EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

Single-channel speech enhancement based on joint constrained dictionary learning

Linhui Sun*, Yunyi Bu, Pingan Li and Zihao Wu

Abstract

To improve the performance of speech enhancement in a complex noise environment, a joint constrained dictionary learning method for single-channel speech enhancement is proposed, which solves the "cross projection" problem of signals in the joint dictionary. In the method, the new optimization function not only constrains the sparse representation of the noisy signal in the joint dictionary, and controls the projection error of the speech signal and noise signal on the corresponding sub-dictionary, but also minimizes the cross projection error and the correlation between the sub-dictionaries. In addition, the adjustment factors are introduced to balance the weight of constraint terms to obtain the joint dictionary more discriminatively. When the method is applied to the single-channel speech enhancement, speech components of the noisy signal can be more projected onto the clean speech sub-dictionary of the joint dictionary without being affected by the noise sub-dictionary, which makes the quality and intelligibility of the enhanced speech higher. The experimental results verify that our algorithm has better performance than the speech enhancement algorithm based on discriminative dictionary learning under white noise and colored noise environments in time domain waveform, spectrogram, global signal-to-noise ratio, subjective evaluation of speech quality, and logarithmic spectrum distance.

Keywords: Single-channel speech enhancement, Joint constraint, Sparse representation, Dictionary learning, Optimization function

1 Introduction

Speech is inevitably affected by the surrounding environment in real life. The background noise, such as mechanical sound, traffic horn, and human voice, seriously affects the intelligibility and clarity of speech signals. Speech enhancement is to extract pure speech signal from the noisy speech signal as far as possible and restrict background noise. It has been widely used in mobile communication, smart home, speech coding, military equipment, and other practical application scenarios [1–5].

Unsupervised speech enhancement algorithms based on short-time spectrum estimation include spectral subtraction [6], statistical model-based method [7], subspace-based method [8], etc. These methods can suppress the stationary noise significantly, but for the non-stationary noise, they often cannot get a good noise reduction effect. In recent years, supervised speech enhancement such as deep neural network and sparse dictionary learning, which uses a pre-trained model to obtain the prior information of the source signal, can get a better denoising effect for non-stationary noise. Deep neural network-based speech enhancement mainly realizes the mapping from noisy signal to clean signal by learning the parameters of a multi-layer network from a large number of sample data. Dictionary learning-based speech enhancement mainly uses some signals to learn dictionaries and get the sparse representation of a clean signal. This paper focuses on the research of the enhancement algorithm based on sparse dictionary learning.

Speech enhancement methods based on sparse dictionary learning have been a research hotspot for several years [9–11]. The sparse dictionary representation model

* Correspondence: sunlh@njupt.edu.cn
College of Telecommunications & Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing, China

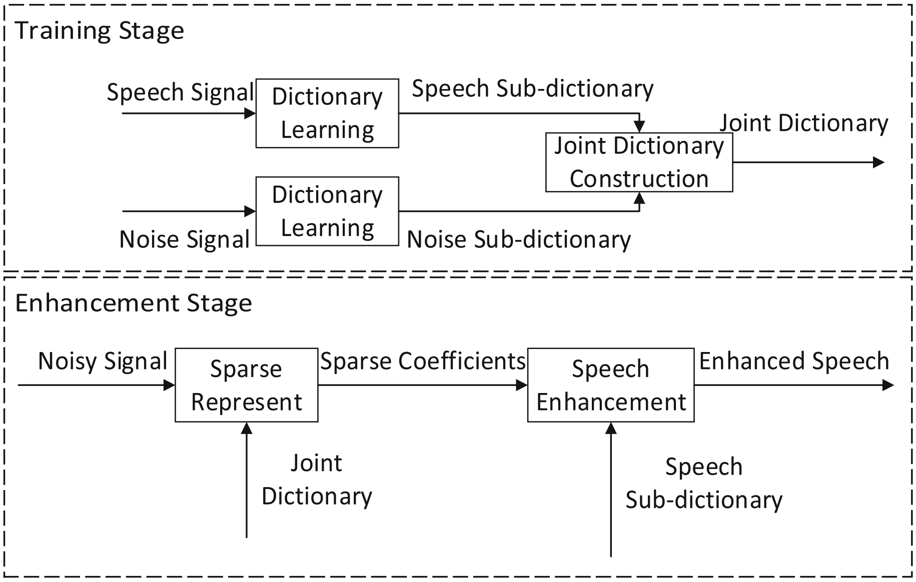


Fig. 1 Speech enhancement based on joint dictionary

assumes that speech signals can be described as a linear combination of several atoms derived from a dictionary matrix. Sparse dictionaries are generally divided into two categories. One is trained with data, such as the K-singular value decomposition dictionary [12]. The other is constructed with a fixed basis [13], such as the discrete cosine transformation dictionary. Many scholars have obtained abundant achievements in speech separation [14] and speech enhancement [15]. Sigg et al. projected the noisy speech on the joint dictionary held together by the clean speech dictionary and the noise dictionary [16]. The speech components and the noise components in the noisy signal were represented by the corresponding sub-dictionaries respectively. Thus, the pure speech signal could be estimated. Mohammadiha et al. proposed a speech enhancement method based on Bayesian non-negative matrix factorization, which combined non-negative matrix factorization with dictionary learning and update methods to adapt to the change of SNR [17]. Baby et al. proposed a speech enhancement and automatic speech recognition algorithm based on double dictionaries [18]. The method used speech signals as dictionary atoms to train the discrete Fourier transform (DFT) domain dictionary and the corresponding feature domain dictionary. In the reconstruction stage, when the noisy signal is projected on the feature domain dictionary, it has less operation dimension and better enhancement effect. Sprechmann et al. proposed a learnable low-rank coefficient model for speech enhancement, which constrained the reconstructed speech and noise to be low rank [19]. With the introduction of neural network technology, the sparse representation coefficients can be obtained more accurately. Whereas the distinction between speech dictionary and noise dictionary is not good enough, some residual noise and distortion existed in the enhanced speech. Zhang et al. trained the dictionary by constraining the relationship among speech, noise, and noisy signal as well as the cross-interference between corresponding dictionaries, thereby improving the discriminability of the joint dictionary [20]. However, due to the fact that some speech components of the noisy signal were still projected on the interference noise sub-dictionary, the speech enhancement performance is not optimal. Fu et al. proposed a two-level complementary joint sparse representation method to enhance single-channel speech [21]. To suppress noise source confusion, a two-level joint sparse representation was constructed using the relationship among speech, noise, noisy signals, and the discriminative property of joint dictionary to estimate a less distorted speech signal. Jia et al. proposed a speech enhancement algorithm with alternate optimization of sparse coefficient and dictionary [22]. The objective function of dictionary learning was constrained by the Fisher criterion, and then the discriminative dictionary and the corresponding sparse coefficient are obtained. In this way, the cross interference between joint dictionaries can be reduced.

To further effectively inhibit the cross projection between the sub-dictionaries of the joint dictionary, a new optimization function is presented in this paper. The optimization function not only jointly controls the reconstruction error of signals and dictionaries but also constrains the cross projection and the correlation between the sub-dictionaries. Furthermore, the adjustment factors are introduced to balance the weight of constraint items, which makes the joint dictionary more discriminative. Thus, the clean speech components would be

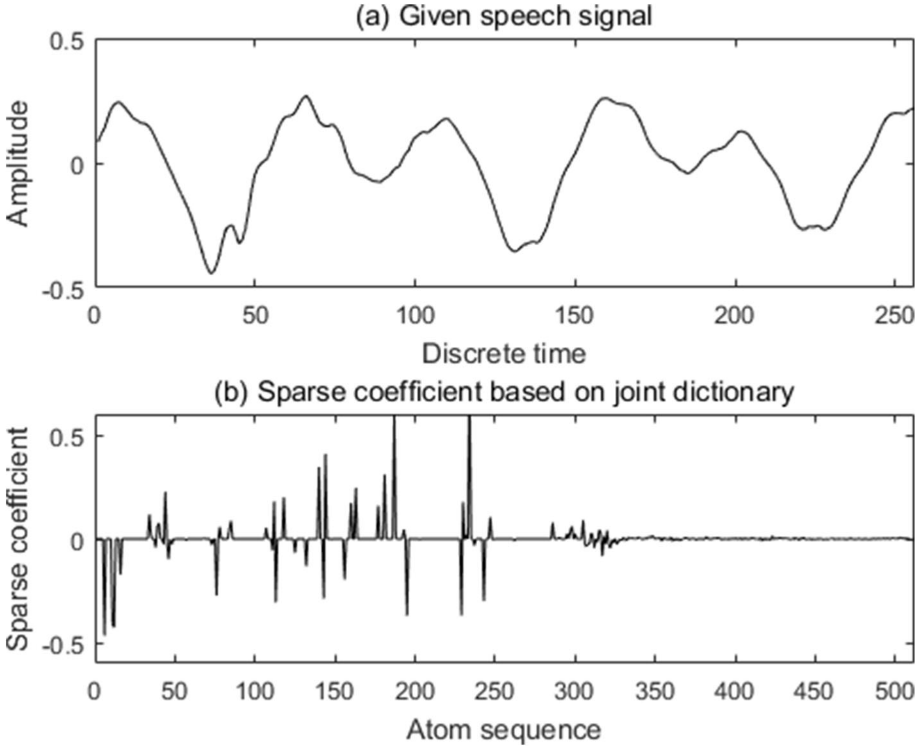


Fig. 2 Projection of a given frame signal onto the joint dictionary

more projected onto the clean speech sub-dictionary of the joint dictionary with being affected as little as possible by the noise sub-dictionary, which makes the enhanced speech quality and intelligibility higher.

The remainder of the paper is organized as follows. Section 2 introduces the joint dictionary learning and the "cross projection" problem. Section 3 mainly elaborates the proposed method. The experiments and results analysis are presented in Section 4. Finally, the conclusion of this work is presented in Section 5.

2 Speech enhancement based on joint dictionary learning

2.1 Algorithm overview

Speech enhancement is to extract speech signals as pure as possible from noisy speech signals. The single-channel speech enhancement model is defined as follows:

$$y(k) = x(k) + n(k), 1 \leq k \leq T, \tag{1}$$

where k is the discrete time sequence number. $y(k)$,$x(k)$, and $n(k)$ represent the discrete time signals of noisy speech, clean speech, and interference noise signal respectively. We aim to reduce the interference of noisy signals and extract as pure speech signals as possible from noisy signals.

For clarity, we define$\mathrm{Y}$as the training set of noisy speech signals, Xas the training set of clean speech signals, and$\mathrm{N}$ as the training set of noise signals in the time domain. The relationship between$\mathrm{Y}$,X, and$\mathrm{N}$can be written as

$$\mathrm{Y} = \mathrm{X} + \mathrm{N}. \tag{2}$$

Single-channel speech enhancement based on joint dictionary learning is divided into two stages: training stage and enhancement stage. In the training stage, the sub-dictionaries corresponding to clean speech signal and noise signal are trained, respectively, and then the two sub-dictionaries are spliced into a joint dictionary. In the enhancement stage, the noisy signal is projected onto the joint dictionary to recover the enhanced speech signal. The specific process is shown in Fig. 1.

In the training stage, the speech sub-dictionary$\mathbf{D}_x$and noise sub-dictionary$\mathbf{D}_n$are usually learned from clean speech signals and noise signals of the training set by the K-SVD algorithm. Then, the two sub-dictionaries are combined into a joint dictionary$\mathbf{D} = [\mathbf{D}_x, \mathbf{D}_n]$. The objective function can be expressed as

$$\mathbf{D} = \arg \min_{\mathbf{D}} \left\| \mathrm{Y} - \mathbf{DC} \right\|_F^2. \tag{3}$$

In the enhancement stage, the sparse coding matrix of the noisy signal training set$\mathrm{Y}$on the joint dictionary$\mathbf{D} =$

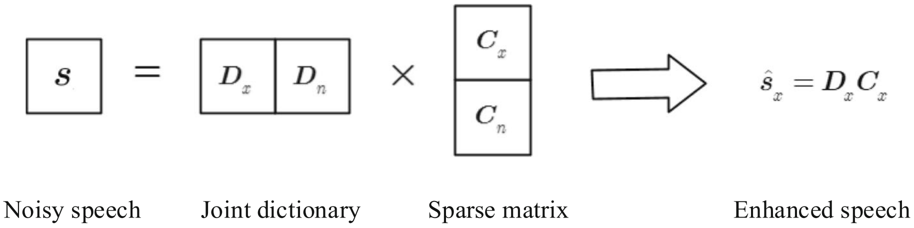


Fig. 3 Schematic diagram of speech enhancement

$[\mathbf{D}_x, \mathbf{D}_n]$ is $\mathbf{C} = [(\mathbf{C}^x)^{\mathrm{T}}, (\mathbf{C}^n)^{\mathrm{T}}]^{\mathrm{T}}$, where $\mathbf{C}^x$and $\mathbf{C}^n$represent the sparse coding coefficients of Y on the signal sub-dictionary $\mathbf{D}_x$and noise sub-dictionary $\mathbf{D}_n$respectively, Tdenotes matrix transposition. The representation of the noisy speech signal **s** on the joint dictionary**D**is

$$\mathbf{S}{=}\mathbf{D} \times \mathbf{C}{=}[\mathbf{D}_x, \mathbf{D}_n] \times \begin{bmatrix} \mathbf{C}_x \\ \mathbf{C}_n \end{bmatrix}. \tag{4}$$

After the sparse coefficient matrix**C**of the noisy signal in the joint dictionary is obtained by sparse coding algorithm, we can reconstruct the desired target source signal $\hat{\mathbf{s}}_x$according to

$$\hat{\mathbf{S}}_x = \mathbf{D}_x \mathbf{C}_x. \tag{5}$$

2.2 "Cross projection" problem

The traditional speech enhancement algorithm based on joint dictionary learning usually only considers the characteristics of the given signal and does not consider the similarity between sub-dictionary. Therefore, some speech components in the noisy signal will be projected on the interference noise sub-dictionary, resulting in "cross projection" problem, which leads to source confusion and poor enhancement effect.

Table 1 Detailed process of the proposed speech enhancement algorithm

In the training stage
Input: Speech signal, noise signal, and noisy speech of training set Output: Trained joint dictionary Step 1: Divide the training signals into frames by a rectangular window. Step 2: Use the K-SVD algorithm to obtain the sub-dictionaries corresponding to the speech signal and noise signal of the training set, and then concatenate them to get the initial joint dictionary. Step 3: Use the BP algorithm to calculate the sparse coefficient matrix of noisy speech on the joint dictionary. Step 4: Use the L-BFGS algorithm to solve the proposed optimization function and update the joint dictionary.
In the enhancement stage
Input: Noisy speech signal of the testing set Output: Reconstructed speech signal Step 1: Preprocess the input signal by framing. Step 2: Use the BP algorithm to calculate the sparse coefficients of noisy speech on the joint dictionary. Step 3: Use the speech sub-dictionary in the joint dictionary and the corresponding sparse coefficients to recover the frame-level speech signals. Step 4: Connect all the frame-level signals to reconstruct speech signals.

The sparse coefficients of a frame of a clean speech signal on the joint dictionary constructed by the method in Section 2.1 are shown in Fig. 2. Figure 2 a represents the time-domain waveform of a frame of clean speech signal, and Fig. 2 b represents the sparse coefficient representation of clean speech signal in the joint dictionary. The abscissa of Fig. 2b represents the sequence numbers of 512 atoms in the joint dictionary. The former 256 represents speech signal atoms, and the latter 256 represents noise signal atoms. From the figure, we can see that there are some coefficients of speech signals on the noise sub-dictionary, which has a bad impact on the reconstructed speech signal. Therefore, it is necessary to further strengthen the distinction between sub-dictionaries with the constraint of the joint dictionary, which helps to reduce the occurrence of "cross interference" problem. Thus, speech components can be projected onto the speech dictionary as much as possible to reconstruct speech signals better.

3 Speech enhancement based on joint constrained dictionary learning

The traditional single-channel speech enhancement algorithm based on joint dictionary learning is easy to cause mutual interference due to the lack of differentiation between sub-dictionaries, which leads to source confusion in the enhancement stage. In order to train a better joint constraint dictionary for speech enhancement, we propose a new optimization function with the joint constraint relationship between the speech sub-dictionary and the noise sub-dictionary.

3.1 New optimization function

The traditional construction method of joint dictionary for speech enhancement uses the source signals training sets to train the corresponding sub-dictionaries, and then combines them to construct the joint dictionary. This method only takes advantage of the characteristics of the source signal itself, but does not consider the similarity and interference between the source signals. When the noisy speech signals are represented on the joint dictionary, the noise has great interference on the speech components.

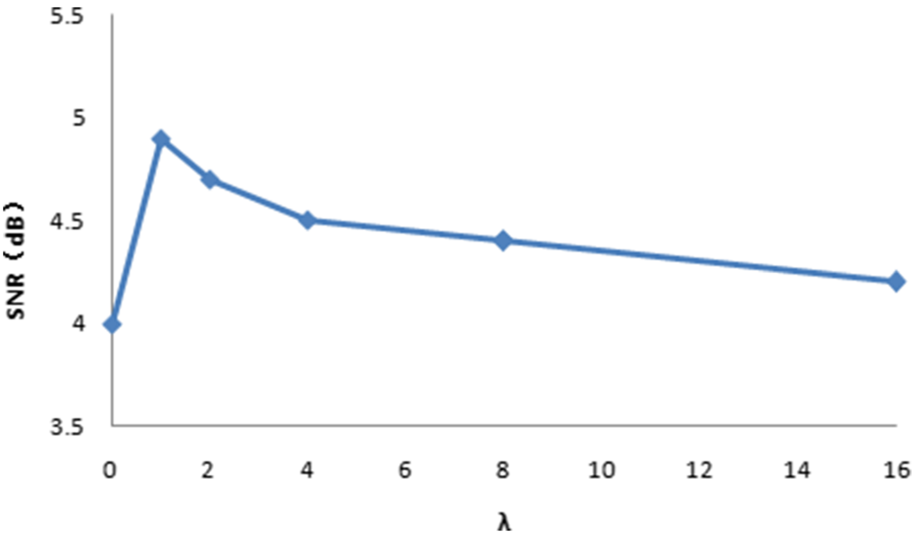


Fig. 4 Influence of balance factor λ on performance

1In this paper, the characteristics of noisy signals, cross interference, and dictionary correlation are fully considered. The discriminative joint dictionary is trained with the discriminative constraint optimization function. The objective function of dictionary learning can be written as

$$\mathbf{D}=\arg\min_{\mathbf{D}} \|\mathrm{Y}-\mathbf{D}\mathbf{C}\|_F^2+\|\mathrm{X}-\mathbf{D}_x\mathbf{C}^x\|_F^2+\|\mathrm{N}-\mathbf{D}_n\mathbf{C}^n\|_F^2 +\alpha\|\mathbf{D}_x\mathbf{C}^n\|_F^2+\beta\|\mathbf{D}_n\mathbf{C}^x\|_F^2+\lambda\left\|\mathbf{D}_x^{\mathrm{T}}\mathbf{D}_n\right\|_F^2, \tag{6}$$

where$\mathbf{D}_x$and$\mathbf{D}_n$are the sub-dictionaries corresponding to speech and noise signals respectively, and$\mathbf{D}$is a joint dictionary composed of$\mathbf{D}_x$and$\mathbf{D}_n$, denoted as $\mathbf{D}=[\mathbf{D}_x, \mathbf{D}_n]$. The coding matrix of the noisy training set Y on the joint dictionary $\mathbf{D}$ is $\mathbf{C}=[(\mathbf{C}_x)^{\mathrm{T}},(\mathbf{C}_n)^{\mathrm{T}}]^{\mathrm{T}}$, where $\mathbf{C}_x$and $\mathbf{C}_n$denote the sparse coding coefficients of the noisy signal Y over the signal sub-dictionary$\mathbf{D}_x$and noise sub-dictionary$\mathbf{D}_n$respectively. α, β, and λ are the adjustment factors of balancing constraint weight.

The first term in (6) is to control the sparse representation of the noisy signal in the joint dictionary, so that the reconstructed signal is close to the source signal. The second and third terms restrict the approximate errors when the speech signal and the noise signal are projected on the corresponding speech sub-dictionary and noise sub-dictionary, respectively, which makes the speech component in the noisy signal be projected on

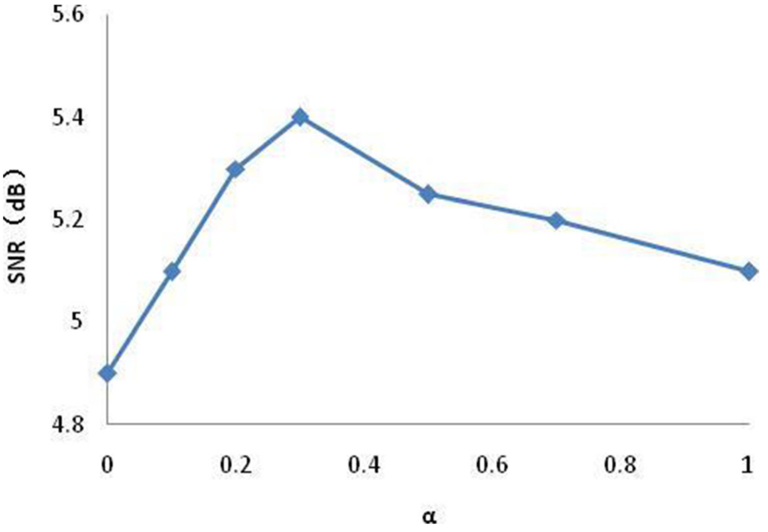


Fig. 5 Influence of balance factor α on performance

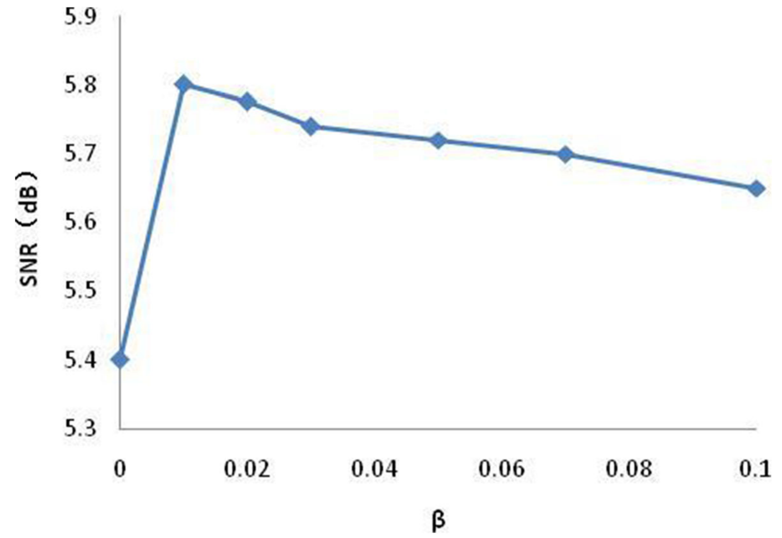


Fig. 6 Influence of balance factor β on performance

the speech sub-dictionary as far as possible. To reduce the cross interference between sub-dictionaries, the fourth and fifth terms restrict the cross projection of the speech component on the noise sub-dictionary and the noise component on the speech sub-dictionary respectively. When the speech components in the noisy signal are sparsely represented on the joint dictionary, the representation of the speech components on the noisy sub-dictionary is controlled as small as possible. In addition, it can be seen from reference [23] that the greater the correlation between the signal and the noise sub-dictionaries is, the more the projection of speech signal on the noise dictionary, which is likely to cause the "cross projection" problem. Therefore, we control the correlation between the speech and noise sub-dictionaries by constraining their inner product in the last term to make the sub-dictionaries more discriminative. In the last three terms of (6), adjustment factors are introduced to balance the weights of constraint terms, which makes the objective function more stable and the signal representation more sparse. The proposed method in this paper not only considers the reconstruction error of the noisy signal and source signal but also considers the cross interference and the correlation between the sub-dictionaries. The joint dictionary is trained by constraining the objective function, so that the speech components in the noisy signal can be sparsely represented in the speech sub-dictionary without being interfered by the noise sub-dictionary. It should be noted that different from [20], we further reduce the cross projection by constraining the cross projection term and introducing adjustment factors into the objective function.

3.2 Joint constrained dictionary learning

Learning the joint constraint dictionary with the new optimization objective function mainly includes the following three stages:

(1) Initialize the joint dictionary: use K-SVD algorithm to train speech sub-dictionary$\mathbf{D}_x$and noise sub-dictionary$\mathbf{D}_n$from the training signals X and N. The two sub-dictionaries are concatenated into the initial joint dictionary$\mathbf{D} = [\mathbf{D}_x, \mathbf{D}_n]$.

(2) Sparse coding update: when the initial joint dictionary$\mathbf{D}$is fixed, we use the objective function (7) to obtain the sparse coefficient matrix$\mathbf{C}$

$$\min \|\mathbf{C}\|_1 s.t. \mathrm{Y} = \mathbf{DC}. \tag{7}$$

Table 2 Increment of SNR in the white noise environment with different frame lengths (dB)

SNR (dB)	Frame length = 128	Frame length = 256	Frame length = 512
− 10	6.7389	7.8283	7.0193
− 5	4.6526	5.8115	5.2284
0	2.8977	4.0693	3.5235
5	1.9062	2.3522	2.2963
10	0.2983	1.0279	0.5137

Table 3 Increment of SNR in the white noise environment with different dictionary redundancy (dB)

SNR (dB)	Redundancy = 1	Redundancy = 2	Redundancy = 3
– 10	7.8283	9.6935	10.0157
– 5	5.8115	7.0714	7.8376
0	4.0693	6.0125	6.8422
5	2.3522	4.6939	5.4729
10	1.0279	2.9905	3.2104

BP algorithm [24] is selected to obtain the sparse representation in this paper.

(3) Dictionary update: when the sparse coefficient matrix $\mathbf{C}$is fixed, the dictionary is updated by the optimization function (6) to obtain the discriminative joint dictionary. To jointly optimize each sub-dictionary, we introduce the matrix$\mathbf{P}_1 = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$, $\mathbf{P}_2 = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}$, $\mathbf{P}_3 = \begin{pmatrix} \mathbf{0} & \mathbf{I} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$, $\mathbf{P}_4 = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{I} & \mathbf{0} \end{pmatrix}$, $\mathbf{P}_5 = \begin{pmatrix} \mathbf{I} \\ \mathbf{0} \end{pmatrix}$ and $\mathbf{P}_6 = \begin{pmatrix} \mathbf{0} \\ \mathbf{I} \end{pmatrix}$, where $\mathbf{0}$ denotes a zero matrix and $\mathbf{I}$ represents an identity matrix. Hence, (6) can be written as

$$\begin{aligned} \mathrm{Q} = {} & \|\mathrm{Y}-\mathbf{DC}\|_F^2 + \|\mathrm{X}-\mathbf{DP}_1\mathbf{C}\|_F^2 \\ & + \|\mathrm{N}-\mathbf{DP}_2\mathbf{C}\|_F^2 + \alpha\|\mathbf{DP}_3\mathbf{C}\|_F^2 \\ & + \beta\|\mathbf{DP}_4\mathbf{C}\|_F^2 + \lambda\|(\mathbf{DP}_5)^{\mathrm{T}}\mathbf{DP}_6\|_F^2. \end{aligned} \tag{8}$$

The problem of (8) can be solved by the limited-memory BFGS algorithm (L-BFGS) [25]. The gradient function of the objective function is

$$\begin{aligned} \frac{\partial \mathrm{Q}}{\partial \mathbf{D}} = {} & 2 \times [(\mathbf{DCC}^{\mathrm{T}}-\mathrm{Y}\mathbf{C}) \\ & + (\mathbf{DP}_1\mathbf{CC}^{\mathrm{T}}\mathbf{P}_1^{\mathrm{T}} + \mathbf{DP}_2\mathbf{CC}^{\mathrm{T}}\mathbf{P}_2^{\mathrm{T}}) \\ & -(\mathrm{X}\mathbf{CP}_1^{\mathrm{T}} + \mathrm{N}\mathbf{CP}_2^{\mathrm{T}}) \\ & + \alpha(\mathbf{DP}_3\mathbf{CC}^{\mathrm{T}}\mathbf{P}_3^{\mathrm{T}}) + \beta(\mathbf{DP}_4\mathbf{CC}^{\mathrm{T}}\mathbf{P}_4^{\mathrm{T}}) \\ & + \lambda(\mathbf{DP}_5\mathbf{P}_5^{\mathrm{T}}\mathbf{D}^{\mathrm{T}}\mathbf{DP}_6\mathbf{P}_6^{\mathrm{T}} + \mathbf{DP}_6\mathbf{P}_6^{\mathrm{T}}\mathbf{D}^{\mathrm{T}}\mathbf{DP}_5\mathbf{P}_5^{\mathrm{T}})]. \end{aligned} \tag{9}$$

Finally, we use the L-BFGS algorithm to solve the optimization function and obtain the discriminative joint dictionary$\mathbf{D}$.

3.3 Speech enhancement

Once we have obtained the discriminative joint dictionary, we can refer to the speech enhancement method in Section 2.1. We use the BP sparse coding algorithm to calculate the coefficient matrix $\mathbf{C} = [(\mathbf{C}_x)^{\mathrm{T}}, (\mathbf{C}_n)^{\mathrm{T}}]^{\mathrm{T}}$ of the noisy signal sin the joint dictionary and estimate the clean target speech signal $\hat{s}_x$. The module can be denoted as Fig. 3. According to $\mathbf{D}_x$and $\mathbf{C}_x$, we recover the estimated clean speech signal by

$$\hat{\mathbf{s}}_x = \mathbf{D}_x\mathbf{C}_x. \tag{10}$$

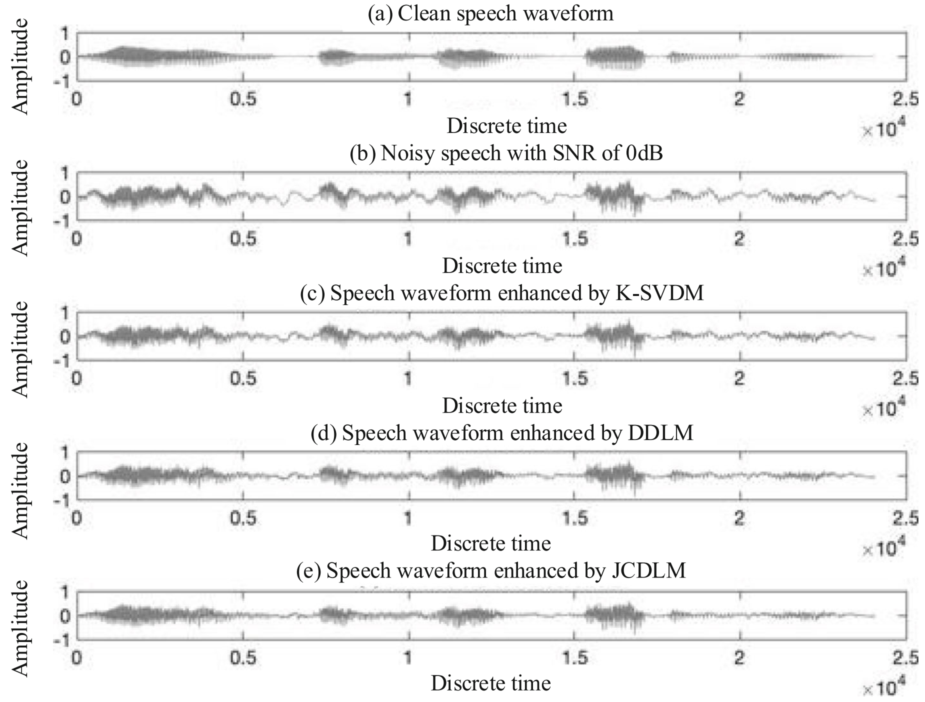


Fig. 7 Enhanced waveforms of different algorithms under 0-dB vehicle noise

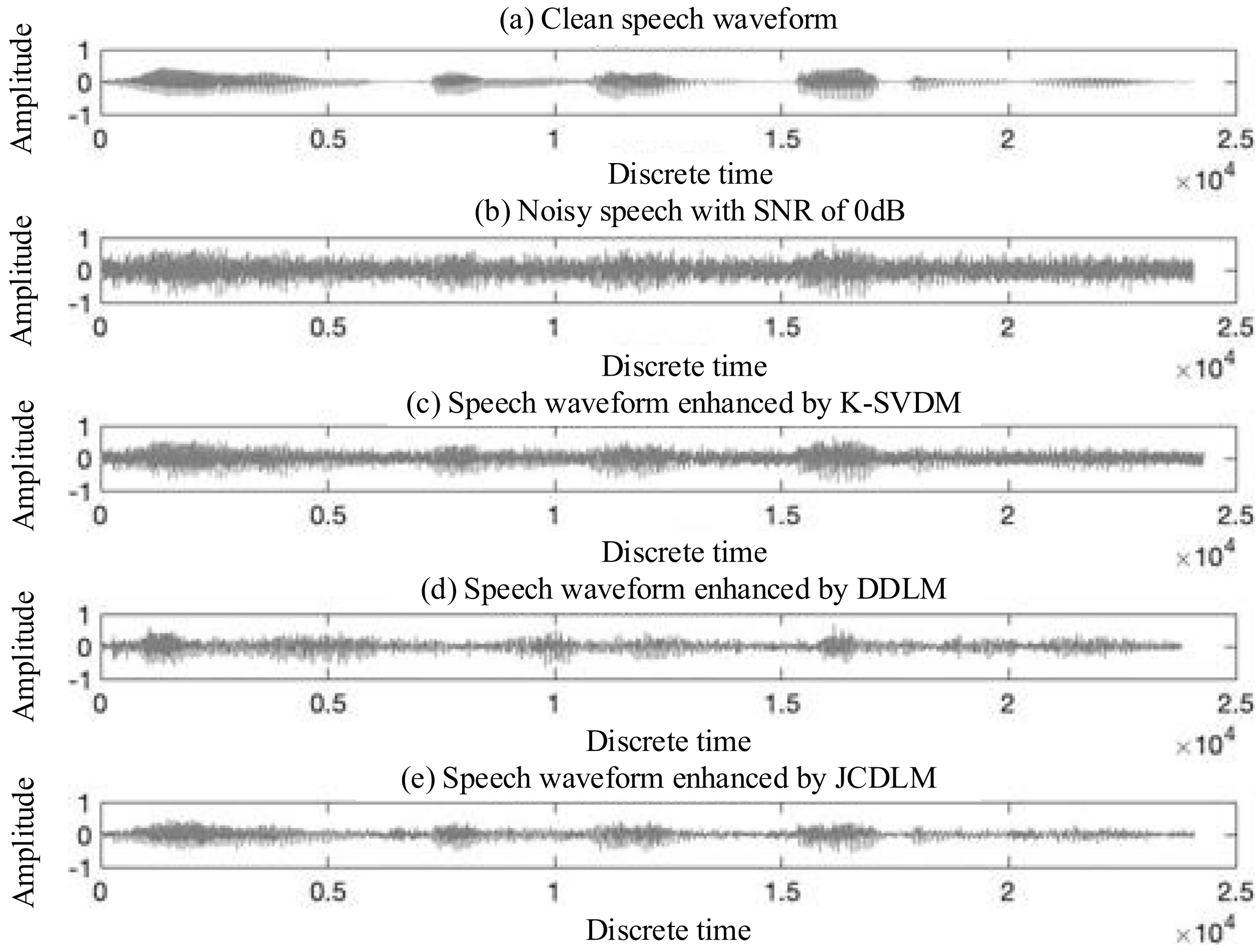


Fig. 8 Enhanced waveforms of different algorithms under 0-dB white noise

The detailed process of the enhancement algorithm is shown in Table 1.

4 Experiment and results analysis

4.1 Dataset and experimental setup

The speech signals used in this paper come from the LibriSpeech corpus with a sampling of 16 kHz, and the noise signals are from the Noisex-92 corpus. Three kinds of noise, including vehicle noise (Volvo), white noise (White), and F-16 cockpit noise (F16), are selected in the experiment. We randomly select 100 speech signals, including 80 sentences as the training set and 20 sentences as the test set. The noisy signal is generated by adding noise to the speech signal. Signal-to-noise ratios (SNR) of the noisy speech is − 10 dB, − 5 dB, 0 dB, 5 dB, and 10 dB.The speech signal and noise signal are divided into frames by a rectangular window, and the experimental results of the test signals are averaged.

4.2 Influence of adjustment factor

In order to measure the influence of the adjustment factors in (6) on the speech enhancement performance, we select the balance factors by averaging the global SNR increment after speech enhancement under three kinds of noise interference with SNR of 0 dB. The increment refers to the increase of the global SNR after speech enhancement. With the increase of the value, the speech enhancement effect is better.

Firstly, we measure the influence of λ on the speech enhancement performance. We assume that the value of the other two balance factors is 1. The mean value of the global SNR increment under different λ is shown in Fig. 4. When λ is 0, the inner product between sub-dictionaries is not constrained, and the global SNR increment is the lowest, which indicates that the correlation between sub-dictionaries can improve the discrimination of dictionaries and the enhancement effect of the system. When λ increases from 0 to 1, the global SNR increment rises rapidly. When λ ranges from

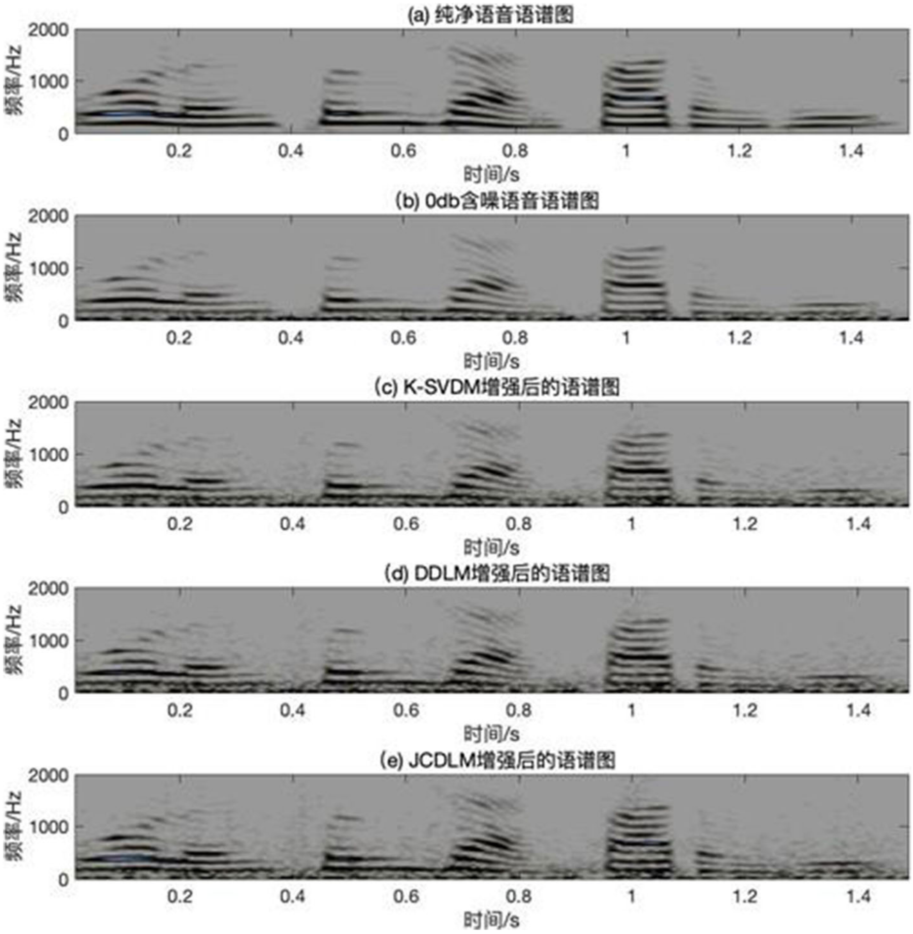


Fig. 9 Spectrum enhanced of different algorithms under 0-dB vehicle noise

1 to 16, the global SNR increment declines slowly. It can be explained by the fact that the dictionary atoms become more distinguishable with the increase of λ. However, when λ > 1, the richness of the atoms is reduced, which weakens the sparse representation ability of the joint dictionary, causing large reconstruction error and poor enhancement performance. Therefore, when the value of α and β is 1, we select $\lambda = 1$ to obtain the optimal sparse dictionary for the best speech enhancement effect in the global SNR increment.

Secondly, we study the effects of balance factors α and β. When λ is the optimal value 1 and β is the initial value 1, the improved average values of global SNR with different α are shown in Fig. 5. We can see when α is from 0 to 0.3, the performance of the enhancement algorithm is improved continuously, and when α goes from 0.3 to 1, the performance of the enhancement algorithm begins to decline. Thus, when $\lambda = 1$ and $\beta = 1$, we choose $\alpha = 0.3$ to get the best performance. As can be seen from Fig. 6, when λ and α are set as the best values of 1 and 0.3, respectively, the mean value of global SNR increment under different β is calculated. We think about β in the range from 0 to 0.1. When β is 0.01, the performance of the enhancement algorithm reaches the peak. Therefore, in the following experiments, we take $\lambda = 1$, $\alpha = 0.3$, and $\beta = 0.01$.

4.3 Influence of parameters

The noisy speech with white noise whose SNR is − 10 dB, − 5 dB, 0 dB, 5 dB, 10 dB is taken as the test set. We analyze the influence of dictionary redundancy and frame length on speech enhancement performance. The average of the global SNR increment is used to make the judgment.

4.3.1 Influence of frame length

The performance of speech enhancement algorithms with the frame lengths of 128, 256, and 512 is discussed when the dictionary redundancy is 1. As can be seen from Table 2, the frame length has a great influence on the enhancement performance. When the frame length is 256, the SNR increment of the proposed algorithm is greater than that of other frame lengths. Therefore, we

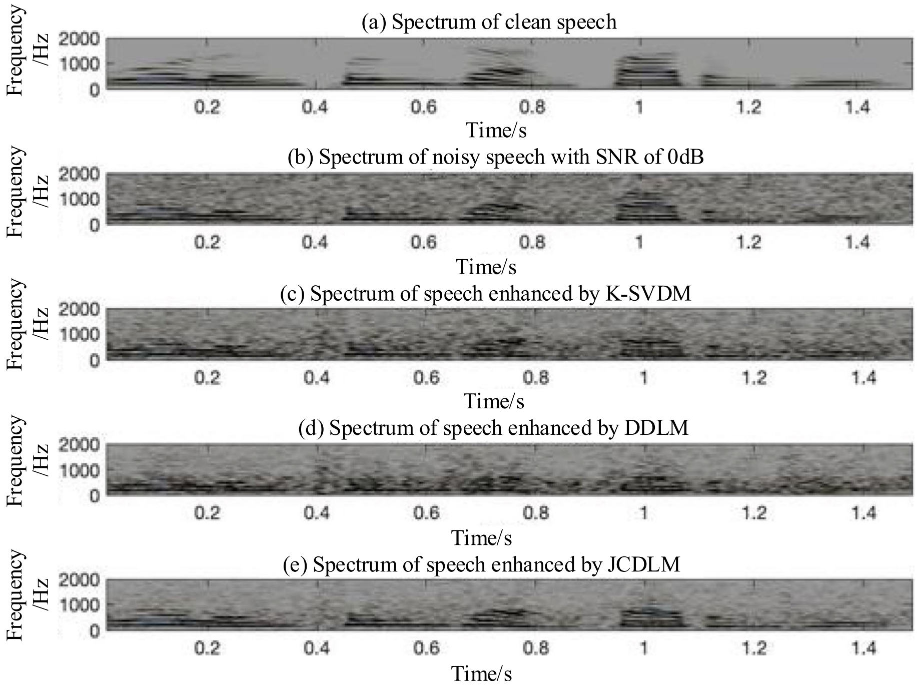


Fig. 10 Spectrum enhanced of different algorithms under 0-dB white noise

choose the frame length of 256 in the subsequent experiments.

4.3.2 Influence of dictionary redundancy

The performance of speech enhancement algorithms for the sub-dictionary with the redundancy of 1, 2, and 3 is discussed when the frame length is 256. According to Table 3, we get the following results. (1) The greater the dictionary redundancy, the better the speech enhancement effect. (2) The SNR increment of the enhanced speech increases greatly when the redundancy is increased from 1 to 2, but the SNR increment of the enhanced speech increases slightly when the redundancy is increased from 2 to 3. (3) It is also found in the actual process that the time of training dictionary in speech enhancement will be longer with the increase of the sub-dictionary redundancy. Therefore, in the actual application, we need to select the dictionary redundancy by considering performance requirements and operation speed. Finally, we choose the sub-dictionary with a redundancy of 2 to achieve speech enhancement in this paper.

4.4 Performance comparison with other algorithms

(1) In order to better verify the enhancement effect of our proposed joint constrained dictionary learning method (JCDLM), we compare it with the K-singular value decomposition method (K-SVDM) [12] and distinguishing dictionary learning method (DDLM) [20] in white and colored noise environments. In the experiment, the frame length is 256 and the redundancy of the sub-dictionary is 2, so the size of the sub-dictionary is256×512.

The waveforms of clean speech, 0-dB speech with vehicle noise, and enhanced speech based on K-SVDM, DDLM, and our proposed JCDLM are shown in Fig. 7. As can be seen from the waveforms, the speech waveform enhanced by K-SVDM is still quite different from the clean speech waveform, and the speech enhancement effect is not ideal. The speech waveform enhanced by DDLM is better than that by K-SVDM, but worse than that by JCDLM. The speech waveform enhanced by JCDLM is closer to the clean speech waveform, and the non-speech segment waveform has fewer defects, which shows a better enhancement effect based on JCDLM. It can be explained that JCDLM not only considers the reconstruction error between the noisy signal and the source signal but also considers the cross interference between the sub-dictionaries. By constraining the objective function to train the joint dictionary, the speech components in the noisy signal can be sparsely expressed in the speech sub-dictionary as far as possible without the interference of the noise sub-dictionary. Therefore, the enhanced speech waveform has less defects than the other two algorithms, which is almost the same as the clean speech waveform.

Figure 8 shows the speech enhancement effects of different algorithms when inputting 0-dB speech with

Table 4 Increment of SNR in different noise environments (dB)

Noise type	Algorithm	SNR/dB				
		− 10	− 5	0	5	10
Volvo	K-SVDM	6.1768	4.8574	2.5872	− 0.2397	− 0.9569
	DDLM	14.9362	12.5592	9.2094	6.4518	5.5384
	JCDLM	17.2179	15.1633	10.7918	7.8586	6.0327
White	K-SVDM	2.4519	2.2735	1.9674	1.7807	1.1582
	DDLM	8.4895	5.887	4.9343	3.9035	2.4761
	JCDLM	9.6935	7.0714	6.0125	4.6939	2.9905
F16	K-SVDM	1.9365	1.0518	0.9635	0.6829	0.2835
	DDLM	7.7137	4.0154	1.493	1.1331	0.3107
	JCDLM	8.0659	4.8726	2.0129	1.3358	0.4882

white noise. From the figure, we can also observe that the speech waveform using the speech enhancement algorithm based on JCDLM is clearer and closer to the clean speech waveform than that of K-SVDM and DDLM. Therefore, it can be proved that the speech enhancement algorithm based on our proposed JCDLM also has a good enhancement effect in the white noise environment.

Under inputting 0-dB speech with vehicle noise, the spectrums of clean speech, noisy speech, and enhanced speech based on K-SVDM, DDLM, and JCDLM are shown in Fig. 9. The energy of vehicle noise is mainly concentrated in the low-frequency band, which affects the intelligibility of speech. From Fig. 9, it can be observed that compared with the other two algorithms, the spectrum enhanced by JCDLM is more similar to that of the clean speech in the low-frequency band, and the information components are more clearly preserved. It indicates that the speech enhancement algorithm based on our proposed JCDLM has a good enhancement effect in the vehicle noise environment.

Figure 10 shows the enhanced spectrums of different algorithms when inputting 0-dB white noise speech. As can be seen from the figure, there is still a lot of residual noise in K-SVDM speech enhancement, and the speech enhancement effect is not ideal. The residual noise of the speech after DDLM enhancement is relatively small, but the harmonic structure in the low-frequency band is not obvious, and some components in the clean speech are lost. The enhanced speech by JCDLM not only removes residual noise but also maintains a good harmonic structure. Compared with the other two algorithms, the enhanced speech quality is greatly improved. The speech enhancement algorithm based on our proposed JCDLM also performs well in a white noise environment.

(2) In order to better compare the speech enhancement effects of different algorithms, Table 4 shows the SNR increment of speech enhancement in the different noise environments. It can be concluded that compared with K-SVDM and DDLM, the SNR increment of JCDLM is higher in Volvo, White, and F16 noise environments. Especially in the vehicle noise environment, the SNR increment based on JCDLM is much higher than the other two speech enhancement algorithms. It can also be seen from the table that JCDLM has a more obvious enhancement effect in the low-SNR environment than in the high-SNR environment.

(3) In order to better compare the speech enhancement effects of various methods, we will further measure the enhancement performance from the MOS calculated by PESQ. The MOS of speech is directly proportional to the intelligibility of speech. The higher the MOS, the better the speech quality. The MOS of each method under vehicle, white, and F16 noise with different input SNR are shown in Fig. 11. It can be found that compared with K-SVDM and DDLM, JCDLM has a higher MOS in any noise environment, and the speech clarity and intelligibility are improved to a certain extent, especially in the vehicle noise environment. Therefore, it can be concluded that JCDLM has a better enhancement effect in the noise environment with SNR of − 10 to 10 dB, especially in the vehicle noise environment.

(4) We also use LSD to measure the enhanced speech distortion. If the value of LSD is smaller, the spectrum distortion of the enhanced speech is smaller and the signal is closer to the clean speech, which shows that the speech enhancement effect is better. The LSDs of each method under vehicle, white, and F16 noise with different input SNR are shown in Fig. 12. Compared with the other two algorithms, JCDLM has smaller LSD in the vehicle noise, white noise, and F16 noise environments, and the enhanced speech distortion is smaller. Therefore, it shows that JCDLM can better describe the distribution of sparse dictionary and coefficient matrix in the same SNR noise environment and can restore clean speech more accurately with a better enhancement effect.

5 Conclusion

In the traditional single-channel speech enhancement algorithm based on joint dictionary learning, clean speech signal and noise signal are used to train the corresponding sub-dictionaries, respectively, and then the two sub-dictionaries are spliced into a joint dictionary. At last, the noisy signal is projected onto the joint dictionary to recover the enhanced speech signal. The traditional speech enhancement algorithm based on joint dictionary learning only considers the characteristics of the given signal itself and does not consider the similarity between them, so that some of the speech components in the noisy signal will still be projected on the interference noise sub-dictionary, which leads to the occurrence of

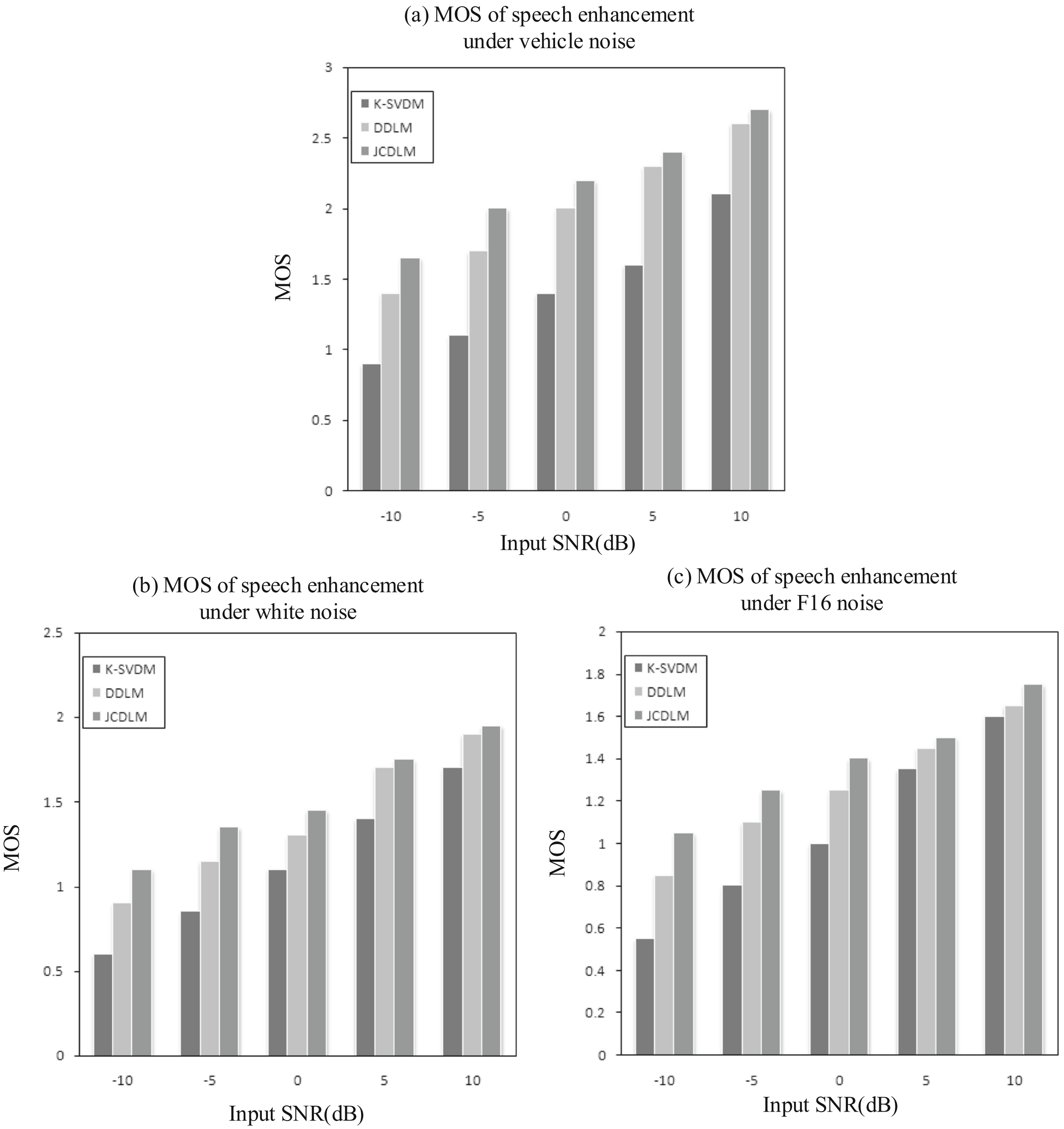
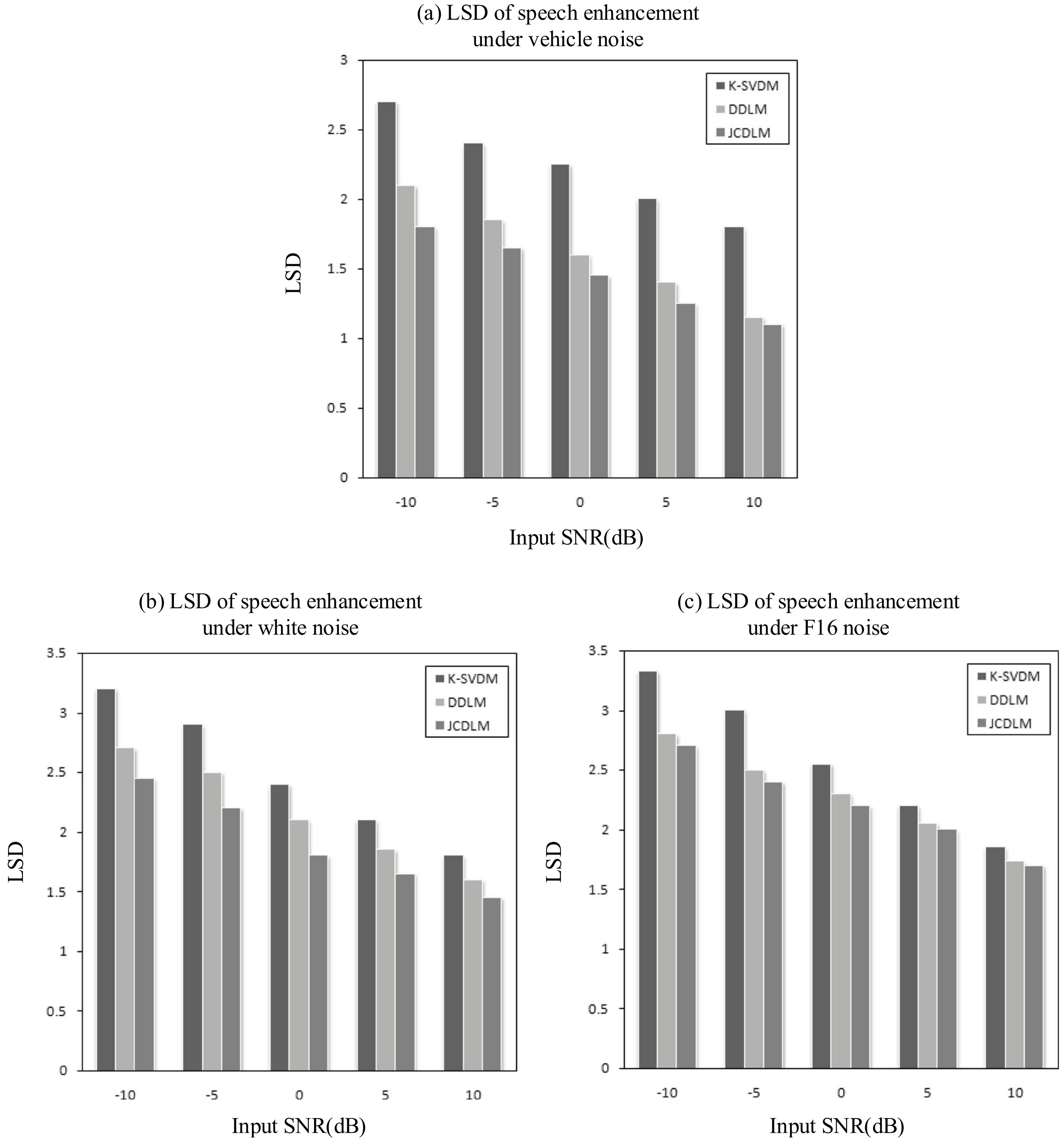


Fig. 11 MOS in different noise environments

cross projection and makes the speech enhancement performance not reach the best. To solve this problem, we consider the joint constraint relationship between speech sub-dictionary and noise sub-dictionary and propose a new optimization function in the training stage. The function not only constrains the approximation error of reconstruction but also improves the discrimination of sub-dictionaries. The weight coefficient is used to allocate the constraints, which greatly reduces the range of solutions and the training time, and makes the signal representation in the joint dictionary sparser. In the enhancement stage, the speech components of the noisy signal can be more projected onto the speech sub-dictionary without being interfered by the noise sub-dictionary, so that the enhanced speech quality and intelligibility are higher. The experimental results show that the speech enhancement algorithm based on our proposed joint constrained dictionary learning has a



(a) LSD of speech enhancement under vehicle noise

(b) LSD of speech enhancement under white noise

(c) LSD of speech enhancement under F16 noise

Fig. 12 LSD in different noise environments

better denoising effect comparing with the traditional K-SVDM and DDLM in time domain waveform, spectrogram, global signal-to-noise ratio, subjective evaluation of speech quality, and logarithmic spectrum distance.

Acknowledgements
Not applicable.

Authors' contributions
Linhui Sun designed the core methodology of the study, carried out the implement, and drafted the manuscript. Yunyi Bu and Zihao Wu carried out the experiments and drafted the manuscript. Pingan Li conducted a formal analysis and validation. All authors read and approved the final manuscript.

Funding
This work is supported by the National Natural Science Foundation of China (No.61901227, No.62071242), and the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (No.19KJB510049).

Availability of data and materials
Not applicable.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 30 May 2021 Accepted: 16 July 2021
Published online: 27 July 2021

References

1. S. Samui, I. Chakrabarti, S.K. Ghosh, Improved single channel phase-aware speech enhancement technique for low signal-to-noise ratio signal. IET Signal Process. **10**(6), 641–650 (2016). https://doi.org/10.1049/iet-spr.2015.0182
2. T. Lavanya, T. Nagarajan, P. Vijayalakshmi, Multi-level single-channel speech enhancement using a unified framework for estimating magnitude and phase spectra, IEEE/ACM Transactions on Audio, Speech, and Language Processing. **28**, 1315-1327 (2020). https://doi.org/10.1109/TASLP.2020.2986877
3. P. Kajla, N.V. George, Speech quality enhancement using a two channel sparse adaptive filtering approach, Applied Acoustics. **158** (2020). https://doi.org/10.1016/j.apacoust.2019.107035
4. N. Saleem, M.I. Khattak, M. Shafi, Unsupervised speech enhancement in low SNR environments via sparseness and temporal gradient regularization. Appl. Acoustics. **141**, 333–347 (2018). https://doi.org/10.1016/j.apacoust.2018.07.027
5. C. You, B. Ma, Spectral-domain speech enhancement for speech recognition. Speech Commun.. **94**, 30–41 (2017). https://doi.org/10.1016/j.specom.2017.08.007
6. W. Zaw, A. T. H. Soe, Speaker identification using power spectral subtraction method, 2019 16th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON). 625-628 (2019). https://doi.org/10.1109/ECTI-CON47248.2019.8955344.
7. J. Choi, J. Chang, On using acoustic environment classification for statistical model-based speech enhancement. Speech Communication. **54**(3), 477–490 (2012). https://doi.org/10.1016/j.specom.2011.10.009
8. A. Salman, E. Muhammad, K. Khurshid, A subspace approach for speech enhancement using frame-level AdaBoost classification, 2007 International Conference on Electrical Engineering. 1-6 (2007). https://doi.org/10.1109/ICEE.2007.4287303
9. M. Sadeghi, M. Babaie-Zadeh, C. Jutten, Dictionary learning for sparse representation: a novel approach. IEEE Signal Process. Lett. **20**(12), 1195–1198 (2013). https://doi.org/10.1109/LSP.2013.2285218
10. R. Rubinstein, A.M. Bruckstein, M. Elad, Dictionaries for sparse representation modeling. Proc. IEEE. **98**(6), 1045–1057 (2010). https://doi.org/10.1109/JPROC.2010.2040551
11. V. Abrol, P. Sharma, A.K. Sao, Greedy double sparse dictionary learning for sparse representation of speech signals. Speech Commun. **85**, 71–82 (2016). https://doi.org/10.1016/j.specom.2016.09.004
12. M. Aharon, M. Elad, A. Bruckstein, K-SVD: an algorithm for designing overcomplete dictionaries for sparse representations. IEEE Transact. Signal Process. **54**(11), 4311–4322 (2006). https://doi.org/10.1109/TSP.2006.881199
13. B.V. Gowreesunker, A.H. Tewfik, Learning sparse representation using iterative subspace identification. IEEE Transact. Signal Process. **58**(6), 3055–3065 (2010). https://doi.org/10.1109/TSP.2010.2044251
14. L. Sun, K. Xie, T. Gu, J. Chen, Z. Yang, Joint dictionary learning using a new optimization method for single-channel blind source separation. Speech Commun. **106**, 85–94 (2019). https://doi.org/10.1016/j.specom.2018.11.008
15. M. Islam, Y. Zhu, M. I. Hossain, R. Ullan, Z. Ye, Supervised single channel dual domains speech enhancement using sparse non-negative matrix factorization, Digital Signal Process. **100** (2020). https://doi.org/10.1016/j.dsp.2020.102697
16. C.D. Sigg, T. Dikk, J.M. Buhmann, Speech enhancement using generative dictionary learning, IEEE Transactions on Audio. Speech Language Process. **20**(6), 1698–1712 (2012). https://doi.org/10.1109/TASL.2012.2187194
17. N. Mohammadiha, P. Smaragdis, A. Leijon, Supervised and unsupervised speech enhancement using nonnegative matrix factorization. IEEE Transact Audio Speech Language Process. **21**(10), 2140–2151 (2017). https://doi.org/10.1109/TASL.2013.2270369
18. D. Baby, T. Virtanen, J. F. Gemmeke, H. Van hamme, Coupled dictionaries for exemplar-based speech enhancement and automatic speech recognition, IEEE Transact. Audio Speech Language Process. **23**(11), 1788-1799 (2015). https://doi.org/10.1109/TASLP.2015.2450491
19. P. Sprechmann, A. Bronstein, M. Bronstein, G. Sapiro, Learnable low rank sparse models for speech denoising, 2013 IEEE International Conference on Acoustics. Speech and Signal Process., 136–140 (2013). https://doi.org/10.1109/ICASSP.2013.6637624
20. L. Zhang, G. Bao, Y. Luo, Z. Ye, Monaural speech enhancement using joint dictionary learning with cross-coherence penalties, international symposium on computational intelligence & design. 518-522 (2015). https://doi.org/10.1109/ISCID.2015.162
21. J. Fu, L. Zhang, Z. Ye, Supervised monaural speech enhancement using two-level complementary joint sparse representations. Appl. Acoustics. **132**, 1–7 (2018). https://doi.org/10.1016/j.apacoust.2017.11.005
22. H. Jia, W. Wang, Y. Wang, J. Pei, Speech enhancement based on discriminative joint sparse dictionary alternate optimization. J. Xidian Univ. **46**(03), 74–81 (2019)
23. L. Sun, C. Zhao, M. Su, F. Wang, Single-channel blind source separation based on joint dictionary with common sub-dictionary, Int. J. Speech Technol. **21** 19–27 (2018). https://doi.org/10.1007/s10772-017-9469-2
24. F. F. Firouzeh, S. Ghorshi, S. Salsabili, Compressed sensing based speech enhancement, 2014 8th International Conference on Signal Processing and Communication Systems (ICSPCS). 1-6 (2014). https://doi.org/10.1109/ICSPCS.2014.7021068
25. P. Qi, W. Zhou, J. Han, A method for stochastic L-BFGS optimization, 2017 IEEE 2nd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA). 156-160 (2017). https://doi.org/10.1109/ICCCBDA.2017.7951902

Publisher's Note