Applied Acoustics 132 (2018) 1–7

Contents lists available at ScienceDirect

Applied Acoustics

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

Supervised monaural speech enhancement using two-level complementary joint sparse representations

Jiafei Fu, Long Zhang, Zhongfu Ye*

Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei, Anhui 230027, China
National Engineering Laboratory for Speech and Language Information Processing, Hefei, Anhui 230027, China

ARTICLE INFO

Keywords:
Joint sparse representation
Joint dictionary learning
Residual weighting
Monaural speech enhancement

ABSTRACT

Recently, a new complementary joint sparse representations (CJSR) method was proposed for monaural speech enhancement, which utilizes the relationships among speech, noise and mixture. One of the joint sparse representations (JSRs) uses the mapping relationship between mixture and speech while the other uses the mapping relationship between mixture and noise. However, since they only use the joint information and overcomplete dictionaries, there may be some confusion components between the estimated speech and noise. In this paper, a novel model with fusion process is proposed, which further using an additional composite dictionary composed of clean speech dictionary and noise dictionary as prior knowledge. When the estimated speech and noise are obtained by JSRs, they are further sparsely represented on the composite dictionary, respectively. Because that the composite dictionary has the discriminative property, the source confusion problem of previous methods is coped well. In order to take advantage of the complementary knowledge of the two estimated signals, we propose a weighting parameter based on the residuals of sparse representation on composite dictionary. If one of the residuals is smaller, the corresponding weighting is greater, and this weighting parameter varies with noise type and speaker flexibly and effectively. Experimental results show that the proposed algorithm has superior performance compared with other tested approaches.

1. Introduction

Enhancing speech degraded by non-stationary real-world interference has been a topic of research in the last few decades, not only because of its difficulty, but also for various applications, including hearing aids, automatic speech recognition, mobile communications, etc. [1]. Conventional single-channel speech enhancement approaches can be categorized into three branches: spectral subtraction (SS) approaches [2–4], statistical-model-based approaches [5–8] and subspace approaches [9–12], the performances of which are mostly dependent on the estimated noise in the absence of speech activity, so their performance for non-stationary noise may not be satisfactory.

Recently, some sparse-model-based speech enhancement approaches have been proposed by more and more researchers. Non-negative matrix factorization (NMF) was first proposed in [13], with clear practical meaning and fast convergence, it has obtained good performance in facial recognition and image denoising. Considering that speech signals' magnitude spectrograms in time-frequency (TF) are non-negative values, supervised NMF [14–16] based speech enhancement approaches have been proposed. These approaches decompose the training data into dictionaries and coefficients for speech and noise, respectively. Given the mixture data, the underlying speech and noise can be represented respectively on the composite dictionary and then the speech is estimated by a Wiener-type filter [16] in the end. The generative dictionary learning (GDL) approach [17,18] learns the overcomplete dictionaries for the speech and noise, then concatenates these two dictionaries for speech enhancement. All these sparse-model-based approaches are shown to be effective in reducing non-stationary noises and have good performance on speech enhancement.

Nevertheless, these approaches have a common drawback, they learn the prior knowledge of speech and noise respectively, but don't utilize potential relative knowledge between original signals. In order to solve this problem, [19] utilizes the mixture and speech simultaneously for JSR in the training stage. Then in the enhancement stage, the noisy speech is sparsely coded over the mixture dictionary to obtain representation coefficients. Whereafter, the speech dictionary and the corresponding representation coefficient are combined to estimate the clean speech. By using the JSR, this method can deal with the nonlinear mapping from the mixture to the speech in the feature domain. Based on this approach [20], proposed an approach using CJSR, and this

* Corresponding author at: Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei, Anhui 230027, China and National Engineering Laboratory for Speech and Language Information Processing, Hefei, Anhui 230027, China.
E-mail addresses: fujiafei@mail.ustc.edu.cn (J. Fu), lonzhang@mail.ustc.edu.cn (L. Zhang), yezf@ustc.edu.cn (Z. Ye).

https://doi.org/10.1016/j.apacoust.2017.11.005
Received 13 April 2017; Received in revised form 31 October 2017; Accepted 7 November 2017

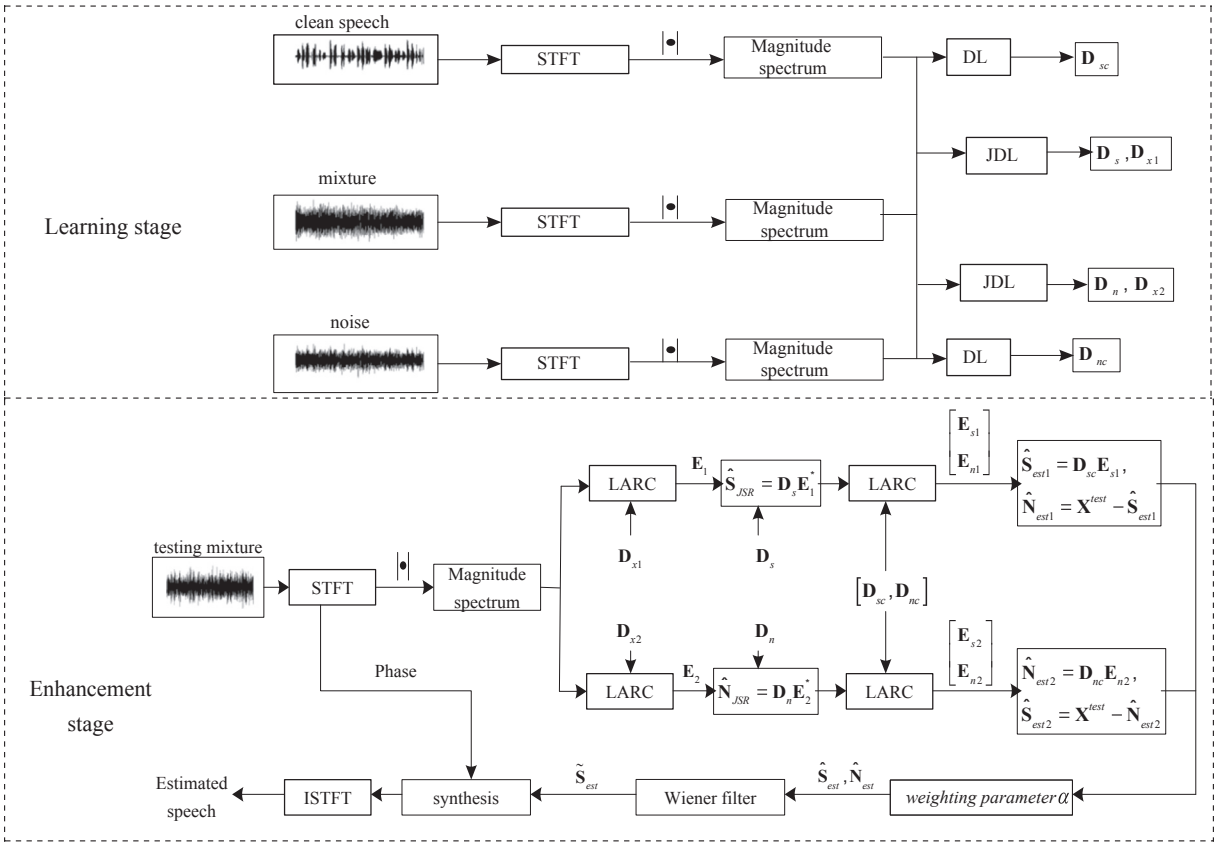

Fig. 1. The overall structure of the proposed algorithm, consisting of learning stage and enhancement stage.

approach uses the linear mapping in the TF domain. Moreover, not only the mapping from mixture to speech is considered, the mapping from mixture to noise is also used. Then the Gini index [21,22] is introduced to combine the complementation of these two JSRs. The existing algorithms assume that clean speech and noise have their unique components, which makes that speech and noise can be explained by the corresponding sub-dictionaries. But the speech and noise have similar components when they are sparsely represented on mixture dictionary, which we denote source confusion [18]. Because the above method only used the joint information among speech, noise and mixture, and the learned dictionaries are overcomplete, the estimated speech and noise still have some confusion components.

The main contribution of this paper is to propose a two-level model scheme with fusion process for single-channel speech enhancement, which using an additional composite dictionary to further reduce the confusion components and weighting parameter to balance the complementation more robustly. The proposed two-level processing model based sparse representation scheme is shown in Fig. 1. The additive signal model is taken into account after the second-level sparse representation to obtain the underlying speech or noise. From Fig. 1, one can see that we first make full use of clean speech and noise database to learn joint dictionaries and composite dictionary respectively. Then the joint dictionaries with the testing mixture signal are used to preliminarily estimate a less distorted speech signal and noise signal in the STFT magnitude domain. After that, the obtained preliminary with the composite dictionary are used to build the second-level processing model which can further reduce the confusion components in estimated speech and noise. In addition, similar to previous work [20], in order to make use of the two-way complementation, weighting parameter based on the residuals of sparse representation on composite dictionary is utilized to drive the optimal estimator in our proposed algorithms. Based on the scheme, a Wiener-type filter is finally designed to enhance the speech. This paper bases on the previous work of our laboratory [20] and enriches the basic idea about the CJSR processing by extending the ideas and doing more experiments. In this paper, an easier and more detailed understanding of the proposed two-level model is given and the composite dictionary is introduced to enhance the speech, which try to take advantage of more information of training data. Experimental results show that the proposed speech enhancement method based on two-level CJSR can achieve almost better performances than some state-of-the-art methods under different situations.

The review of CJSR-based speech enhancement is presented in Section 2. Section 3 details the proposed two-level algorithm. Simulation results are given in Section 4. Finally, conclusions are summarized in Section 5.

2. Review of CJSR-based speech enhancement

In this section, we review the CJSR-based speech enhancement approach in [20].

Straightforwardly, the signal can be modeled as a linear additive mixture of clean speech and noise

$$x(t) = s(t) + n(t), \quad 1 \leqslant t \leqslant T, \tag{1}$$

where $x(t)$ is the time-domain mixture signal at sample t, and $s(t)$ and $n(t)$ are the time-domain speech and noise signals.

After applying a short time Fourier transform (STFT) on both sides of (1) and only using the magnitude [23,24], it can be expressed in the TF domain as

$$|X(t,f)| \approx |S(t,f)| + |N(t,f)|, \tag{2}$$

where t and f are the time and frequency index, $|S(t,f)|$ and $|N(t,f)|$ are the STFT magnitudes of clean speech and noise in the mixed signal. The magnitude spectra can be written in a matrix form as

$$\mathbf{X} \approx \mathbf{S} + \mathbf{N}, \tag{3}$$