang), yezf@ustc.edu.cn (Z. Ye).

http://dx.doi.org/10.1016/j.specom.2016.06.001
0167-6393/

minimizes the mean-square error of the log-spectra, is obtained as a weighted geometric mean of the hypothetical gains associated with the speech presence uncertainty. The idea behind subspace approaches is to project the noisy signal onto two subspaces: the signal-plus-noise subspace and the noise subspace. The noise subspace only contain signals from the noise process, hence an estimate of the clean speech can be made by removing the components of the signal in the noise subspace and retaining only the components in the signal subspace. However, the assumption of orthogonality between the speech and interferer subspaces is too strict in some cases. The conventional approaches are applicable to the situation where the interferers are stationary. However, in practical applications, the interferers are often non-stationary and potentially similar to the clean speeches, which cause the conventional approaches incompetent.

Recently, some supervised speech enhancement/source separation approaches based on sparsity model have been proposed by making use of the fact that speech and interferer signals have approximately sparse coding (SC) in suitably chosen dictionaries, respectively. In these approaches, the underlying speech is reconstructed from the single-channel mixture signal by sparsely representing the mixture signal over a composite dictionary consisting of the concatenation of the speech and interferer dictionaries, which can dominantly represent the speech and interferer signals, respectively. Supervised nonnegative matrix factorization (NMF) based speech enhancement algorithms (Grais and Erdogan, 2013; Mohammadiha et al., 2013; Simsekli et al., 2014) learn the speech and the interferer dictionaries using different objective functions and constraints including nonnegative representation and sparseness (Lee and Seung, 2001; Le Roux et al., 2015), and sparsely represent the mixture signals on the learned composite dictionary to extract the speech and interferer components and then estimate the speech by a Wiener-type filter (Grais and Erdogan, 2013; Mohammadiha et al., 2013) in the end. The idea of supervised probabilistic latent component analysis (PLCA) based enhancement algorithm (Smaragdis et al., 2007; Smaragdis and Raj, 2007) is similar to the supervised NMF (SNMF) based algorithms, however, it explores the PLCA models (Smaragdis and Raj, 2007) to learn the speech and interferer dictionaries in time-frequency (TF) domain. Generative dictionary learning (GDL) algorithm (Sigg et al., 2010; Sigg et al., 2012) learns the speech and interferer dictionaries based on approximate K-SVD dictionary update algorithm (Rubinstein et al., 2008) in the TF domain and sparsely represents the mixture signal over the composite dictionary using batch least angle regression with coherence criterion (LARC) algorithm (Sigg et al., 2012; Efron et al., 2004). These sparsity model based speech enhancement algorithms are shown to reduce non-stationary interferers effectively and perform well on speech enhancement and source separation.

Deep learning based methods have recently started to attract much more attention in the speech enhancement research community by modeling the nonlinear mapping relationship between the mixture and enhanced speech. Prior works on deep learning based speech enhancement can be categorized into two categories depending on the interaction between input mixture and output targets, mainly including deep neural networks (DNN) based mask algorithms (Williamson et al., 2014; Wang et al., 2014) and DNN based on regression algorithms (Xu et al., 2015). The output targets of DNN based mask estimators are always different masks such as ideal binary mask (IBM), while the output targets of DNN based on regression estimators are different speech features such as speech TF magnitude. They both have shown great abilities to enhance the speech signal from the mixture.

Estimating speech magnitude by directly multiplying the learned speech dictionary and corresponding SC, GDL sometimes causes large errors compared to the clean speech magnitude because of producing quite a few outliers and negative magnitudes. Here, "outlier" means the estimate speech magnitude larger than the mixture's. In order to overcome these shortcomings and make full use of information provided by training data sets, a novel structure based on a new continuous valued mask named ratio mask (RM) is developed. Similar to the definition in Williamson et al. (2014), RMs for speech/interferer are introduced to be the ratios between the speech/interferer magnitude and their mixture's in the TF domain. Therefore, RMs have the discriminative information to extract the speech/interferer magnitude from the mixture's. Here, "discriminative" means that RMs have abilities to indicate the magnitude ratios and extract the speech/interferer magnitude from the mixture. RMs have been shown to be effective on improving the speech intelligibility and quality (Srinivasan et al., 2006; Williamson et al., 2014; Wang et al., 2014). Noted that the clean data sets are existed in the training stage to calculate the ideal RMs (IRMs) for speech and interferer, while the estimated RMs (ERMs) of speech and interferer in the mixture are need to calculated in the enhancement stage.

Considering the discriminative information provided by the RMs, the proposed novel structure combines joint dictionary learning (JDL) (Zhang and Li, 2010; Jiang et al., 2011) and joint SC (JSC) to recover the underlying speech from the mixture for the first time. The JDL means using the clean signals including speech and interferer data and their corresponding IRMs to jointly learn the signal and IRM dictionaries in the training stage, which benefit from both information provided by the signals and IRMs. The JSC means sparsely representing the mixture signal and mixture RM over the learned composite signal and IRM dictionaries to estimate the SCs in the enhancement stage and will benefit for calculating the ERMs of speech and interferer from the mixture. Based on the ERMs, two soft mask (SM) filters which combine ideal binary mask (IBM) technique (Srinivasan et al., 2006; Li and Loizou, 2008) and Wiener-type filter (Grais and Erdogan, 2013; Mohammadiha et al., 2013), are proposed to recover the speech from the mixture in this paper. The proposed SM filters, which take the advantages of both IBM and Wiener-type filter, strengthen the speech part while suppress the interferer part and relieve the weaknesses of GDL to reduce the outliers and negative magnitudes in the TF domain effectively. Two proposed algorithms based on the SM filters are proven to promote both the speech intelligibility and quality efficiently. The novel structure is named RMJDL because of the crucial role played by the RM and JDL.

The major contribution of this paper can be summarized as followed three points. 1) RMs of the speech and interferer are introduced to provide discriminative information for the signals, then a JDL framework is formed to learn the signal and IRM dictionaries jointly for the first time; 2) A JSC framework is proposed for representing the mixture signals and mixture RM over the learned composite signal and IRM dictionaries to improve the calculated ERMs from the mixture; 3) Two SM filters combining the IBM and Wiener-type filter are elaborately developed to estimate the speech magnitude from the mixture, which make full use of the discriminative information provided by the ERMs.

The rest of this paper is organized as follows. Section 2 briefly describes sparsity model based speech enhancement and gives an overview of the overall RMJDL structure and proposed algorithms. The signal model and details of proposed algorithms will be presented in Section 3. Then, experiment evaluations and discussions are presented in Section 4. Finally, conclusions and systematic evaluations are drawn in Section 5.

Notation: Upper (lower) bold face letters are used to denote matrices (column vectors). $(\cdot)^T$, $|\cdot|$, $\|\cdot\|_0$, and $\|\cdot\|_F$ denote the transpose, absolute value, l_0 norm and Frobenius norm, respectively. **A**./**B** means element-wise division.