IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 8, 2018, accepted November 27, 2018, date of publication December 17, 2018,
date of current version January 11, 2019.

Digital Object Identifier 10.1109/ACCESS.2018.2886991

Speech Enhancement Based on Dictionary Learning and Low-Rank Matrix Decomposition

YUNYUN JI[1,2], WEI-PING ZHU[2], (Senior Member, IEEE), AND BENOIT CHAMPAGNE[3], (Senior Member, IEEE)

[1]School of Electronics and Information, Nantong University, Nantong 226019, China
[2]Department of Electrical and Computer Engineering, Concordia University, Montreal H3G 2W1, Canada
[3]Department of Electrical and Computer Engineering, McGill University, Montreal H3A 0E9, Canada

Corresponding author: Yunyun Ji (ellaji1985@gmail.com)

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada under Grant CRDPJ 515072-17, in part by the National Natural Science Foundation of China under Grant 61601248 and Grant 61871241, and in part by the University Natural Science Research Foundation of Jiangsu Province under Grant 16KJB510037.

ABSTRACT In this paper, a new speech enhancement method is proposed based on dictionary learning and low-rank matrix decomposition with the objective to improve speech quality and intelligibility. In both the learning stage and the enhancement stage of the proposed method, a new noise model is employed to capture the noise characteristics with a combination of a low-rank matrix and an overcomplete noise dictionary. In the learning stage, an unsupervised dictionary learning algorithm is first proposed based on this new noise model to train a noise dictionary with low mutual coherence to the clean speech. Then, a supervised low-rank matrix decomposition algorithm is proposed to extract the low-rank component of noise using the clean speech, noise, and noisy speech training data sets. The estimated low-rank matrix is clustered through the K-means method to generate a codebook which provides useful information for the enhancement stage. Finally, in the enhancement stage, by using the well-trained dictionaries and the reference codebook from the learning stage, a decomposition algorithm for the low-rank matrix and sparse component is developed to effectively estimate the clean speech from the noisy speech. The experimental results show that our proposed method achieves better enhancement performance than some state-of-the-art reference methods in terms of four objective performance evaluation measures especially in low signal-to-noise ratio adverse environments.

INDEX TERMS Clustering, noise model, speech enhancement, supervised low-rank matrix decomposition, unsupervised dictionary learning.

I. INTRODUCTION

With the pervasive use of wireless networks and smart home devices, speech enhancement has become an indispensable part of voice communication systems and human-machine interfaces, as a means to improve both quality and intelligibility of speech signals contaminated by background noises [1]. In the last decade, sparse-model based speech enhancement methods have been proposed to achieve effective denoising, especially under non-stationary noisy environment, by using dictionary learning approaches [2]–[4] or the low-rank matrix decomposition method [5]. These methods can be divided into two categories: the first group is based on the sparsity of both speech and noise while the second group is based on the sparsity of speech and the low-rank property of noise.

The work in the first category was initiated in [6] where a generative dictionary learning (GDL) method was proposed for estimating the speech magnitude spectra. In the learning stage, the GDL method employs the K-singular value decomposition (K-SVD) algorithm [7] to carry out an unsupervised learning of redundant dictionaries for both speech and noise in the time-frequency domain. Then, in the enhancement stage, the noisy speech magnitude spectra are sparsely represented with a composite dictionary, i.e. a concatenation of the learned speech and noise dictionaries, through a sparse coding algorithm [8]. The speech magnitude spectra are recovered through the product of the speech dictionary and the estimated sparse coefficient vectors of speech. Finally, the time-domain speech signals are synthesized via inverse Fourier transform (IFT) of the estimated speech spectra.

Different from the GDL method, the complementary joint sparse representation (CJSR) method [9] utilizes the noisy speech magnitude spectra in the learning stage together with the clean speech magnitude spectra and the noise magnitude spectra, respectively, to form two different training datasets

to train two distinct mixture dictionaries. The trained mixture dictionaries offer latent mappings from the noisy speech to the clean speech and noise, respectively, which enable the sparse coding algorithm in the enhancement stage to improve the accuracy of both speech and noise estimations in the frequency domain.

Thus, the methods in the first category rely only on the intra-frame characteristic of the speech and noise (i.e. the sparsity) to learn redundant dictionaries and seek for a robust sparse coefficient representation in order to achieve the intended separation of speech and noise in adverse environments. However, the robustness of sparse representation is highly related to the intrinsic distinctiveness of the speech and noise spectrograms [6]. Consequently, the enhancement performance is subject to mutual coherence between the speech and the noise dictionary.

The methods in the second category decouple the mixture signal in the time-frequency domain into a sparse component for speech and a low-rank component for noise. The constrained low-rank and sparse matrix decomposition (CLSMD) method [10] can achieve effective separation of speech and noise by constraining the number of the non-zero entries in the speech magnitude spectral matrix and the rank of the noise magnitude spectral matrix. Different from the CLSMD method, the robust principal component analysis (RPCA) based speech enhancement method [11] employs a speech dictionary to sparsely represent the speech magnitude spectra and utilizes the nuclear norm to control the rank of the noise magnitude spectral matrix. The optimization problem developed in this method is then solved by the alternating direction method of multiplier (ADMM) algorithm [12]. Similar to the RPCA method in [11], the learnable sparse and low-rank decomposition (LSLD) algorithm [13] guarantees the low-rank property of the noise magnitude spectral matrix by minimizing its nuclear norm. Hence, the methods in this category exploit the low-rank model to describe non-stationary noise, focusing on the inter-frame coherence of the noise spectrogram.

In this paper, we propose a new method for monaural speech enhancement that is based on unsupervised dictionary learning and supervised low-rank matrix decomposition. Our method differs from the aforementioned approaches in four aspects. Firstly, we combine the sparse model and the low-rank model to develop a new mathematical model for background noise, which underlies both the learning stage and enhancement stage of the new method. Secondly, based on this new noise model, we propose a new optimization technique in the learning stage to train a noise dictionary in an unsupervised fashion. Thirdly, we propose a supervised low-rank matrix decomposition algorithm in the learning stage to extract a low-rank noise component, which is used to design a representative noise codebook for the enhancement stage. Finally, in the enhancement stage, the well-trained dictionaries and the codebook are utilized along with a new low-rank matrix and sparse component decomposition algorithm to decouple the noisy speech magnitude spectral matrix into two separate parts, namely, the sparse component and the low-rank component. The speech magnitude spectral matrix is then estimated from the sparse component. The joint use of the inter-frame coherence (i.e. the low-rank property) and the intra-frame characteristic (i.e. the sparsity) of noisy speech magnitude spectra can account for the performance improvement of our proposed approach.

The rest of the paper is organized as follows: Section II provides a brief review of the GDL and the CLSMD which serve as basis for our proposed method. In Section III, we develop the new speech enhancement method in detail. Section IV presents experimental results of the proposed method with comparison to five reference methods, and for four objective performance measures. Finally, we conclude the paper in Section V. In addition, in this paper, the notation $\|\cdot\|_0$ denotes the ℓ_0 norm, counting the number of nonzero elements in a vector or a matrix. $\|\cdot\|_1$ denotes the ℓ_1 norm, referring to the sum of the absolute elements in a vector or a matrix. $\|\cdot\|_\infty$ denotes the maximum norm, referring to the maximum absolute element in a vector. And $\|\cdot\|_F$ denotes the Frobenius norm of a matrix [14].

II. RELATED WORK

In single-channel speech enhancement, the noisy speech signal $x(t)$ at time sample t can be expressed as

$$x(t) = s(t) + n(t) \tag{1}$$

where $s(t)$ and $n(t)$ denote, respectively, the clean speech and noise. In the context of sparse-model based speech enhancement, with the goal to recover $s(t)$ from $x(t)$, a suitable feature space should first be selected for the representation of the time-domain signals. An appropriate and common candidate to realize this representation is the short time Fourier transform (STFT) [8]. After being segmented into frames, the mixture signal $x(t)$ is transformed into the time-frequency domain by means of the STFT. Consequently, (1) can be converted as

$$X(f, m) = S(f, m) + N(f, m) \tag{2}$$

where $X(f, m)$, $S(f, m)$ and $N(f, m)$ represent the STFT spectra respectively of the noisy speech, clean speech and noise, $f \in \{1, 2, \cdots, M\}$ and $m \in \{1, 2, \cdots, N\}$ denote the frequency bin and frame index respectively. Regardless of the phases of complex STFT coefficients, existing sparse-model based speech enhancement approaches exploit the sparsity of the speech magnitude spectrum $|S(f, m)|$ to effectively separate it from the noisy speech magnitude spectrum $|X(f, m)|$.

The GDL method [6] provides a basic framework for dictionary learning based speech enhancement methods. The GDL method is divided into two stages: the learning stage and the enhancement stage. In the learning stage, the GDL method exploits the K-SVD algorithm to solve the following two optimization problems for training the speech dictionary and noise dictionary, respectively, i.e.,

$$\min_{\boldsymbol{D}_s, \boldsymbol{\Theta}_s} \left\| \boldsymbol{S} - \boldsymbol{D}_s \boldsymbol{\Theta}_s \right\|_F^2 \quad \text{s.t.} \ \left\| \boldsymbol{\theta}_{s,i} \right\|_0 \leq K, \ \forall i, \tag{3}$$

and

$$\min_{\boldsymbol{D}_n,\boldsymbol{\Theta}_n} \left\|\boldsymbol{N}-\boldsymbol{D}_n\boldsymbol{\Theta}_n\right\|_F^2 \quad \text{s.t. } \left\|\boldsymbol{\theta}_{n,i}\right\|_0 \le K,\ \forall i. \tag{4}$$

In this formulation: $\boldsymbol{S} \in \mathbb{R}^{M\times N_1}$ and $\boldsymbol{N} \in \mathbb{R}^{M\times N_2}$ represent respectively the speech and noise magnitude spectral matrices for training; $\boldsymbol{D}_s \in \mathbb{R}^{M\times P}$ and $\boldsymbol{D}_n \in \mathbb{R}^{M\times P}$ denote respectively the speech dictionary and the noise dictionary to be trained; $\boldsymbol{\Theta}_s \in \mathbb{R}^{P\times N_1}$ and $\boldsymbol{\Theta}_n \in \mathbb{R}^{P\times N_2}$ are the sparse coefficient matrices in the representation of the speech magnitude spectral matrix $\boldsymbol{S}$ and the noise magnitude spectral matrix $\boldsymbol{N}$ in terms of their dictionaries $\boldsymbol{D}_s$ and $\boldsymbol{D}_n$; and $\boldsymbol{\theta}_{s,i}$ and $\boldsymbol{\theta}_{n,i}$ represent the i^{th} column vector in $\boldsymbol{\Theta}_s$ and $\boldsymbol{\Theta}_n$.

In the enhancement stage, the noisy speech magnitude spectral matrix $\boldsymbol{X}$ can be sparsely represented by a composite dictionary $\boldsymbol{D}$ which is a concatenation of the speech dictionary $\boldsymbol{D}_s$ and the noise dictionary $\boldsymbol{D}_n$ from the learning stage, i.e.,

$$\boldsymbol{X} = \boldsymbol{D}\boldsymbol{\Theta} = \begin{bmatrix}\boldsymbol{D}_s & \boldsymbol{D}_n\end{bmatrix}\begin{bmatrix}\boldsymbol{\Theta}_s \\ \boldsymbol{\Theta}_n\end{bmatrix} \tag{5}$$

where $\boldsymbol{D} = \begin{bmatrix}\boldsymbol{D}_s\ \boldsymbol{D}_n\end{bmatrix}$ and $\boldsymbol{\Theta} = \begin{bmatrix}\boldsymbol{\Theta}_s \\ \boldsymbol{\Theta}_n\end{bmatrix}$. It is noted that $\boldsymbol{\Theta}_s$ and $\boldsymbol{\Theta}_n$ in (5) are the submatrices of the matrix $\boldsymbol{\Theta}$ to be estimated. For convenience and simplicity, we do not use any superscript or subscript to distinguish the sparse coefficient matrices estimated from the data during the training and enhancement stages. Here, $\boldsymbol{\Theta}$ is the sparse coefficient matrix of $\boldsymbol{X}$ with respect to the composite dictionary $\boldsymbol{D}$. It can be estimated by using the least angle regression with coherence criterion (LARC) algorithm [6] to solve the following optimization problem,

$$\widehat{\boldsymbol{\Theta}} = \arg\min_{\boldsymbol{\Theta}} \frac{1}{2}\left\|\boldsymbol{X}-\boldsymbol{D}\boldsymbol{\Theta}\right\|_F^2 + \lambda\left\|\boldsymbol{\Theta}\right\|_1. \tag{6}$$

Then, the speech magnitude spectral matrix can be estimated as

$$\widehat{\boldsymbol{S}} = \boldsymbol{D}_s\widehat{\boldsymbol{\Theta}}_s \tag{7}$$

where $\widehat{\boldsymbol{\Theta}}_s$ is the corresponding submatrix of $\widehat{\boldsymbol{\Theta}}$ as in (5). The speech spectra are then obtained using the estimated speech magnitude spectra in (7) and the phases of the noisy speech spectra. Finally, the time-domain speech signals are synthesized by applying the IFT to the estimated speech spectra.

The GDL method utilizes the unsupervised K-SVD algorithm to train both speech and noise dictionaries in the learning stage and employs the LARC algorithm to estimate the sparse coefficient matrix of speech in the enhancement stage. As such, it is able to achieve complete separation of speech and noise when the learned speech and noise dictionaries are coherent to their respective signal classes and incoherent to the other signal classes in the noisy speech. However, this condition cannot be satisfied in a real-world environment due to the non-stationarity of noise, which accounts for the primary drawback of the GDL, namely the so-called source confusion [6].

Different from the GDL, the CLSMD method uses the low-rank matrix decomposition to extract the noise from the noisy speech without using a learning scheme [10]. Based on the sparsity of the speech magnitude spectral matrix and the low-rank property of the noise magnitude spectral matrix, the CLSMD method considers the following two subproblems to alternately and iteratively estimate the speech and noise components from the noisy speech magnitude spectral matrix.

$$\boldsymbol{N}^t = \arg\min_{\text{rank}(\boldsymbol{N})\le r} \left\|\boldsymbol{X}-\boldsymbol{N}-\boldsymbol{S}^{t-1}\right\|_F^2. \tag{8}$$

$$\boldsymbol{S}^t = \arg\min_{\left\|\boldsymbol{S}\right\|_0\le K} \left\|\boldsymbol{X}-\boldsymbol{N}^t-\boldsymbol{S}\right\|_F^2. \tag{9}$$

Here, the parameter r is the upper bound on the rank of $\boldsymbol{N}$ while K is the sparsity level of $\boldsymbol{S}$. The optimization problems in (8) and (9) are solved respectively through the singular value hard thresholding algorithm and the hard thresholding algorithm in [10].

None of the existing sparse-model based speech enhancement methods has incorporated the low-rank property of the noise magnitude spectral matrix into the learning stage. In this work, we exploit the low-rank component of noise both in the learning stage and the enhancement stage, which, as shown in Section III, can largely improve the enhancement performance.

III. PROPOSED SPEECH ENHANCEMENT SYSTEM

In this section, we first describe the operation of the complete speech enhancement system based on our proposed method. We subsequently present the details of the proposed algorithms for the learning stage in Subsection III-B and the enhancement stage in Subsection III-C.

A. SYSTEM OVERVIEW

The block diagram of our proposed system, depicted in Fig. 1, is divided into two stages: the learning stage and the enhancement stage. In order to reduce source confusion and provide more useful information from the learning stage to the enhancement stage, we establish a new model for the noise in the time-frequency domain which considers both intra-frame (sparsity) and inter-frame (low-rank) characteristics. Specifically, we model the noise magnitude spectral matrix as

$$\boldsymbol{N} = \boldsymbol{L} + \boldsymbol{D}_n\boldsymbol{\Theta}_n \tag{10}$$

which jointly employs the low-rank matrix $\boldsymbol{L}$ and an overcomplete noise dictionary $\boldsymbol{D}_n$. This model will be employed in both the learning and the enhancement stages of our proposed approach.

The learning stage aims at training appropriate dictionaries for both speech and noise and learning the low-rank component of noise. At first, all the three available time-domain training datasets, consisting of clean speech, noise and noisy speech (obtained as the sum of the clean speech and noise with a predetermined signal-to-noise ratio (SNR)), are transformed into the time-frequency domain through STFT. Then,

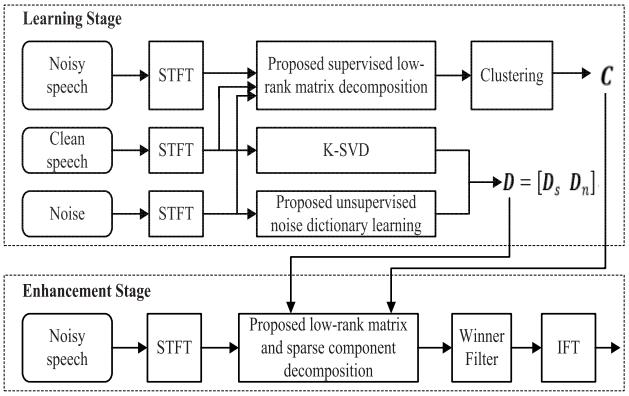


FIGURE 1. Block diagram of proposed speech enhancement method.

the unsupervised K-SVD algorithm is directly applied to the clean speech magnitude spectral matrix to build a speech dictionary $\boldsymbol{D}_s$. Furthermore, based on the new noise model in (10), we propose a new unsupervised dictionary learning algorithm to train the noise dictionary $\boldsymbol{D}_n$. It is worth noting that the trained speech dictionary $\boldsymbol{D}_s$ is input into this new noise dictionary learning algorithm as a reference to reduce the mutual coherence between the noise dictionary $\boldsymbol{D}_n$ and the speech dictionary $\boldsymbol{D}_s$. Finally, a supervised low-rank matrix decomposition algorithm is proposed to extract the low-rank component of the noise magnitude spectral matrix based on the trained dictionaries and all three training datasets. This estimated low-rank matrix $\boldsymbol{L}$ is then input into a clustering module to produce a reference codebook $\boldsymbol{C}$ for the enhancement stage. The codebook will provide the enhancement stage with useful information to further improve the speech quality and intelligibility.

In the enhancement stage, the noisy speech magnitude spectral matrix is decoupled via the proposed low-rank matrix and sparse component decomposition algorithm into the sum of a structured component, which is sparse in regards of the composite dictionary $\boldsymbol{D}$ and a low-rank component which is refined using the codebook $\boldsymbol{C}$. The speech magnitude spectral matrix is extracted from the estimated sparse component. A Wiener filter [1] is then employed to further improve the estimation of the speech magnitude spectra. Finally, the time-domain speech signal is synthesized via IFT of the estimated speech spectra using the estimated speech magnitude spectra and the unprocessed phases of the noisy speech spectra.

B. LEARNING STAGE

1) UNSUPERVISED DICTIONARY LEARNING

In the learning stage, we first utilize the unsupervised K-SVD algorithm to train the speech dictionary $\boldsymbol{D}_s$. Next, we propose to train the noise dictionary $\boldsymbol{D}_n$ based on the new noise model in (10), which can effectively reduce source confusion [6]. Before training the noise dictionary, we need to extract the low-rank component from the noise magnitude spectral matrix $\boldsymbol{N}$, which is considered to be highly coherent to the speech.

We apply the singular value decomposition (SVD) to the noise magnitude spectral matrix $\boldsymbol{N}$ and obtain

$$\boldsymbol{N} = \boldsymbol{U}_N \boldsymbol{\Lambda}_N \boldsymbol{V}_N{}^T \tag{11}$$

where both $\boldsymbol{U}_N$ and $\boldsymbol{V}_N$ are unitary matrices and $\boldsymbol{\Lambda}_N$ is a rectangular diagonal matrix with the singular values $\lambda_{N,i}$ ($i = 1, 2, \cdots, M$) of $\boldsymbol{N}$ on its principal diagonal. The average mutual coherence between speech, and noise components in the frequency domain is measured in our work as

$$\rho(i) = \frac{1}{P}\left\|\boldsymbol{U}_{N,i}^T \boldsymbol{D}_s\right\|_1, \quad i = 1, 2, \cdots, M \tag{12}$$

where $\boldsymbol{U}_{N,i}$ is the i^{th} column vector of $\boldsymbol{U}_N$. We then define the mutual coherence vector as

$$\boldsymbol{\rho} = \begin{bmatrix}\rho(1) & \rho(2) & \cdots & \rho(M)\end{bmatrix}^T. \tag{13}$$

It is evident to see a larger value of the average mutual coherence in (12) indicates a higher level of coherence between the corresponding noise component vector and the speech dictionary. We select the components of $\boldsymbol{N}$ corresponding to the indices of the first I largest entries in $\boldsymbol{\rho}$ as basis vectors in the construction of the low-rank matrix $\boldsymbol{L}$. In the sequel, the corresponding index set with cardinality I is denoted as A. Then the low-rank matrix can be approximated as

$$\widehat{\boldsymbol{L}} = \sum_{i \in A} \lambda_{N,i} \left\|\boldsymbol{U}_{N,i}^T \boldsymbol{D}_s\right\|_\infty \boldsymbol{U}_{N,i} \boldsymbol{V}_{N,i}^T. \tag{14}$$

We utilize the maximum mutual coherence values between the selected noise components and the speech dictionary as weights to approximate the low-rank component of the noise magnitude spectral matrix, and then use the residual noise magnitude spectral matrix to train the noise dictionary. This way will effectively reduce the coherence between the speech and the noise dictionary. Specifically, training of the noise dictionary $\boldsymbol{D}_n$ can be formulated as the following optimization problem,

$$\min_{\boldsymbol{D}_n, \boldsymbol{\Theta}_n} \left\|\boldsymbol{N} - \widehat{\boldsymbol{L}} - \boldsymbol{D}_n \boldsymbol{\Theta}_n\right\|_F^2 \text{ s.t. } \left\|\boldsymbol{\theta}_{n,i}\right\|_0 \leq K, \quad \forall i. \tag{15}$$

We can employ the K-SVD algorithm to solve this problem.

The complete procedure for implementing the above unsupervised noise dictionary learning is summarized in Algorithm 1.

2) SUPERVISED LOW-RANK MATRIX DECOMPOSITION FOR CODEBOOK CONSTRUCTION

In this part, we will take advantage of all three training datasets, i.e. the clean speech, noise and noisy speech, to learn the desired low-rank matrix to construct a codebook. This low-rank matrix decomposition will be carried out in a supervised fashion. We should mention that the low-rank matrix estimation in this part is somewhat different from that in III-B1. The low-rank matrix in (14) represents the components of noise which are highly coherent to speech. However, the low-rank matrix extraction here aims at capturing highly correlated components among the magnitude spectra of various noise frames and provides useful information through

Algorithm 1 Proposed Unsupervised Noise Dictionary Learning Algorithm

Input: $\boldsymbol{N}, \boldsymbol{D}_s, I$;

- SVD: $\boldsymbol{N} = \boldsymbol{U}_N \boldsymbol{\Lambda}_N \boldsymbol{V}_N{}^T$;
- Average mutual coherence values:
  1) $\rho(i) = \frac{1}{P}\left\|\boldsymbol{U}_{N,i}^T \boldsymbol{D}_s\right\|_1,\ i = 1, 2, \cdots, M$;
  2) $\boldsymbol{\rho} = \left[\rho(1)\ \rho(2) \cdots \rho(M)\right]^T$;
- Noise component selection:

$$A = \{\text{indices corresponding to the first } I \text{ largest magnitude elements in } \boldsymbol{\rho}\};$$

- Low-rank matrix approximation: $\widehat{\boldsymbol{L}} = \sum_{i\in A} \lambda_{N,i} \left\|\boldsymbol{U}_{N,i}^T \boldsymbol{D}_s\right\|_\infty \boldsymbol{U}_{N,i} \boldsymbol{V}_{N,i}^T$;
- Noise dictionary training: $\boldsymbol{D}_n$ =K-SVD($\boldsymbol{N} - \widehat{\boldsymbol{L}}$).

Output: $\boldsymbol{D}_n$.

construction of a codebook about the mapping of the noisy speech to noise during the enhancement stage.

With the new noise model in (10), the noisy speech magnitude spectral matrix $\boldsymbol{X}$ in the learning stage can be expressed as

$$\boldsymbol{X} = \boldsymbol{S} + \boldsymbol{N} = \boldsymbol{D}_s \boldsymbol{\Theta}_s + \boldsymbol{D}_n \boldsymbol{\Theta}_n + \boldsymbol{L} = \boldsymbol{D}\boldsymbol{\Theta} + \boldsymbol{L}, \quad (16)$$

where $\boldsymbol{D}_s$ and $\boldsymbol{D}_n$ are obtained respectively from the unsupervised K-SVD algorithm and the above proposed unsupervised noise dictionary learning algorithm, $\boldsymbol{D} = \left[\boldsymbol{D}_s\ \boldsymbol{D}_n\right]$ and $\boldsymbol{\Theta} = \left[\boldsymbol{\Theta}_s^T\ \boldsymbol{\Theta}_n^T\right]^T$. Based on (16), the estimation of the low-rank matrix $\boldsymbol{L}$ can be formulated as the following problem,

$$\begin{aligned} \min_{\boldsymbol{\Theta},\boldsymbol{L}}\ & \left\|\boldsymbol{X} - \boldsymbol{D}\boldsymbol{\Theta} - \boldsymbol{L}\right\|_F^2 + \tau_1 \left\|\boldsymbol{N} - \boldsymbol{L} - \boldsymbol{D}_n\boldsymbol{\Theta}_n\right\|_F^2 \\ & + \tau_2 \left\|\boldsymbol{S} - \boldsymbol{D}_s\boldsymbol{\Theta}_s\right\|_F^2 + \tau_3 \left\|\boldsymbol{\Theta}\right\|_1 \\ \text{s.t. }\ & \text{rank}(\boldsymbol{L}) \le r \end{aligned} \quad (17)$$

where τ_1, τ_2 and τ_3 are regularization factors, providing a tradeoff between the sparsity and the approximation errors in the noisy speech, noise and clean speech. The parameter r is used to guarantee the low-rank property of the matrix $\boldsymbol{L}$. The minimization problem in (17) can be addressed by alternate iteration between the following two subproblems.

$$\begin{aligned} \boldsymbol{\Theta}^t = \arg\min_{\boldsymbol{\Theta}}\ & \left\|\boldsymbol{X} - \boldsymbol{D}\boldsymbol{\Theta} - \boldsymbol{L}^{t-1}\right\|_F^2 \\ & + \tau_1 \left\|\boldsymbol{N} - \boldsymbol{L}^{t-1} - \boldsymbol{D}_n\boldsymbol{\Theta}_n\right\|_F^2 \\ & + \tau_2 \left\|\boldsymbol{S} - \boldsymbol{D}_s\boldsymbol{\Theta}_s\right\|_F^2 + \tau_3 \left\|\boldsymbol{\Theta}\right\|_1 \end{aligned} \quad (18)$$

$$\boldsymbol{L}^t = \arg\min_{\text{rank}(\boldsymbol{L})\le r} \left\|\boldsymbol{X} - \boldsymbol{D}\boldsymbol{\Theta}^t - \boldsymbol{L}\right\|_F^2 + \tau_1 \left\|\boldsymbol{N} - \boldsymbol{L} - \boldsymbol{D}_n\boldsymbol{\Theta}_n^t\right\|_F^2 \quad (19)$$

where the superscript t is the iteration index. With a fixed value of $\boldsymbol{L}$, (18) seeks to estimate the sparse coefficient matrix $\boldsymbol{\Theta}$ based on the ℓ_1-norm minimization. With a fixed estimate of the matrix $\boldsymbol{\Theta}$ from (18), the subproblem (19) seeks to estimate the low-rank matrix $\boldsymbol{L}$ with an explicit rank constraint. In the following, we expose the details of our solution approach for these two optimization problems.

By combining the first three terms in the objective function of problem (18) into a single term, the latter problem is seen to be equivalent to

$$\boldsymbol{\Theta}^t = \arg\min_{\boldsymbol{\Theta}} \left\|\boldsymbol{Y}^t - \boldsymbol{W}\boldsymbol{\Theta}\right\|_F^2 + \tau_3 \left\|\boldsymbol{\Theta}\right\|_1 \quad (20)$$

where $\boldsymbol{Y}^t = \begin{bmatrix} \boldsymbol{X} - \boldsymbol{L}^{t-1} \\ \sqrt{\tau_1}(\boldsymbol{N} - \boldsymbol{L}^{t-1}) \\ \sqrt{\tau_2}\boldsymbol{S} \end{bmatrix}$ and $\boldsymbol{W} = \begin{bmatrix} \boldsymbol{D}_s & \boldsymbol{D}_n \\ \boldsymbol{0} & \sqrt{\tau_1}\boldsymbol{D}_n \\ \sqrt{\tau_2}\boldsymbol{D}_s & \boldsymbol{0} \end{bmatrix}$. The above convex optimization problem (20) is a typical ℓ_1-norm based sparse decomposition problem and accordingly, we can exploit the LARC algorithm to find the globally optimal solution.

The subproblem (19) can be converted into the following optimization problem by combining the two terms in its objective:

$$\boldsymbol{L}^t = \arg\min_{\text{rank}(\boldsymbol{L})\le r} \left\| \left[\boldsymbol{X} - \boldsymbol{D}\boldsymbol{\Theta}^t\ \ \sqrt{\tau_1}(\boldsymbol{N} - \boldsymbol{D}_n\boldsymbol{\Theta}_n^t)\right] - \left[\boldsymbol{L}\ \ \sqrt{\tau_1}\boldsymbol{L}\right] \right\|_F^2 \quad (21)$$

Let $\boldsymbol{H}^t = \left[\boldsymbol{X} - \boldsymbol{D}\boldsymbol{\Theta}^t\ \sqrt{\tau_1}(\boldsymbol{N} - \boldsymbol{D}_n\boldsymbol{\Theta}_n^t)\right]$ and $\boldsymbol{Q} = \left[\boldsymbol{L}\ \sqrt{\tau_1}\boldsymbol{L}\right]$. Then problem (21) can be converted into

$$\boldsymbol{Q}^t = \arg\min_{\text{rank}(\boldsymbol{Q})\le r} \left\|\boldsymbol{H}^t - \boldsymbol{Q}\right\|_F^2. \quad (22)$$

We now utilize the singular value hard thresholding algorithm [10] to solve this problem. Firstly, SVD is applied to the matrix $\boldsymbol{H}^t$, i.e.,

$$\boldsymbol{H}^t = \boldsymbol{U}_H^t \boldsymbol{\Lambda}_H^t \boldsymbol{V}_H^t{}^T \quad (23)$$

where $\boldsymbol{U}_H^t$ and $\boldsymbol{V}_H^t$ are both unitary matrices and $\boldsymbol{\Lambda}_H^t$ is a rectangular diagonal matrix with singular values $\lambda_{H,i}^t$ ($i = 1, 2, \cdots, M$) of $\boldsymbol{H}^t$ on its principal diagonal. And then $\boldsymbol{Q}^t$ can then be approximated as

$$\boldsymbol{Q}^t = \sum_{i=1}^{r} \lambda_{H,i}^t \boldsymbol{U}_{H,i}^t \boldsymbol{V}_{H,i}^t{}^T \quad (24)$$

where $\lambda_{H,i}^t$ ($i = 1, 2, \cdots, r$) are the first r dominant singular values of $\boldsymbol{\Lambda}_H^t$, and $\boldsymbol{U}_{H,i}^t$ ($i = 1, 2, \cdots, r$) and $\boldsymbol{V}_{H,i}^t$ ($i = 1, 2, \cdots, r$) respectively represent the first r left-singular and right-singular vectors in $\boldsymbol{U}_H^t$ and $\boldsymbol{V}_H^t$. In accordance with the relationship between $\boldsymbol{Q}$ and $\boldsymbol{L}$, we can obtain the estimate $\boldsymbol{L}^t$ from $\boldsymbol{Q}^t$. Let matrix $\boldsymbol{Q}^t$ be partitioned into blocks of identical size, i.e.,

$$\boldsymbol{Q}^t = \left[\boldsymbol{Q}_1^t\ \boldsymbol{Q}_2^t\right]. \quad (25)$$

Recalling that $\boldsymbol{Q} = \left[\boldsymbol{L}\ \sqrt{\tau_1}\boldsymbol{L}\right]$, we have

$$\boldsymbol{L}^t = \mu\boldsymbol{Q}_1^t + \frac{(1-\mu)}{\sqrt{\tau_1}}\boldsymbol{Q}_2^t \quad (26)$$

where $\mu \in [0, 1]$ is a weight controlling the relative contribution of $\boldsymbol{Q}_1^t$ and $\boldsymbol{Q}_2^t$ to the estimation. Since matrix $\boldsymbol{Q}_1^t$ is related

Algorithm 2 Proposed Supervised Low-Rank Matrix Decomposition Algorithm

Input: Magnitude spectral matrices of the noisy speech $\boldsymbol{X}$, clean speech $\boldsymbol{S}$, noise $\boldsymbol{N}$, speech dictionary $\boldsymbol{D}_s$ and noise dictionary $\boldsymbol{D}_n$, rank r, weight parameter μ.

Initialization: Set initial estimate of the low-rank matrix $\boldsymbol{L}^0 = \boldsymbol{0}$. Set $\boldsymbol{W} = \begin{bmatrix} \boldsymbol{D}_s & \boldsymbol{D}_n \\ \boldsymbol{0} & \sqrt{\tau_1}\boldsymbol{D}_n \\ \sqrt{\tau_2}\boldsymbol{D}_s & \boldsymbol{0} \end{bmatrix}$.

Iteration: For $t = 1, 2, \cdots t_{\text{stop}}$:

- *Sparse Coding Stage*: Employ the LARC algorithm to estimate the sparse coefficient matrix.
  1) $\boldsymbol{Y}^t = \begin{bmatrix} \boldsymbol{X} - \boldsymbol{L}^{t-1} \\ \sqrt{\tau_1}(\boldsymbol{N} - \boldsymbol{L}^{t-1}) \\ \sqrt{\tau_2}\boldsymbol{S} \end{bmatrix}$;
  2) $\boldsymbol{\Theta}^t = \text{LARC}(\boldsymbol{Y}^t, \boldsymbol{W})$.
- *Low-Rank Matrix Estimation Stage*: Update the low-rank component of the noise magnitude spectral matrix.
  1) Partition $\boldsymbol{\Theta}^t = \left[\boldsymbol{\Theta}_s^{t\,T}\ \boldsymbol{\Theta}_n^{t\,T}\right]^T$, where two blocks of equal size represent the sparse coefficient matrix estimates of the speech and noise, respectively.
  2) $\boldsymbol{H}^t = \left[\boldsymbol{X} - \boldsymbol{D}\boldsymbol{\Theta}^t\ \sqrt{\tau_1}(\boldsymbol{N} - \boldsymbol{D}_n\boldsymbol{\Theta}_n^t)\right]$;
  3) Apply SVD to the matrix $\boldsymbol{H}^t$: $\boldsymbol{H}^t = \boldsymbol{U}_H^t \boldsymbol{\Lambda}_H^t \boldsymbol{V}_H^{t\,T}$;
  4) Approximate the matrix $\boldsymbol{Q}^t$ by $\boldsymbol{Q}^t = \sum_{i=1}^{r} \lambda_{H,i}^t \boldsymbol{U}_{H,i}^t \boldsymbol{V}_{H,i}^{t\,T}$;
  5) Partition $\boldsymbol{Q}^t$ to obtain two column blocks of the same size, $\boldsymbol{Q}_1^t$ and $\boldsymbol{Q}_2^t$;
  6) Update the low-rank matrix as $\boldsymbol{L}^t = \mu\boldsymbol{Q}_1^t + \frac{(1-\mu)}{\sqrt{\tau_1}}\boldsymbol{Q}_2^t$.

Output: $\widehat{\boldsymbol{\Theta}} = \boldsymbol{\Theta}^t, \widehat{\boldsymbol{L}} = \boldsymbol{L}^t$.

to the noisy speech and matrix $\boldsymbol{Q}_2^t$ is related to the noise, the parameter μ offers a tradeoff between the approximation of noisy speech and noise. In effect, (26) underlies an implied mapping from the noisy speech to the low-rank component of noise in the frequency domain, which can provide auxiliary information to the enhancement stage.

The pseudo-code of the proposed supervised low-rank matrix decomposition algorithm is presented in Algorithm 2.

The codebook, denoted as a matrix $\boldsymbol{C} \in \mathbb{R}^{M \times N_3}$, is established with the column vectors of the estimated low-rank matrix via the traditional K-means clustering method [15]. Specifically, we apply the K-means method to the column vectors in the estimated low-rank matrix from Algorithm 2 to obtain multiple centroids and then store these centroids as the column vectors in $\boldsymbol{C}$.

C. ENHANCEMENT STAGE

Based on both the noise and speech models established in the learning stage, here we propose to use the composite dictionary $\boldsymbol{D}$, which is a concatenation of the speech dictionary $\boldsymbol{D}_s$ and the noise dictionary $\boldsymbol{D}_n$, to sparsely represent the structured component of the mixture signal, and exploit the rank deficient matrix $\boldsymbol{L}$ to capture the low-rank component among the noisy speech segments. Thus, the noisy speech magnitude spectral matrix $\boldsymbol{X}$ can be modeled as

$$\boldsymbol{X} = \boldsymbol{D}\boldsymbol{\Theta} + \boldsymbol{L}, \tag{27}$$

where $\boldsymbol{D}\boldsymbol{\Theta}$ refers to the structured component and $\boldsymbol{L}$ represents the low-rank component.

The estimate of the sparse coefficient matrix $\boldsymbol{\Theta}$ and the low-rank matrix $\boldsymbol{L}$ can be formulated in terms of the following optimization problem,

$$\begin{aligned} &\min_{\boldsymbol{\Theta},\boldsymbol{L}} \left\|\boldsymbol{X} - \boldsymbol{D}\boldsymbol{\Theta} - \boldsymbol{L}\right\|_F^2 + \tau_4 \left\|\boldsymbol{\Theta}\right\|_1 \\ &\text{s.t. } \text{rank}(\boldsymbol{L}) \le r \end{aligned} \tag{28}$$

where τ_4 is a regulation factor, allowing a trade-off between the sparsity level of $\boldsymbol{\Theta}$ and the approximation error of the noisy speech magnitude spectra. This problem can be addressed through alternating iterations between the estimation of the low-rank matrix $\boldsymbol{L}$ and the sparse coefficient matrix $\boldsymbol{\Theta}$, as per the following subproblems,

$$\boldsymbol{L}^t = \underset{\text{rank}(\boldsymbol{L}) \le r}{\arg\min} \left\|\boldsymbol{X} - \boldsymbol{D}\boldsymbol{\Theta}^{t-1} - \boldsymbol{L}\right\|_F^2, \tag{29}$$

$$\boldsymbol{\Theta}^t = \underset{\boldsymbol{\Theta}}{\arg\min} \left\|\boldsymbol{X} - \boldsymbol{D}\boldsymbol{\Theta} - \boldsymbol{L}^t\right\|_F^2 + \tau_4 \left\|\boldsymbol{\Theta}\right\|_1. \tag{30}$$

Within this iterative process in the enhancement stage, the codebook $\boldsymbol{C}$ constructed in the learning stage is utilized as a reference to refine the low-rank matrix estimation. Specifically, the optimization problem in (29) can be solved through SVD of the residual matrix $\boldsymbol{R}^t = \boldsymbol{X} - \boldsymbol{D}\boldsymbol{\Theta}^{t-1}$, i.e.,

$$\boldsymbol{R}^t = \boldsymbol{X} - \boldsymbol{D}\boldsymbol{\Theta}^{t-1} = \boldsymbol{U}_R^t \boldsymbol{\Lambda}_R^t \boldsymbol{V}_R^{t\,T}. \tag{31}$$

where $\boldsymbol{U}_R^t$ and $\boldsymbol{V}_R^t$ are two unitary matrices and $\boldsymbol{\Lambda}_R^t$ is a rectangular diagonal matrix, containing the singular values of $\boldsymbol{R}^t$ on its principal diagonal. The solution to (29) can then be expressed as

$$\boldsymbol{L}^t = \sum_{i=1}^{r} \lambda_{R,i}^t \boldsymbol{U}_{R,i}^t \boldsymbol{V}_{R,i}^{t\,T} \tag{32}$$

where $\lambda_{R,i}^t$ $(i = 1, 2, \cdots, r)$ are the r largest singular values of $\boldsymbol{R}^t$ and $\boldsymbol{U}_{R,i}^t$ and $\boldsymbol{V}_{R,i}^t$ are, respectively, the i^{th} column vectors in the matrix $\boldsymbol{U}_R^t$ and $\boldsymbol{V}_R^t$. We then refine the columns of the matrix $\boldsymbol{L}^t$ according to the codebook $\boldsymbol{C}$ to obtain an improved matrix estimate $\widetilde{\boldsymbol{L}}^t$, which can be described as

$$\widetilde{\boldsymbol{L}}^t = \text{codebooktuning}(\boldsymbol{L}^t, \boldsymbol{C}, k). \tag{33}$$

In detail, for each column vector of $\boldsymbol{L}$, we find the k nearest column vectors in the codebook $\boldsymbol{C}$, average these k vectors, and use this average to replace the corresponding column in $\boldsymbol{L}^t$. Subsequently, we replace the matrix $\boldsymbol{L}^t$ in (30) with the matrix $\widetilde{\boldsymbol{L}}^t$ and the optimization problem in (30) can then be solved with the LARC algorithm, i.e.,

$$\boldsymbol{\Theta}^t = \text{LARC}(\boldsymbol{X} - \widetilde{\boldsymbol{L}}^t, \boldsymbol{D}). \tag{34}$$

Algorithm 3 Proposed Low-Rank Matrix and Sparse Component Decomposition Algorithm

Input: Composite dictionary $\boldsymbol{D}$, noisy speech magnitude spectral matrix $\boldsymbol{X}$, codebook $\boldsymbol{C}$, parameter k, rank r;
Initialization: Set $\boldsymbol{\Theta}^0 = \boldsymbol{0}$.
Iteration: For $t = 1, 2 \cdots, t_{\text{stop}}$:

- *Low-Rank Matrix Approximation Stage:*
  1) $\boldsymbol{R}^t = \boldsymbol{X} - \boldsymbol{D}\boldsymbol{\Theta}^{t-1}$;
  2) Apply SVD to $\boldsymbol{R}^t$: $\boldsymbol{R}^t = \boldsymbol{U}_R^t \boldsymbol{\Lambda}_R^t \boldsymbol{V}_R^{t\,T}$;
  3) $\boldsymbol{L}^t = \sum_{i=1}^{r} \lambda_{R,i}^t \boldsymbol{U}_{R,i}^t \boldsymbol{V}_{R,i}^{t\,T}$
  4) Codebook tuning: $\widetilde{\boldsymbol{L}}^t = \text{codebooktuning}(\boldsymbol{L}^t, \boldsymbol{C}, k)$;
- *Sparse Coding Stage:* $\boldsymbol{\Theta}^t = \text{LARC}(\boldsymbol{X} - \widetilde{\boldsymbol{L}}^t, \boldsymbol{D})$.

Output: $\widehat{\boldsymbol{\Theta}} = \boldsymbol{\Theta}^t, \widehat{\boldsymbol{L}} = \widetilde{\boldsymbol{L}}^t$.

The pseudo-code of the proposed algorithm for jointly estimating the sparse coefficient matrix $\boldsymbol{\Theta}$ and the low-rank matrix $\boldsymbol{L}$ in the enhancement stage is presented in Algorithm 3. As explained, the mapping from $\boldsymbol{L}^t$ to $\widetilde{\boldsymbol{L}}^t$ in step 4 of this algorithm aims to reduce the mismatch in noise information between the learning stage and the enhancement stage.

With the estimated sparse coefficient matrix $\widehat{\boldsymbol{\Theta}}$ and low-rank matrix $\widehat{\boldsymbol{L}}$ from Algorithm 3, the submatrices $\widehat{\boldsymbol{\Theta}}_s$ and $\widehat{\boldsymbol{\Theta}}_n$ are extracted via block row partitioning of matrix $\widehat{\boldsymbol{\Theta}}$ and the speech magnitude spectral matrix is estimated as

$$\widehat{\boldsymbol{S}} = \boldsymbol{D}_s \widehat{\boldsymbol{\Theta}}_s. \tag{35}$$

Moreover, we can obtain the noise magnitude spectral matrix as

$$\widehat{\boldsymbol{N}} = \boldsymbol{D}_n \widehat{\boldsymbol{\Theta}}_n + \widehat{\boldsymbol{L}}. \tag{36}$$

Wiener filtering is then applied to improve the estimation of speech magnitude spectra as

$$\bar{S}_{i,j} = \frac{\widehat{S}_{i,j}^2}{\widehat{S}_{i,j}^2 + \widehat{N}_{i,j}^2} X_{i,j} \tag{37}$$

where $\bar{S}_{i,j}$ represents the entry in the i^{th} row and j^{th} column of the enhanced speech magnitude spectral matrix $\bar{\boldsymbol{S}}$ while $\widehat{S}_{i,j}$, $\widehat{N}_{i,j}$ and $X_{i,j}$ are corresponding entries of matrices $\widehat{\boldsymbol{S}}$, $\widehat{\boldsymbol{N}}$ and $\boldsymbol{X}$. Finally, we apply IFT to the estimated speech spectra, namely the estimated speech magnitude spectra and mixture phases, to produce the target speech signal in the time domain.

IV. EXPERIMENTAL RESULTS

In this section, we evaluate our proposed speech enhancement method based on the Grid corpus [16] and NOISEX-92 corpus [17]. From the training stage, we randomly select 5 female and 5 male speakers from the Gird corpus and for each speaker, we randomly select 70 utterances. For the enhancement stage, we select another 4 female and 4 male speakers from the same corpus, and for each speaker, we select 5 utterances to form the test dataset. We emphasize that in all experiments described below, the speakers in the enhancement stage are different from those in the learning stage, i.e. speaker independent application [6]. Six types of noise are selected from the NOISEX-92 corpus, including babble noise (bab), factory noise (fct), f16 cockpit noise (f16), buccaneer cockpit noise (buc), hfchannel noise (hf) and white noise (wht). All signals are downsampled to 8KHz and segmented with a Hanning window, with a length of 512 and an overlap of 50%. The dictionary size is three times the frame length. The various parameters involved in the optimization techniques of our method (i.e. Algorithm 1, 2 and 3) are set as listed in Table 1. All the experiments were performed with MATLABR2018a (64bit) in a desktop with an Intel i7-8700 CPU (3.2GHz) and 16GB RAM.

TABLE 1. Parameters configuration in the proposed speech enhancement method.

Algorithm 1	Algorithm 2		Algorithm 3	
I	r	μ	k	r
4	2	0.3	4	2

The segmental SNR (SSNR) [18] and perceptual evaluation of speech quality (PESQ) [19] are employed in this section to evaluate the resynthesized speech quality, while the short-time objective intelligibility (STOI) [20] score is used for speech intelligibility evaluation. In addition, we employ a composite measure denoted as OVL in [19] to rate the overall speech quality. OVL is a linear combination of three objective measures including PESQ, log likelihood ratio (LLR) [21] and weighted-slope spectral (WSS) distance [22]. Moreover, we compare all the experimental results of our proposed method with five state-of-the-art speech enhancement methods involving GDL [6], CJSR [9], RPCA [11], LSLD [13] and CLSMD [10].

As the low-frequency dominance in magnitude spectra of male speakers is more conspicuous than that of female speakers, frequency components of male speakers are sparser than that of the opposite gender. Thus, we conduct simulation experiments and present corresponding performance evaluation results respectively for male and female speakers.

A. SPEECH ENHANCEMENT PERFORMANCE EVALUATION FOR MALE SPEAKERS

In this part, we compare the performance of our proposed method in improving the speech quality and intelligibility for male speakers with 5 different reference methods. Table 2 presents average PESQ scores of our proposed method and all the reference methods for the 6 types of noise at 3 different levels of SNR, i.e., 0dB, 5dB and 10dB. Our proposed method outperforms all the reference methods with regard to the PESQ scores of the processed male speech in the different noise scenarios. Moreover, it can achieve a higher gain in PESQ scores under low-SNR noise conditions. For example, the average improvement in PESQ score of the processed male speech with our proposed method across the six types

TABLE 2. PESQ scores for male speakers of six different speech enhancement methods for six different types of noise at three different levels of SNRs.

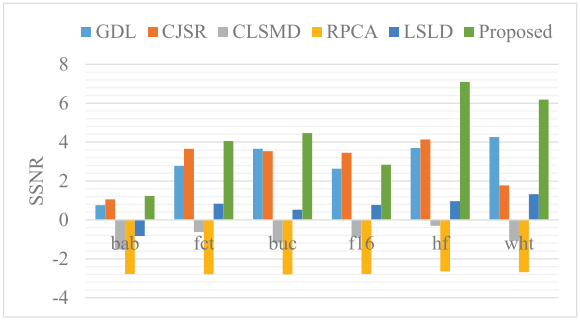
		GDL	CJSR	CLSMD	RPCA	LSLD	Proposed
	bab	2.01	1.85	1.86	1.91	1.90	2.07
	fct	2.25	2.39	2.06	1.97	2.13	2.45
	buc	2.14	2.07	1.64	1.62	1.73	2.47
0dB	f16	2.16	2.16	1.84	1.79	1.92	2.57
	wht	1.99	1.78	1.53	1.52	1.64	2.45
	hf	2.06	1.97	1.61	1.58	1.66	2.39
	Ave	2.10	2.04	1.76	1.73	1.83	2.4
	bab	2.30	2.21	2.08	2.17	2.14	2.37
	fct	2.51	2.68	2.22	2.24	2.36	2.75
	buc	2.38	2.45	1.87	1.83	2.01	2.69
5dB	f16	2.48	2.50	2.05	2.02	2.20	2.59
	wht	2.22	2.17	1.72	1.71	1.88	2.63
	hf	2.28	2.22	1.84	1.74	1.83	2.69
	Ave	2.36	2.37	1.96	1.95	2.07	2.62
	bab	2.57	2.60	2.23	2.44	2.38	2.68
	fct	2.74	2.82	2.28	2.50	2.55	2.99
	buc	2.57	2.64	2.11	2.09	2.33	2.82
10dB	f16	2.68	2.69	2.23	2.28	2.46	2.88
	wht	2.42	2.60	2.00	1.97	2.10	2.77
	hf	2.48	2.46	2.18	1.96	1.98	2.87
	Ave	2.58	2.64	2.17	2.21	2.3	2.84

of noise at 0dB SNR is 0.3, 0.36, 0.64, 0.67 and 0.57, respectively, over GDL, CJSR, CLSMD, RPCA and LSLD.

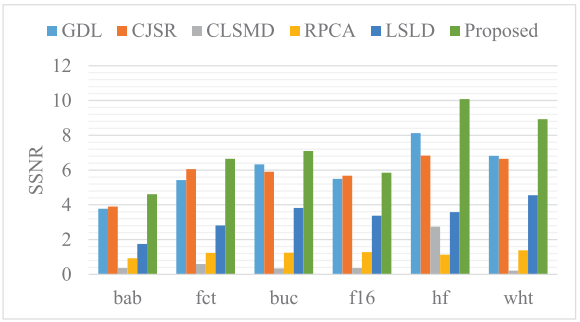
Fig. 2 depicts the average SSNR results of the processed male speech from our proposed method and the five reference methods for the six types of noise at SNR=0dB (a), 5dB (b) and 10dB (c). Our proposed method can achieve higher gain in SSNR than the reference methods irrespective of the noise type and the SNR level, except when the clean speech was corrupted by the f16 noise at the 0dB SNR. For example, in the case of hfchannel noise at 0dB SNR, the average SSNR gain of the processed male speech with our proposed method is 3.4 dB, 2.96dB, 7.41dB, 9.75dB and 6.14dB, respectively, over GDL, CJSR, CLSMD, RPCA and LSLD.

Fig. 3 presents the average OVL results of the processed speech from all the methods used in this paper. The results in all the subplots reveal that the enhanced speech from our proposed method exhibits better overall quality than those from the benchmark approaches. For instance, in the case of the factory noise at 5dB SNR, the OVL of the enhanced speech obtained from our proposed method is 3.16, compared to 2.78, 2.99, 2.17, 2.77 and 2.35, respectively, for GDL, CJSR, CLSMD, RPCA and LSLD.

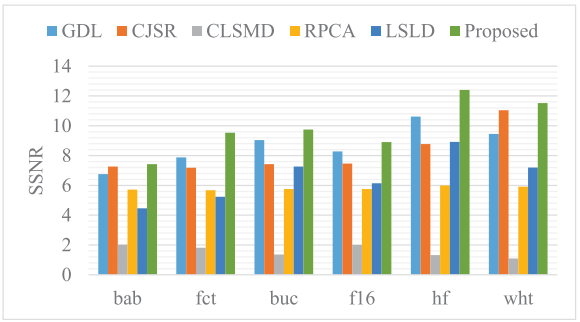
In Table 3, the average STOI scores of the processed male speech from our proposed method are higher than those of the reference methods. For example, when the SNR is 10dB, our proposed method can achieve an average gain in the STOI score across the six different types of noise as 0.02, 0.06, 0.14, 0.02 and 0.05, respectively, over GDL, CJSR, CLSMD, RPCA and LSLD. It is worth mentioning that compared with



(a)



(b)



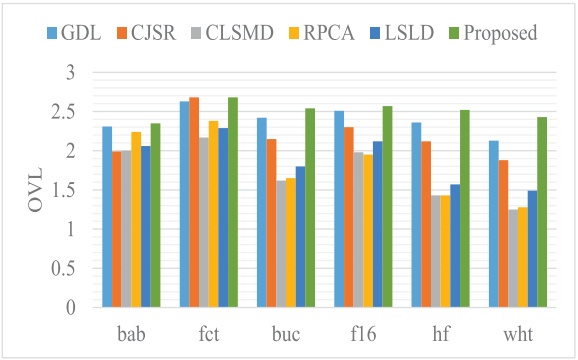
(c)

FIGURE 2. Average SSNR results of male speech with six different methods for six different types of noise. (a) SNR=0dB, (b) SNR=5dB, (c) SNR=10dB.

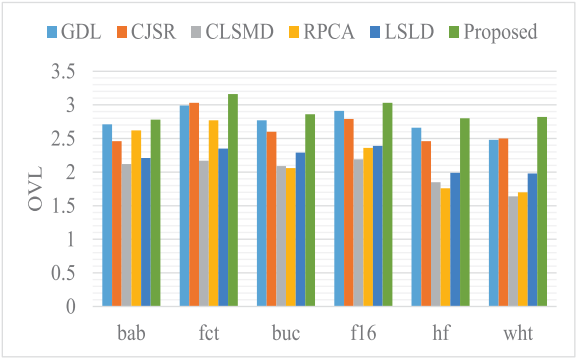
GDL, the CJSR method can further improve the speech quality but the intelligibility of the processed speech is reduced. However, our proposed method can improve both the speech quality and intelligibility.

B. SPEECH ENHANCEMENT PERFORMANCE EVALUATION FOR FEMALE SPEAKERS

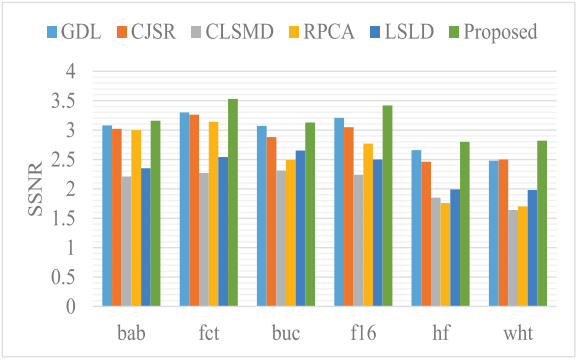
Here, we present the corresponding performance evaluation results of the processed female speech utterances based on the above four measures. As illustrated in Table 4, our proposed method still achieves the highest average PESQ scores across the six different types of noise irrespective of the SNR levels. The average PESQ score for the proposed method at 0dB is 2.16, an increase of 0.34, 0.22, 0.71, 0.76 and 0.51, respectively, over GDL, CJSR, CLSMD, RPCA and LSLD. Fig. 4 shows the average SSNR results of the processed



(a)



(b)



(c)

FIGURE 3. Average OVL results of the processed male speech from six different speech enhancement methods for six different types of noise. (a) SNR=0dB, (b) SNR=5dB, (c) SNR=10dB.

female speech from the proposed and reference methods for the six different types of noise and three SNR levels. The proposed method can achieve higher SSNRs than all the reference methods in the babble, factory, buccaneer, hfchannel and white noise environments. However, when female speech is corrupted by the f16 noise, the SSNR for our proposed method is lower than that for the CJSR method. Fig. 5 shows that our proposed method can achieve better overall speech quality for female speakers than the reference methods under different noise environments. Moreover, as illustrated in Table 5, the average STOI scores of the enhanced female speech from our proposed method are higher than those from

TABLE 3. STOI results of the processed male speech from six different speech enhancement methods for six different types of noise at three different levels of SNRs.

		GDL	CJSR	CLSMD	RPCA	LSLD	Proposed
0dB	bab	0.68	0.63	0.62	0.68	0.64	0.72
	fct	0.80	0.80	0.72	0.79	0.76	0.81
	buc	0.75	0.65	0.59	0.64	0.66	0.74
	f16	0.78	0.75	0.68	0.70	0.73	0.78
	wht	0.74	0.75	0.61	0.68	0.71	0.76
	hf	0.77	0.62	0.66	0.70	0.70	0.78
	Ave	0.75	0.70	0.65	0.70	0.70	0.77
5dB	bab	0.80	0.76	0.75	0.80	0.75	0.83
	fct	0.88	0.87	0.79	0.87	0.83	0.89
	buc	0.83	0.75	0.70	0.76	0.77	0.83
	f16	0.86	0.82	0.74	0.81	0.81	0.88
	wht	0.80	0.83	0.67	0.77	0.78	0.83
	hf	0.83	0.69	0.74	0.77	0.81	0.86
	Ave	0.83	0.79	0.73	0.65	0.79	0.85
10dB	bab	0.88	0.87	0.79	0.90	0.84	0.89
	fct	0.92	0.91	0.82	0.93	0.88	0.94
	buc	0.88	0.79	0.75	0.86	0.85	0.88
	f16	0.90	0.89	0.79	0.90	0.87	0.93
	wht	0.86	0.88	0.74	0.85	0.83	0.89
	hf	0.89	0.76	0.73	0.87	0.87	0.90
	Ave	0.89	0.85	0.77	0.89	0.86	0.91

TABLE 4. PESQ scores for female speakers of six different speech enhancement methods for six different types of noise at three different levels of SNRs.

		GDL	CJSR	CLSMD	RPCA	LSLD	Proposed
0dB	bab	1.70	1.76	1.50	1.49	1.63	1.89
	fct	1.89	2.16	1.70	1.59	1.96	2.25
	buc	1.84	2.13	1.34	1.27	1.56	2.21
	f16	1.94	2.24	1.55	1.45	1.76	2.21
	wht	1.72	1.59	1.29	1.25	1.47	2.28
	hf	1.80	1.76	1.34	1.32	1.51	2.14
	Ave	1.82	1.94	1.45	1.40	1.65	2.16
5dB	bab	2.11	2.20	1.72	1.84	1.91	2.23
	fct	2.22	2.66	1.83	1.95	2.20	2.64
	buc	2.14	2.46	1.56	1.54	1.92	2.44
	f16	2.27	2.60	1.73	1.76	2.08	2.51
	wht	2.03	2.08	1.47	1.48	1.75	2.47
	hf	2.10	2.15	1.50	1.54	1.80	2.46
	Ave	2.15	2.36	1.64	1.69	1.94	2.45
10dB	bab	2.40	2.55	1.88	2.19	2.11	2.48
	fct	2.47	2.90	1.88	2.28	2.38	2.90
	buc	2.37	2.61	1.79	1.88	2.22	2.63
	f16	2.52	2.81	1.86	2.10	2.33	2.76
	wht	2.29	2.56	1.73	1.80	2.00	2.67
	hf	2.39	2.44	1.72	1.83	2.07	2.65
	Ave	2.41	2.65	1.81	2.01	2.19	2.67

the reference methods. At 0dB SNR, the average gain of our proposed method across the six noise types in STOI scores is 0.01, 0.03, 0.08, 0.04 and 0.03, respectively, over GDL,

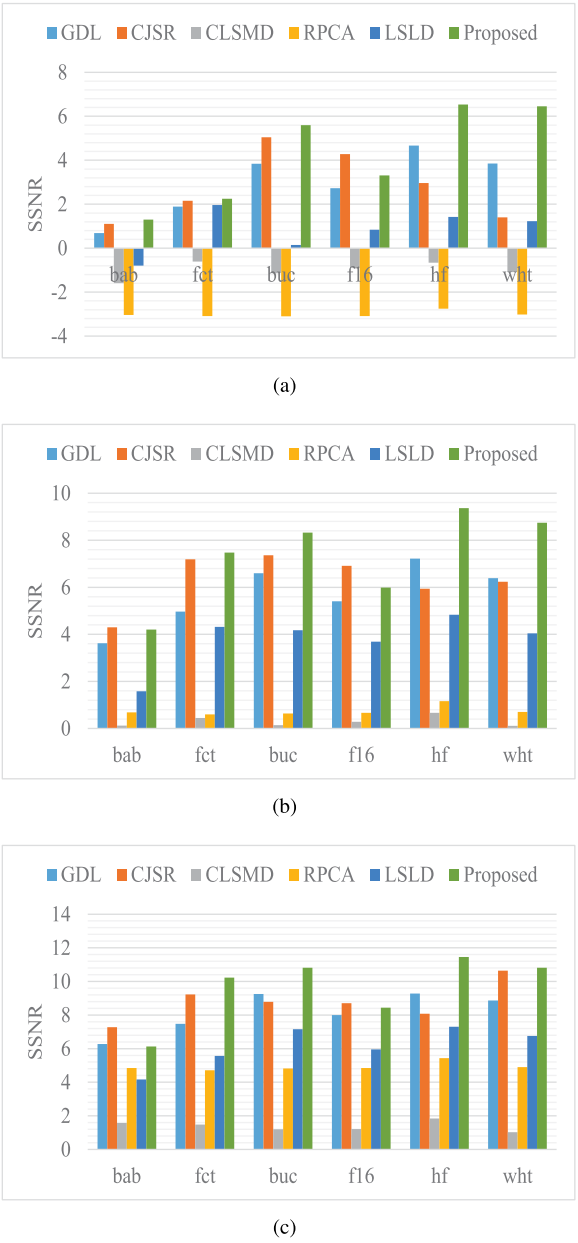


FIGURE 4. Average SSNR results of the processed female speech from six different methods for six different types of noise. (a) SNR=0dB, (b) SNR=5dB, (c) SNR=10dB.

FIGURE 5. Average OVL results of the processed female speech from six different methods for six different types of noise. (a) SNR=0dB, (b) SNR=5dB, (c) SNR=10dB.

CJSR, CLSMD, RPCA and LSLD. In general, our proposed method can achieve better speech quality and intelligibility for female speakers than the reference methods, especially under low-SNR noise conditions.

C. COMPUTATIONAL EFFICIENCY

In this part, we investigate the computational complexity of the speech enhancement stage of the proposed method. The dominant operations of Algorithm 3 are the sparse coding algorithm LARC and the SVD. According to [6], the computational cost of the LARC algorithm is $O(P^2\ M)$. Provided that the number of the frames which are processed simultaneously in the enhancement stage is denoted as F, the full SVD is performed at $O(MF^2)$ flops. However, in consideration of the r-rank approximation of the matrix $\boldsymbol{L}^t$, we can just conduct the partial SVD with the overall cost $O(rMF)$. Thus, the total computational cost of Algorithm 3 is $O(P^2\ M + rMF)$. As $r \ll F \ll P$, it is about $O(P^2\ M)$. And the theoretical computational cost is $O(P^2M)$ for GDL, $O(P^2M)$ for CJSR, $O(rMF)$ for CLSMD, $O(P^2F)$ for RPCA and $O(P^3)$ for LSLD, respectively. The average running time is 3.66s for the proposed method, 3.05s for the GDL, 3.80s for the CJSR, 0.14s for the CLSMD, 5.9s for the RPCA and 12.88s for the LSLD, respectively, which agrees with the theoretical computational complexity.

TABLE 5. STOI results of the processed female speech from six different speech enhancement methods for six different types of noise at three different levels of SNRs.

		GDL	CJSR	CLSMD	RPCA	LSLD	Proposed
0dB	bab	0.71	0.68	0.66	0.70	0.68	0.72
	fct	0.78	0.80	0.74	0.77	0.78	0.78
	buc	0.71	0.68	0.66	0.70	0.68	0.72
	f16	0.77	0.77	0.68	0.71	0.74	0.78
	wht	0.71	0.74	0.63	0.69	0.70	0.74
	hf	0.74	0.66	0.65	0.70	0.73	0.74
	Ave	0.74	0.72	0.67	0.71	0.72	0.75
5dB	bab	0.82	0.76	0.70	0.75	0.78	0.81
	fct	0.86	0.87	0.79	0.86	0.84	0.88
	buc	0.82	0.76	0.70	0.75	0.78	0.81
	f16	0.84	0.85	0.73	0.81	0.82	0.85
	wht	0.78	0.82	0.68	0.77	0.78	0.80
	hf	0.82	0.73	0.71	0.80	0.81	0.83
	Ave	0.82	0.80	0.71	0.79	0.80	0.83
10dB	bab	0.87	0.86	0.80	0.89	0.84	0.86
	fct	0.90	0.91	0.83	0.92	0.89	0.93
	buc	0.87	0.79	0.77	0.86	0.85	0.86
	f16	0.89	0.89	0.79	0.89	0.88	0.90
	wht	0.83	0.87	0.75	0.85	0.83	0.88
	hf	0.85	0.78	0.77	0.88	0.87	0.87
	Ave	0.87	0.85	0.79	0.88	0.86	0.88

V. CONCLUSION

In contrast to the existing sparse-model based speech enhancement methods, a new monaural speech enhancement framework has been proposed in this paper by introducing a new noise model to reduce the mutual coherence between speech and the noise dictionary and constructing a codebook to provide auxiliary information for noise estimation. The new noise model is utilized during both the learning and enhancement stages to decompose the noise into a sum of a low-rank component and a sparsely structured component. In the learning stage, the unsupervised Algorithm 1 is proposed to train a new noise dictionary with the extracted sparse component of the noise magnitude spectral matrix, which presents low mutual coherence to the speech. Then, the supervised Algorithm 2 is developed to carry out the low-rank matrix decomposition using all the training datasets consisting of clean speech, noise and noisy speech. The K-means clustering algorithm has subsequently been applied to this low-rank matrix to construct a reference codebook. In the enhancement stage, Algorithm 3 is proposed to estimate the magnitude spectra of the clean speech through decomposing the noisy speech magnitude spectral matrix into a structured component, which is sparsely represented with a composite dictionary, and a low-rank component which is refined by the reference codebook. Experimental results show that our proposed method outperforms the five reference methods in terms of PESQ, SSNR, OVL and STOI scores.

REFERENCES

[1] P. C. Loizou, "Evaluating of speech enhancement algorithms," in *Speech Enhancement: Theory and Practice*, 2nd ed. New York, NY, USA: CRC Press, 2007, pp. 439–475.
[2] R. Rubinstein, A. M. Bruckstein, and M. Elad, "Dictionaries for sparse representation modeling," *Proc. IEEE*, vol. 98, no. 6, pp. 1045–1057, Jun. 2010.
[3] Z. Zhang, Y. Xu, J. Yang, X. Li, and D. Zhang, "A survey of sparse representation: Algorithms and applications," *IEEE Access*, vol. 3, pp. 490–530, May 1965.
[4] I. Tosic and P. Frossard, "Dictionary learning," *IEEE Signal Process. Mag.*, vol. 28, no. 2, pp. 27–38, Mar. 2011.
[5] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky, "Sparse and low-rank matrix decompositions," in *Proc. Allerton*, Monticello, IL, USA, 2009, pp. 962–967.
[6] C. D. Sigg, T. Dikk, and J. M. Buhmann, "Speech enhancement using generative dictionary learning," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 6, pp. 1698–1712, Aug. 2012.
[7] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.
[8] A. M. Bruckstein, D. L. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM Rev.*, vol. 51, no. 1, pp. 34–81, 2009.
[9] Y. Luo, G. Bao, Y. Xu, and Z. Ye, "Supervised monaural speech enhancement using complementary joint sparse representations," *IEEE Signal Process. Lett.*, vol. 23, no. 2, pp. 237–241, Feb. 2016.
[10] C. Sun, Q. Zhu, and M. Wan, "A novel speech enhancement method based on constrained low-rank and sparse matrix decomposition," *Speech Commun.*, vol. 60, pp. 44–55, May 2014.
[11] Z. Chen and D. P. W. Ellis, "Speech enhancement by sparse, low-rank, and dictionary spectrogram decomposition," in *Proc. WASPAA*, New Paltz, NY, USA, Oct. 2013, pp. 1–4.
[12] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, Jan. 2011.
[13] S. Mavaddaty, S. M. Ahadi, and S. Seyedin, "A novel speech enhancement method by learnable sparse and low-rank decomposition and domain adaptation," *Speech Commun.*, vol. 76, pp. 42–60, Feb. 2016.
[14] S. Liu, J. Jia, Y. D. Zhang, and Y. Yang, "Image reconstruction in electrical impedance tomography based on structure-aware sparse Bayesian learning," *IEEE Trans. Med. Imag.*, vol. 37, no. 9, pp. 2090–2102, Sep. 2018.
[15] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1982.
[16] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *J. Acoust. Soc. Amer.*, vol. 120, no. 5, pp. 2421–2424, Nov. 2006.
[17] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–251, Jul. 1993.
[18] J. Du and Q. Huo, "A speech enhancement approach using piecewise linear approximation of an explicit model of environmental distortions," in *Proc. Interspeech*, Brisbane, QLD, Australia, 2008, pp. 569–572.
[19] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.
[20] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.
[21] S. R. Quackenbush, "Objective measures of speech quality (subjective)," Ph.D. dissertation, School Elect. Eng., Georgia Inst. Technol., Altanta, GA, USA, 1985.
[22] D. Klatt, "Prediction of perceived phonetic distance from critical-band spectra: A first step," in *Proc. ICASSP*, Paris, France, May 1982, pp. 1278–1281.

YUNYUN JI received the B.E. degree in communication engineering from the Hunan Institute of Engineering, Xiangtan, China, in 2009, and the Ph.D. degree in signal and information processing from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2014. Since 2014, she has been a Lecturer with Nantong University, Nantong, China. Since 2017, she has been a Post-Doctoral Fellow with Concordia University, Montreal, Canada.

WEI-PING ZHU (SM'97) received the B.E. and M.E. degrees from the Nanjing University of Posts and Telecommunications, in 1982, and the Ph.D. degree from Southeast University, Nanjing, China, in 1985 and 1991, respectively, all in electrical engineering. He was a Post-Doctoral Fellow and a Research Associate with the Department of Electrical and Computer Engineering, Concordia University, Montreal, Canada, from 1991 to 1992 and from 1996 to 1998, respectively. From 1993 to 1996, he was an Associate Professor with the Department of Information Engineering, Nanjing University of Posts and Telecommunications. From 1998 to 2001, he was with hi-tech companies in Ottawa, Canada, including Nortel Networks and SR Telecom, Inc. Since 2001, he has been with the Electrical and Computer Engineering Department, Concordia University, as a full-time Faculty Member, where he is currently a Full Professor. Since 2008, he has been an Adjunct Professor with the Nanjing University of Posts and Telecommunications, Nanjing. His research interests include digital signal processing fundamentals, speech and statistical signal processing, and signal processing for wireless communication with a particular focus on MIMO systems and cooperative communication.

Dr. Zhu was the Secretary of the Digital Signal Processing Technical Committee (DSPTC) of the IEEE Circuits and System Society, from 2012 to 2014, and the Chair of the DSPTC, from 2014 to 2016. He served as an Associate Editor for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS PART I: FUNDAMENTAL THEORY AND APPLICATIONS, from 2001 to 2003, *Circuits, Systems and Signal Processing* from 2006 to 2009, and the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS PART II: Transactions Briefs, from 2011 to 2015. He was also a Guest Editor of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS for the special issues of Broadband Wireless Communications for High Speed Vehicles and Virtual MIMO, from 2011 to 2013. He is currently an Associate Editor of the *Journal of The Franklin Institute*.

BENOIT CHAMPAGNE received the B.Ing. degree in engineering physics from the École Polytechnique de Montréal, in 1983, the M.Sc. degree in physics from the Université de Montréal, in 1985, and the Ph.D. degree in electrical engineering from the University of Toronto, in 1990. From 1990 to 1999, he was an Assistant and, then, an Associate Professor with INRS-Telecommunications, Université du Quebec, Montréal. In1999, he joined McGill University, Montreal, where he is currently a Full Professor with the Department of Electrical and Computer Engineering; he also served as an Associate Chairman of Graduate Studies with the Department, from 2004 to 2007. His research focuses on the study of advanced algorithms for the processing of communication signals by digital means. His interests span many areas of statistical signal processing, including detection and estimation, sensor array processing, adaptive filtering, and applications thereof to broadband communications and audio processing, in which he has co-authored about 300 referred publications. His research has been funded by the Natural Sciences and Engineering Research Council of Canada, the Fonds de Recherche sur la Nature et les Technologies from the Government of Quebec, and some major industrial sponsors, including Nortel Networks, Bell Canada, InterDigital, and Microsemi.

Dr. Champagne was a Registration Chair of IEEE ICASSP 2004, a Co-Chair of the Antenna and Propagation Track of IEEE VTC-Fall 2004, a Co-Chair of the Wide Area Cellular Communications Track of IEEE PIMRC 2011, a Co-Chair of the Workshop on D2D Communications of IEEE ICC 2015, and a Publicity Chair of IEEE VTC-Fall 2016. He has served on the Technical Committees of several international conferences in the fields of communications and signal processing. He has been an Associate Editor of the *EURASIP Journal on Applied Signal Processing*, from 2005 to 2007, the IEEE SIGNAL PROCESSING LETTERS, from 2006 to 2008, and the IEEE TRANSACTIONS ON SIGNAL PROCESSING, from 2010 to 2012, and a Guest Editor of two special issues of the *EURASIP Journal on Applied Signal Processing* published, in 2007 and 2014.