INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Adversarial Dictionary Learning for Monaural Speech Enhancement

Yunyun Ji[1], *Longting Xu*[2], *Wei-Ping Zhu*[3]

[1]Agora IO, Inc, Shanghai, China
[2]College of Information Science and Technology, Donghua University, Shanghai, China
[3]Concordia University, Montreal, Canada

jiyunyun@agora.io, xlt@dhu.edu.cn, weiping@ece.concordia.ca

Abstract

In this paper, we propose an adversarial dictionary learning method to train a speaker independent speech dictionary and a universal noise dictionary for improving the generality of the dictionary learning based speech enhancement system. In the learning stage, two discriminators are employed separately to identify the components in speech and noise which are highly correlated with each other. The residuals in the speech and noise magnitude spectral matrices are then utilized to train the speech and noise dictionaries via the alternating direction method of multiplier algorithm, which can effectively reduce the mutual coherence between speech and noise. In the enhancement stage, a new optimization technique is proposed for enhancing the speech based on the low-rank decomposition and sparse coding. Experimental results show that our proposed method achieves better performance in improving the speech quality and intelligibility than the reference methods in terms of three objective performance evaluation measures.

Index Terms: speech enhancement, dictionary learning, adversarial training, sparse coding, low rank matrix decomposition

1. Introduction

With a pervasive use of mobile devices and smart home products, speech enhancement has become increasingly important as a frontend of speech recognition, which can produce high-quality speech signals for the backend system under adverse environments. Conventional monaural speech enhancement methods such as the minimum mean-square error (MMSE) based log-spectral magnitude estimator (LogMMSE) [1] balance a tradeoff between reducing the residual noise and speech distortion especially under the low signal-to-noise ratio (SNR) environment due to the interaction between the sophisticated statistical models of speech and noise. Moreover, these conventional speech enhancement approaches suffer from obvious performance degradation when the speech is contaminated by the non-stationary noise under the real-world adverse environments.

The generative dictionary learning (GDL) [2] based speech enhancement method was proposed for noise reduction based on the sparse representation of the speech and noise magnitude spectra. Specifically, the K-singular value decomposition (K-SVD) algorithm [3] is employed in the learning stage to train the speech and noise dictionaries respectively based on the speech and noise training datasets in the time-frequency domain. At the enhancement stage, the noisy speech is able to be sparsely represented by a composite dictionary which is constructed by concatenating the speech and noise dictionaries to produce the sparse coefficient matrices of speech and noise. The speech magnitude spectral matrix can be estimated through the product of the speech dictionary and its corresponding sparse coefficient matrix. Finally, the time-domain speech signals are synthesized by applying inverse Fourier transform (IFT) to the estimated speech spectrum. The enhancement performance of the GDL method relies on the coherence of the dictionaries to their corresponding signal classes and their mutual coherence to the other signal classes. The high mutual coherence results in obvious distortion to the speech quality, named source confusion[2].

In this case, a modified coherence based dictionary learning (MCDL) method [4] was proposed to introduce a post-processing step, named atom correlation, to the trained dictionaries from the GDL method so as to reduce source confusion. However, both the GDL and the MCDL are based on the noise-dependent case, i.e., different noise dictionaries are trained for different types of noise in the learning stage and the corresponding dictionary is chosen from a bunch of noise dictionaries for the enhancement stage. On the one hand, dictionary selection via the voice activity detection requires a large memory and results in time delay. On the other hand, these methods are faced with sharp performance degradation when dealing with unseen noise types.

In order to solve these problems, we propose to train a speaker-independent speech dictionary and a universal noise dictionary by utilizing an adversarial training scheme. Specifically, we employ two discriminators to remove the components from the speech and noise magnitude spectral matrices which are highly correlated with each other. The residuals are utilized for training corresponding dictionaries. In the enhancement stage, we propose a new optimization technique for recovering the speech from the mixture signal. All the optimization problems involved in the proposed speech enhancement method are addressed via the alternating direction method of multiplier (ADMM) algorithm [5]. Compared with the LogMMSE, GDL and MCDL algorithms, the proposed method has the robustness and generalization capability to varying background noise.

The rest of the paper is organized as follows: Section 2 gives a brief introduction to the ADMM algorithm for sparse coding and dictionary learning, which are the tools for our proposed method. Section 3 presents the details of the proposed method. Section 4 demonstrates the experimental results of the proposed method with comparison to three reference methods. Section 5 concludes the whole paper.

2. Related Work

2.1. Sparse coding

The ADMM algorithm was proposed in [5] to solve the optimization problem of the following style,

$$\min_{\boldsymbol{x},\boldsymbol{y}} f(\boldsymbol{x}) + g(\boldsymbol{y}) \quad s.t.\ \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{y} = \boldsymbol{c} \tag{1}$$

with variables $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{y} \in \mathbb{R}^p$, where $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, $\boldsymbol{B} \in \mathbb{R}^{m\times p}$, $\boldsymbol{c} \in \mathbb{R}^n$ and both f and g are convex. The

http://dx.doi.org/10.21437/Interspeech.2020-2500

ADMM alternates among the following steps for estimating the variables,

$$\boldsymbol{x}^{i+1} = \underset{\boldsymbol{x}}{arg\,min}\ f(\boldsymbol{x}) + \frac{\rho}{2}\left\|\boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{y}^i - \boldsymbol{c} + \boldsymbol{u}^i\right\|_2^2, \quad (2)$$

$$\boldsymbol{y}^{i+1} = \underset{\boldsymbol{y}}{arg\,min}\ g(\boldsymbol{y}) + \frac{\rho}{2}\left\|\boldsymbol{A}\boldsymbol{x}^{i+1} + \boldsymbol{B}\boldsymbol{y} - \boldsymbol{c} + \boldsymbol{u}^i\right\|_2^2, \quad (3)$$

$$\boldsymbol{u}^{i+1} = \boldsymbol{u}^i + \boldsymbol{A}\boldsymbol{x}^{i+1} + \boldsymbol{B}\boldsymbol{y}^{i+1} - \boldsymbol{c}, \quad (4)$$

where $\|\cdot\|_2$ denote the ℓ_2 norm of a vector, referring to the square root of the sum of the squares of all the elements in a vector, and $\rho > 0$ is the augmented Lagrangian factor.

Sparse coding [6] refers to sparsely represent the signal vector $\boldsymbol{d}$ or the signal matrix $\boldsymbol{D}$ with respect to an overcomplete dictionary $\boldsymbol{\Psi}$. In this paper, we focus on the sparse coding of the data matrix. The sparse coefficient matrix $\boldsymbol{\Theta}$ of the data matrix $\boldsymbol{D}$ with respect to the dictionary $\boldsymbol{\Psi}$ is estimated through solving the following convex optimization problem,

$$\min_{\boldsymbol{\Theta}} \frac{1}{2}\|\boldsymbol{D} - \boldsymbol{\Psi}\boldsymbol{\Theta}\|_F^2 + \lambda\|\boldsymbol{\Theta}\|_1, \quad (5)$$

where $\|\cdot\|_F$ denotes the Frobenius norm of a matrix, referring to the square root of the sum of the squares of all the entries in a matrix, and $\|\cdot\|_1$ denotes the ℓ_1 norm, referring to the sum of the absolute values of all the elements in a matrix or a vector. The ADMM algorithm can be used to solve this optimization problem [7] and the solution is denoted as

$$\boldsymbol{\Theta} = \mathrm{SC} - \mathrm{ADMM}(\boldsymbol{D}, \boldsymbol{\Psi}). \quad (6)$$

2.2. Dictionary learning

Dictionary learning aims at utilizing a large-scale training dataset $\boldsymbol{D}$ to train an overcomplete dictionary $\boldsymbol{\Psi}$ so as to capture the internal structure of the data for sparse representation [8], which is usually based on the following optimization problem,

$$\min_{\boldsymbol{\Psi},\boldsymbol{\Theta}} \frac{1}{2}\|\boldsymbol{D} - \boldsymbol{\Psi}\boldsymbol{\Theta}\|_F^2 + \lambda\|\boldsymbol{\Theta}\|_1 \quad \text{s.t. } \|\boldsymbol{\psi}_j\|_2 = 1, \quad \forall j, \quad (7)$$

where $\boldsymbol{\psi}_j$ is the j^{th} atom, namely the j^{th} column vector, in the dictionary $\boldsymbol{\Psi}$. This optimization problem can be addressed by alternating between the following two subproblems, i.e., at the $(k+1)^{th}$ iteration,

$$\boldsymbol{\Theta}^{k+1} = \arg\min_{\boldsymbol{\Theta}} \frac{1}{2}\left\|\boldsymbol{D} - \boldsymbol{\Psi}^k\boldsymbol{\Theta}\right\|_F^2 + \lambda\|\boldsymbol{\Theta}\|_1, \quad (8)$$

$$\boldsymbol{\Psi}^{k+1} = \arg\min_{\boldsymbol{\Psi}} \frac{1}{2}\left\|\boldsymbol{D} - \boldsymbol{\Psi}\boldsymbol{\Theta}^{k+1}\right\|_F^2 \ s.t.\ \|\boldsymbol{\psi}_j\|_2 = 1,\ \forall j. \quad (9)$$

The subproblem in (8) is solved via the ADMM based sparse coding algorithm in 2.1 and then

$$\boldsymbol{\Theta}^{k+1} = \mathrm{SC} - \mathrm{ADMM}(\boldsymbol{D}, \boldsymbol{\Psi}^k). \quad (10)$$

The above ADMM algorithm for updating the dictionary in (9) is denoted as UD-ADMM [9] and hence

$$\boldsymbol{\Psi}^{k+1} = \mathrm{UD} - \mathrm{ADMM}(\boldsymbol{D}, \boldsymbol{\Theta}^{k+1}). \quad (11)$$

The ADMM based dictionary learning algorithm alternates between the SC-ADMM and UD-ADMM, and is then denoted as DL-ADMM, i.e., the trained dictionary $\boldsymbol{\Psi}$ is expressed as

$$\boldsymbol{\Psi} = \mathrm{DL} - \mathrm{ADMM}(\boldsymbol{D}). \quad (12)$$

3. Proposed Speech Enhancement Method

In this section, we will present the details of the proposed speech enhancement method. The proposed method includes two stages: the learning stage and the enhancement stage.

3.1. Adversarial dictionary learning for speech and noise

In this part, we propose an adversarial learning based on the following speech and noise models.

The speech magnitude spectral matrix $\boldsymbol{S}$ is modeled as

$$\boldsymbol{S} = \boldsymbol{L}_s + \boldsymbol{\Psi}_s\boldsymbol{\Theta}_s, \quad (13)$$

where $\boldsymbol{L}_s$ is rank deficient and denotes the component highly correlated with noise, $\boldsymbol{\Psi}_s$ represents the speech dictionary and $\boldsymbol{\Theta}_s$ is the corresponding sparse coefficient matrix. The noise magnitude spectral matrix $\boldsymbol{N}$ is modeled as

$$\boldsymbol{N} = \boldsymbol{L}_n + \boldsymbol{\Psi}_n\boldsymbol{\Theta}_n, \quad (14)$$

where the low-rank matrix $\boldsymbol{L}_n$ denotes the component highly similar to speech, $\boldsymbol{\Psi}_n = \begin{bmatrix}\boldsymbol{\psi}_{n,1} & \boldsymbol{\psi}_{n,2} & \boldsymbol{\psi}_{n,M}\end{bmatrix}$ represents the noise dictionary and $\boldsymbol{\Theta}_n$ is the corresponding sparse coefficient matrix. In order to train the speech dictionary which presents lower mutual coherence to noise, we should extract the matrix $\boldsymbol{L}_s$ from the speech magnitude spectral matrix and utilize the residual matrix denoted as $\boldsymbol{R}_s = \boldsymbol{S} - \boldsymbol{L}_s$ to train the speech dictionary. We propose to employ a discriminative scheme to capture the components of $\boldsymbol{S}$, exhibiting high coherence to the noise dictionary $\boldsymbol{\Psi}_n$, to constitute the matrix $\boldsymbol{L}_s$. The discriminator exploits the following optimization problem to pick out the atoms in the noise dictionary which are highly correlated with speech, i.e.,

$$\hat{\boldsymbol{\Theta}}_{sn} = \arg\min_{\boldsymbol{\Theta}_{sn}} \frac{1}{2}\|\boldsymbol{S} - \boldsymbol{\Psi}_n\boldsymbol{\Theta}_{sn}\|_F^2 + \lambda_1\|\boldsymbol{\Theta}_{sn}\|_1. \quad (15)$$

The parameter λ_1 is a regulation factor, and the above optimization problem can be solved through the SC-ADMM algorithm. The i^{th} row vector in the estimated matrix $\hat{\boldsymbol{\Theta}}_{sn}$ is denoted as $\hat{\boldsymbol{\theta}}_{sn,i}$ and its energy is able to reflect the mutual coherence between the speech magnitude spectral matrix $\boldsymbol{S}$ and the i^{th} atom $\boldsymbol{\psi}_{n,i}$ in the noise dictionary. The set Λ consists of the indices corresponding to the first I-largest energy row vectors in the matrix $\hat{\boldsymbol{\Theta}}_{sn}$. The matrix $\boldsymbol{L}_s$ is approximated as

$$\boldsymbol{L}_s = \sum_{i\in\Lambda} \boldsymbol{\psi}_{n,i}\hat{\boldsymbol{\theta}}_{sn,i}. \quad (16)$$

The residual matrix $\boldsymbol{R}_s$ is estimated as

$$\boldsymbol{R}_s = \boldsymbol{S} - \boldsymbol{L}_s = \boldsymbol{S} - \sum_{i\in\Lambda} \boldsymbol{\psi}_{n,i}\hat{\boldsymbol{\theta}}_{sn,i}. \quad (17)$$

Then we utilize this residual matrix to train the generative speech dictionary by solving the subsequent optimization problem through the DL-ADMM algorithm,

$$\min_{\boldsymbol{\Psi}_s,\boldsymbol{\Theta}_s} \frac{1}{2}\|\boldsymbol{R}_s - \boldsymbol{\Psi}_s\boldsymbol{\Theta}_s\|_F^2 + \lambda_2\|\boldsymbol{\Theta}_s\|_1, \quad (18)$$

where λ_2 is a regulation factor. In this case, the speech dictionary will present relatively low mutual coherence to the noise magnitude spectral matrix $\boldsymbol{N}$, resulting from the low mutual coherence between the matrix $\boldsymbol{R}_s$ and $\boldsymbol{N}$.

Subsequently, we apply a similar discriminative scheme to

the noise magnitude spectral matrix $\boldsymbol{N}$ to extract the matrix $\boldsymbol{L}_n$ which is highly correlated with the speech dictionary and then use the residual $\boldsymbol{R}_n = \boldsymbol{N} - \boldsymbol{L}_n$ to train the generative noise dictionary so as to reduce the mutual coherence between the speech and noise dictionary. The following optimization technique is employed to estimate the sparse coefficient matrix $\boldsymbol{\Theta}_{ns}$ of the noise magnitude spectral matrix $\boldsymbol{N}$ with respect to the speech dictionary $\boldsymbol{\Psi}_s = \begin{bmatrix}\boldsymbol{\psi}_{s,1} & \boldsymbol{\psi}_{s,2} & \cdots & \boldsymbol{\psi}_{s,M}\end{bmatrix}$,

$$\hat{\boldsymbol{\Theta}}_{ns} = \arg\min_{\boldsymbol{\Theta}_{ns}} \frac{1}{2}\|\boldsymbol{N} - \boldsymbol{\Psi}_s\boldsymbol{\Theta}_{ns}\|_F^2 + \lambda_3\|\boldsymbol{\Theta}_{ns}\|_1, \quad (19)$$

which is also solved via the SC-ADMM algorithm and λ_3 is the regulation factor. Then the low-rank matrix $\boldsymbol{L}_n$ can be approximated as

$$\boldsymbol{L}_n = \sum_{i\in\Omega} \boldsymbol{\psi}_{s,i}\hat{\boldsymbol{\theta}}_{ns,i}, \quad (20)$$

where $\hat{\boldsymbol{\theta}}_{ns,i}$ is the i^{th} row vector in $\hat{\boldsymbol{\Theta}}_{ns}$ and Ω is a set consisting of the indices of the atoms with respect to the first $I-$ largest energy row vectors in $\hat{\boldsymbol{\Theta}}_{ns}$. The residual of the noise magnitude spectral matrix is obtained as

$$\boldsymbol{R}_n = \boldsymbol{N} - \boldsymbol{L}_n = \boldsymbol{N} - \sum_{i\in\Omega} \boldsymbol{\psi}_{s,i}\hat{\boldsymbol{\theta}}_{ns,i}. \quad (21)$$

Then, the noise dictionary is estimated based on the following optimization problem,

$$\min_{\boldsymbol{\Psi}_n,\boldsymbol{\Theta}_n} \frac{1}{2}\|\boldsymbol{R}_n - \boldsymbol{\Psi}_n\boldsymbol{\Theta}_n\|_F^2 + \lambda_4\|\boldsymbol{\Theta}_n\|_1 \quad (22)$$

which is solved by the DL-ADMM algorithm and λ_4 is the regulation factor. We implement the above process iteratively to continuously reduce the mutual coherence between speech and noise dictionary and vise versa.

3.2. Enhancement stage

At the enhancement stage, the noisy speech is firstly transformed into time-frequency domain via short-time Fourier transform (STFT) and the noisy speech magnitude spectral matrix $\boldsymbol{X}$ can be written as

$$\boldsymbol{X} = \boldsymbol{S}+\boldsymbol{N} = \boldsymbol{L}_s+\boldsymbol{\Psi}_s\boldsymbol{\Theta}_s+\boldsymbol{L}_n+\boldsymbol{\Psi}_n\boldsymbol{\Theta}_n = \boldsymbol{L}+\boldsymbol{\Psi}\boldsymbol{\Theta}, \quad (23)$$

where $\boldsymbol{L} = \boldsymbol{L}_s + \boldsymbol{L}_n$, $\boldsymbol{\Psi} = \begin{bmatrix}\boldsymbol{\Psi}_s & \boldsymbol{\Psi}_n\end{bmatrix}$, and $\boldsymbol{\Theta} = \begin{bmatrix}\boldsymbol{\Theta}_s \\ \boldsymbol{\Theta}_n\end{bmatrix}$. In accordance with (16) and (20), we have

$$\mathrm{rank}(\boldsymbol{L}_s) \le I \quad \text{and} \quad \mathrm{rank}(\boldsymbol{L}_n) \le I.$$

Hence, the rank of the matrix $\boldsymbol{L}$ satisfies $\mathrm{rank}(\boldsymbol{L}) \le 2I$. Therefore, $\boldsymbol{L}$ is also a low-rank matrix. In this case, we propose the following optimization problem to estimate the low-rank matrix $\boldsymbol{L}$ and the sparse coefficient matrix $\boldsymbol{\Theta}$, i.e.,

$$\min_{\boldsymbol{L},\boldsymbol{\Theta}} \frac{1}{2}\|\boldsymbol{X}-\boldsymbol{L}-\boldsymbol{\Psi}\boldsymbol{\Theta}\|_F^2+\tau\|\boldsymbol{\Theta}\|_1 \;\; s.t.\,\mathrm{rank}(\boldsymbol{L}) \le 2I, \quad (24)$$

where τ is a regulation factor. This optimization problem can be solved through iteratively alternating between the two subsequent problems, namely, at the t^{th} iteration,

$$\boldsymbol{\Theta}^t = \arg\min_{\boldsymbol{\Theta}} \frac{1}{2}\left\|\boldsymbol{X} - \boldsymbol{L}^{t-1} - \boldsymbol{\Psi}\boldsymbol{\Theta}\right\|_F^2 + \tau\|\boldsymbol{\Theta}\|_1 \quad (25)$$

$$\boldsymbol{L}^t = \arg\min_{\mathrm{rank}(\boldsymbol{L})\le 2I} \frac{1}{2}\left\|\boldsymbol{X} - \boldsymbol{L} - \boldsymbol{\Psi}\boldsymbol{\Theta}^t\right\|_F^2. \quad (26)$$

The subproblem (25) can be solved via the SC-ADMM algorithm. The subproblem (26) can be solved through SVD, i.e.,

$$\boldsymbol{X} - \boldsymbol{\Psi}\boldsymbol{\Theta}^t = \boldsymbol{U}\boldsymbol{\Xi}\boldsymbol{V}^T, \quad (27)$$

where both $\boldsymbol{U}$ and $\boldsymbol{V}$ are unitary matrices and $\boldsymbol{\Xi}$ is a rectangular diagonal matrix. And then at the t^{th} iteration, the matrix $\boldsymbol{L}$ can be approximated as

$$\boldsymbol{L}^t = \sum_{i=1}^{2I} \xi_i\boldsymbol{u}_i\boldsymbol{v}_i^T, \quad (28)$$

where ξ_i is the i^{th} element on the principal diagonal of $\boldsymbol{\Xi}$, $\boldsymbol{u}_i$ and $\boldsymbol{v}_j$ are the i^{th} left and right singular vectors, respectively.

According to (23), the estimated low-rank matrix can be expressed as $\hat{\boldsymbol{L}} = \hat{\boldsymbol{L}}_s + \hat{\boldsymbol{L}}_n$. As mentioned above, the matrix $\hat{\boldsymbol{L}}_s$ represents the component in the speech magnitude spectral matrix which is highly coherent to noise while the matrix $\hat{\boldsymbol{L}}_n$ is the component in the noise magnitude spectral matrix highly coherent to speech. In this case, we can estimate the matrices $\hat{\boldsymbol{L}}_s$ and $\hat{\boldsymbol{L}}_n$ as

$$\hat{\boldsymbol{L}}_s = \eta\hat{\boldsymbol{L}} \;\text{ and }\; \hat{\boldsymbol{L}}_n = (1-\eta)\hat{\boldsymbol{L}}, \quad (29)$$

where η is a factor playing a tradeoff between the speech distortion and noise reduction. Then, the speech and noise magnitude spectral matrices can be estimated as

$$\hat{\boldsymbol{S}} = \hat{\boldsymbol{L}}_s + \boldsymbol{\Psi}_s\hat{\boldsymbol{\Theta}}_s, \text{and } \hat{\boldsymbol{N}} = \hat{\boldsymbol{L}}_n + \boldsymbol{\Psi}_n\hat{\boldsymbol{\Theta}}_n. \quad (30)$$

The Wiener filter is employed to further improve the estimation of the speech magnitude spectrum which is utilized in combination with the phase of the complex noisy speech spectrum to synthesize the speech in time domain.

4. Experimental Results

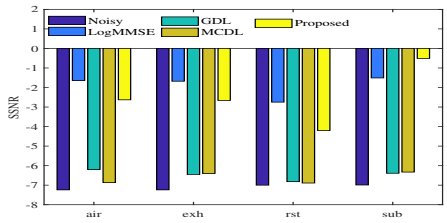
In this part, the experiments are implemented to evaluate the enhancement performance of the proposed method. We randomly selected 10 speakers and 10 utterances of each speaker from the TIMIT speech corpus, to constitute the training dataset of clean speech. Four types of noise including babble (bab), hfchannel (hf), white (wht) and factory (fct), from the NOISEX-92 database are utilized for training a universal noise dictionary. Another 5 speakers from the TIMIT corpus who are distinct from the speakers in the learning stage and 5 utterances of each speaker are used at the enhancement stage. As the speakers in the enhancement stage are different from those in the learning stage, it is the speaker independent case.

We consider the quality and intelligibility of the enhanced speech under the unseen noise environment. In this scenario, four types of noise, namely subway (sub), restaurant (rst), airport (air) and exhibition (exh), from the Aurora2 database are used in the enhancement stage to build the test dataset with the clean speech at four SNR levels, namely, -5dB, 0dB, 5dB and 10dB. Here, the noise types are distinct from the noise types in the learning stage, which can verify the generality of the proposed method. Three objective measures are utilized in this paper. Perceptual evaluation of speech quality (PESQ) is highly related to the subjective quality rating scores and suitable for evaluating the overall speech quality [10]. Segmental SNR (SSNR) is utilized to measure the level of noise distortion [10]. Short-time objective intelligibility (STOI) score is utilized to measure the speech intelligibility [11].

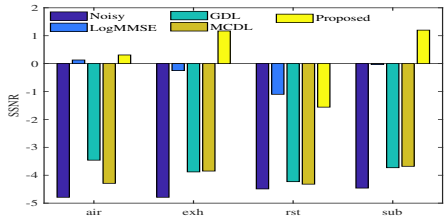
All the clean speech and noise signals are down sampled to 8KHz. The Hamming window with the length of 512 is employed for framing with the overlap of 50%.

Table 1, Table 2 and Figure 1 present the PESQ, STOI and SSNR results of various speech enhancement methods under

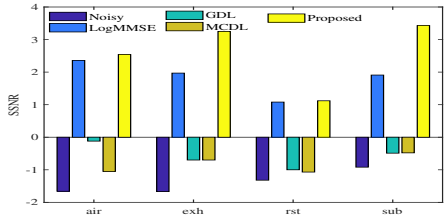
unseen noise environments, respectively. Generally speaking, the proposed method can achieve much better performance in enhancing the speech quality and intelligibility than the reference methods in this scenario. In accordance with Table 1, the average PESQ improvement of the traditional LogMMSE is trivial as result of the high non-stationarity of the noise. Due to the mismatch of the noise types in the learning and enhancement stages, neither of the GDL and the MCDL algorithms can effectively improve the PESQ scores as the average improvement is around 0.05. However, by using the adversarial learning, the gain of our proposed method in the average PESQ score is around 0.35. In Table 2, none of the reference methods can improve the average STOI score, while the proposed technique can still guarantee the improvement. In Figure 1, the weak capacity of the GDL and the MCDL algorithms in improving the SSNR is recognized. And the improvement of the proposed method in the average SSNR is about 0.4dB larger than that of the LogMMSE. Thus, our proposed method can better improve the overall quality and intelligibility of the speech than the reference methods under unseen noise environments.



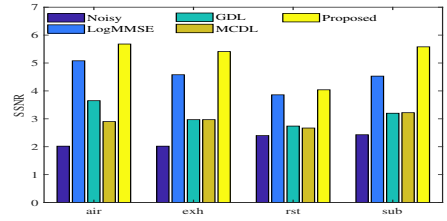
(a)



(b)



(c)



(d)

Figure 1: *Average SSNR results of different speech enhancement methods for unseen noise types. (a) SNR=-5dB, (b) SNR=0dB, (c) SNR=5dB, (d) SNR=10dB.*

Table 1: *PESQ scores of different speech enhancement methods for four unseen noise types at four Levels of SNR.*

		Noisy	LogMMSE	GDL	MCDL	Proposed
-5dB	sub	1.47	1.51	1.52	1.55	1.80
	air	1.76	1.67	1.78	1.77	1.97
	exh	1.53	1.34	1.59	1.59	1.78
	rst	1.72	1.53	1.73	1.72	1.88
	Ave	1.62	1.51	1.66	1.66	1.86
0dB	sub	1.68	1.90	1.75	1.77	2.17
	air	2.02	2.07	2.06	2.02	2.28
	exh	1.67	1.78	1.76	1.79	2.14
	rst	1.95	1.89	1.96	1.96	2.13
	Ave	1.83	1.91	1.88	1.89	2.18
5dB	sub	1.96	2.26	2.05	2.06	2.50
	air	2.31	2.45	2.36	2.30	2.58
	exh	1.94	2.20	2.03	2.06	2.48
	rst	2.23	2.24	2.25	2.24	2.41
	Ave	2.11	2.29	2.17	2.17	2.49
10dB	sub	2.28	2.64	2.35	2.36	2.79
	air	2.61	2.82	2.65	2.60	2.88
	exh	2.25	2.56	2.33	2.36	2.78
	rst	2.53	2.61	2.54	2.53	2.71
	Ave	2.42	2.66	2.47	2.46	2.79

Table 2: *STOI scores under different types of unseen noise at four different levels of SNRs.*

		Noisy	LogMMSE	GDL	MCDL	Proposed
-5dB	sub	0.51	0.47	0.51	0.53	0.57
	air	0.60	0.56	0.59	0.58	0.60
	exh	0.53	0.49	0.54	0.56	0.57
	rst	0.55	0.47	0.55	0.55	0.55
	Ave	0.55	0.50	0.55	0.56	0.57
0dB	sub	0.64	0.61	0.65	0.66	0.72
	air	0.72	0.68	0.71	0.70	0.75
	exh	0.68	0.62	0.69	0.70	0.74
	rst	0.68	0.61	0.68	0.68	0.69
	Ave	0.68	0.63	0.68	0.69	0.73
5dB	sub	0.78	0.74	0.78	0.79	0.83
	air	0.82	0.78	0.82	0.80	0.85
	exh	0.81	0.76	0.81	0.82	0.84
	rst	0.80	0.75	0.80	0.80	0.81
	Ave	0.80	0.76	0.80	0.80	0.83
10dB	sub	0.89	0.85	0.89	0.89	0.90
	air	0.90	0.87	0.90	0.89	0.92
	exh	0.90	0.87	0.90	0.90	0.91
	rst	0.90	0.85	0.89	0.89	0.90
	Ave	0.90	0.86	0.90	0.89	0.91

5. Conclusions

In contrast to the existing dictionary learning based speech enhancement method, an adversarial training scheme is employed in the learning stage to train a speaker-independent speech dictionary and a universal noise dictionary. In detail, two discriminators based on adversarial sparse coding are utilized to remove the components from speech and noise magnitude spectral matrices, reducing the mutual coherence between speech and noise dictionaries. Then, the well-trained dictionaries are applied in the enhancement stage to effectively separate speech and noise based on a new optimization technique which is solved through the ADMM based sparse coding and low-rank matrix decomposition. Experimental results demonstrate that our proposed method effectively enhances the speech in terms of the PESQ, SSNR and STOI scores.

6. References

[1] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE signal processing letters*, vol. 9, no. 1, pp. 12–15, 2002.

[2] C. D. Sigg, T. Dikk, and J. M. Buhmann, "Speech enhancement using generative dictionary learning," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1698–1712, 2012.

[3] M. Aharon, M. Elad, A. Bruckstein *et al.*, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on signal processing*, vol. 54, no. 11, p. 4311, 2006.

[4] S. Mavaddaty, S. M. Ahadi, and S. Seyedin, "Modified coherence-based dictionary learning method for speech enhancement," *IET Signal Processing*, vol. 9, no. 7, pp. 537–545, 2015.

[5] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein *et al.*, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine learning*, vol. 3, no. 1, pp. 1–122, 2011.

[6] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM review*, vol. 43, no. 1, pp. 129–159, 2001.

[7] Q. Lyu, D. Ruan, J. Hoffman, R. Neph, M. McNitt-Gray, and K. Sheng, "Iterative reconstruction for low dose ct using plug-and-play alternating direction method of multipliers (admm) framework," in *Medical Imaging 2019: Image Processing*, vol. 10949. International Society for Optics and Photonics, 2019, p. 1094906.

[8] I. Tosic and P. Frossard, "Dictionary learning," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 27–38, 2011.

[9] Y. Sun, T. Fei, L. Zhang, X. Liu, and J. Zhang, "Improving medical ct image blind restoration algorithm based on dictionary learning by alternating direction method of multipliers," *Automatic Control and Computer Sciences*, vol. 52, no. 1, pp. 49–59, 2018.

[10] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on audio, speech, and language processing*, vol. 16, no. 1, pp. 229–238, 2008.

[11] M. Kolbæk, Z.-H. Tan, and J. Jensen, "On the relationship between short-time objective intelligibility and short-time spectral-amplitude mean-square error for speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 2, pp. 283–295, 2019.