IEICE TRANS. INF. & SYST., VOL.E100–D, NO.3 MARCH 2017

PAPER

A Speech Enhancement Method Based on Multi-Task Bayesian Compressive Sensing

Hanxu YOU[†a)], ***Nonmember***, **Zhixian MA**[†], ***Student Member***, **Wei LI**[†], ***and*** **Jie ZHU**[†], ***Nonmembers***

SUMMARY Traditional speech enhancement (SE) algorithms usually have fluctuant performance when they deal with different types of noisy speech signals. In this paper, we propose multi-task Bayesian compressive sensing based speech enhancement (MT-BCS-SE) algorithm to achieve not only comparable performance to but also more stable performance than traditional SE algorithms. MT-BCS-SE algorithm utilizes the dependence information among compressive sensing (CS) measurements and the sparsity of speech signals to perform SE. To obtain sufficient sparsity of speech signals, we adopt overcomplete dictionary to transform speech signals into sparse representations. K-SVD algorithm is employed to learn various overcomplete dictionaries. The influence of the overcomplete dictionary on MT-BCS-SE algorithm is evaluated through large numbers of experiments, so that the most suitable dictionary could be adopted by MT-BCS-SE algorithm for obtaining the best performance. Experiments were conducted on well-known NOIZEUS corpus to evaluate the performance of the proposed algorithm. In these cases of NOIZEUS corpus, MT-BCS-SE is shown that to be competitive or even superior to traditional SE algorithms, such as optimally-modified log-spectral amplitude (OMLSA), multi-band spectral subtraction (SSMul), and minimum mean square error (MMSE), in terms of signal-noise ratio (SNR), speech enhancement gain (SEG) and perceptual evaluation of speech quality (PESQ) and to have better stability than traditional SE algorithms.

key words: *speech enhancement, compressive sensing, overcomplete dictionary, sparse representation*

1. Introduction

In the past decades, speech enhancement (SE) has been one of the most active research topics in speech and audio signal processing applications such as speech recognition, speaker recognition. SE usually aims to reduce a particular type of interference like additive noise, so that SE algorithms normally utilize some dynamic characteristics of the noise to improve their performance [1]. In this paper, we focus on denoising additive noise.

SE algorithms that aim to suppress additive noise can be divided into some categories: spectral subtractive methods, subspace methods, and statistical based methods [2]. Spectral subtractive methods, such as multi-band spectral subtraction (SSMul), normally estimate noise spectra first and then subtract them from noisy observation spectra to derive the clean speech spectra. Subspace methods decompose an observed noisy signal into mutually orthogonal speech and noise subspaces. Methods of this type normally assume that speech and noise signals are located in two orthogonal subspaces [3], [4]. Statistical modelling methods then improve their denoising capacities for nonstationary noises by employing some statistical models to maximize certain statistical criteria, such as maximum likelihood (ML) and minimum mean square error (MMSE). Algorithms of this type normally are based on an assumption that speech and noise signals always obey certain probability distributions [5], [6]. It could be observed that these traditional SE algorithms which are proposed based on different denoising strategies usually have fluctuant performance when they deal with different types of noisy speech signals. For example, SSMul is effective for denoising stationary noise, but it's limited to non-stationary noises [7]. In real world situations, without prior knowledge of noise type, we cannot judiciously choose the most suitable SE algorithm from a set of viable algorithms. In this paper, we try to propose an SE method to achieve not only comparable to but also more stable performance than traditional SE algorithms in different environments.

Compressive sensing, proposed by Donoho [8], Candes and Tao [9], has been established on solving an l_0-norm minimization problem to recover a sparse or compressible signal from its downsamples at a low rate [10]. Though CS has been used in digital image processing [11] for many years, it also raises researchers' concern in speech and audio signal processing recently. As for speech enhancement, D. Wu et al. already proposed CS-based methods to address speech enhancement in sparse noisy [12] and non-sparse noisy environments [13]. They also proposed an advanced CoSaMP recovery algorithm termed Tdn-CoSaMP to achieve SE in [14]. Those CS-based SE algorithms usually adopt some pre-specific dictionaries (such as DCT, and wavelet transform) to be as the sparse basis.

In this paper, multi-task Bayesian compressive sensing based speech enhancement (MT-BCS-SE) algorithm is proposed. Two important points of MT-BCS-SE algorithm are that the measurements of clean speech signals are not statistically independent and speech signals are sparse enough in certain domain such as frequency domain, wavelet domain, and others. MT-BCS-SE algorithm utilizes the two characteristics of speech signals to improve SE performance. For the first point, from a statistical standpoint, multiple compressive sensing measurements are related to each other. Therefore, if taking those statistical information into consideration, MT-BCS-SE may improve the SE performance. For the second point, because the additive noise is normally

Manuscript received August 17, 2016.
Manuscript revised October 28, 2016.
Manuscript publicized November 30, 2016.
[†]The authors are with the Dept. of Electronic Engineering, Shanghai Jiao Tong University (SJTU), Shanghai, 200240, China.
a) E-mail: gongzihan@sjtu.edu.cn (Corresponding author)
DOI: 10.1587/transinf.2016EDP7350

not sparse at all, so that sparsity of speech signals guarantees that MT-BCS-SE can recover the clean speech signals from the measurements. Overcomplete dictionary is used to obtain sparse representations of noisy speech signals. We adopt K-SVD algorithm to learn various overcomplete dictionaries, and the influence of dictionary is also evaluated in this paper. The most suitable dictionary could be adopted by the proposed algorithm for obtaining the best SE performance.

This CS-based SE algorithm may alleviate some difficulties that the traditional SE methods are confronted with, because it avoids some of their assumptions such as the noise invariant (spectral subtractive methods) and the mutual orthogonality (subspace methods). MT-BCS-SE algorithm also has some advantages over the traditional SE methods: (i) no necessity of noise spectrum; (ii) various suitable dictionaries are available; (iii) taking advantage of the dependence information among compressive sensing measurements.

The rest of this article is organized as follows. In Sect. 2, we briefly introduce CS theory. And then we introduce MT-BCS-SE algorithm in Sect. 3. In Sect. 4, K-SVD algorithm is adopted to learn the overcomplete dictionaries for obtaining sparse representations of speech signals. We carried out experiments to evaluate the performance of MT-BCS-SE algorithm and several results are presented in Sect. 5. Finally, we draw our conclusions in Sect. 6.

2. Compressive Sensing

Considering a real-valued, one-dimensional, discrete-time signal $\boldsymbol{x}$ of size $n\times 1$ and a basis $\boldsymbol{\Psi}$ of size $n\times N$, thus $\boldsymbol{x} = \Psi \boldsymbol{s}$ can be obtained, where $\boldsymbol{s}$ is an $N \times 1$ vector. Supposing $k\,(k \ll N)$ elements of $\boldsymbol{s}$ are nonzero or largest, and the $n-k$ remaining elements are zero or negligible, then we say that $\boldsymbol{x}$ is k-sparse or compressible. After measuring or sampling this sparse signal, a $m \times 1$ $(k \le m \ll N)$ measurement $\boldsymbol{y}$ can be obtained as follows:

$$\boldsymbol{y} = \boldsymbol{\Phi x} = \boldsymbol{\Phi\Psi s} = \boldsymbol{\Theta s} \tag{1}$$

Where $\boldsymbol{\Phi}$ is a $m\times n$ measurement matrix (MM) and $\boldsymbol{\Theta} = \boldsymbol{\Phi\Psi}$ satisfies restricted isometry property (RIP) [9] which ensures the sparse representation $\boldsymbol{s}$ has a unique solution. $\boldsymbol{\Theta}$ is of size $m \times N$. Because the dimension of $\boldsymbol{y}$ is far lower than $\boldsymbol{s}$ $(m \ll N)$, so that obtaining $\boldsymbol{s}$ from measurement $\boldsymbol{y}$ by solving the linear equation in (1) is NP-hard. Many algorithms including OMP, CoSaMP, BP, etc, were proposed to solve this problem. In Sect. 4, we adopted BatchOMP [15], a fast algorithm for multiple OMP decompositions, to implement sparse decomposition over the same dictionary in sparse coding stage of K-SVD algorithm.

3. Multi-Task Bayesian Compressive Sensing Based Speech Enhancement

In this section, multi-task Bayesian compressive sensing based speech enhancement (MT-BCS-SE) algorithm is proposed to achieve speech enhancement from noisy speech signals.

3.1 MT-BCS-SE

The speech enhancement tasks can be formulated in the framework of CS theory as below:

Let $\boldsymbol{x}_i^* = \boldsymbol{x}_i + \boldsymbol{n}_i$ be the i-th noisy speech signal, where $\boldsymbol{x}_i$ is clean speech signal and $\boldsymbol{n}_i$ is additive noise, so that the i-th compressive sensing measurement $\boldsymbol{y}_i$ can be given as:

$$\boldsymbol{y}_i = \boldsymbol{\Phi x}_i^* = \boldsymbol{\Phi}(\boldsymbol{x}_i + \boldsymbol{n}_i) = \boldsymbol{\Phi\Psi s}_i + \boldsymbol{n}_i^* \tag{2}$$

where $\boldsymbol{s}_i$ is the sparse representation of clean speech signal $\boldsymbol{x}_i$ and $\boldsymbol{n}_i^*$ is the noisy measurement. We assume that $\boldsymbol{n}_i^*$ draws a zero-mean Gaussian random variable with a variance $1/\alpha$. Parameter α is associated with additive noise $\boldsymbol{n}^*$. We therefore have a Gaussian likelihood function for $\boldsymbol{s}_i$ and α based on the measurement $\boldsymbol{y}_i$:

$$p(\boldsymbol{y}_i|\boldsymbol{s}_i, \alpha) = (2\pi/\alpha)^{-m/2} \exp\left(-\frac{\alpha}{2}\|\boldsymbol{y}_i - \boldsymbol{\Theta s}_i\|_2^2\right) \tag{3}$$

where $\boldsymbol{\Theta} = \boldsymbol{\Phi\Psi}$ satisfies RIP and m represents the dimension of measurement vector $\boldsymbol{y}_i$.

We define that recovering clean representation $\boldsymbol{s}_i$ from $\boldsymbol{y}_i$ as Task-i, and the coefficients of $\boldsymbol{s}_i$ are assumed to be drawn from a common zero-mean Gaussian prior that is shared among all sparse representations $\{\boldsymbol{s}_i\}$ $(i = 1, 2, \ldots, M_t)$. Taking the information that M_t tasks are statistically dependent into consideration, MT-BCS-SE processes M_t tasks jointly and simultaneously. Let s_{ij} represent the j-th sparse coefficient of $\boldsymbol{s}_i$ for Task-i, we then have

$$p(\boldsymbol{s}_i|\boldsymbol{\beta}) = \prod_{j=1}^{n} \mathcal{N}(s_{ij}|0, 1/\beta_j) \tag{4}$$

The parameter $\boldsymbol{\beta}$ is shared among all M_t tasks. That's to say, all measurements $\{\boldsymbol{y}_i\}$ $(i = 1, 2, \ldots, M_t)$ will contribute to learning $\boldsymbol{\beta}$. Because the noise $\boldsymbol{n}_i^*$ obeys zero-mean Gaussian distribution and the sparse coefficients of $\boldsymbol{s}_i$ are assumed to be drawn from a common zero-mean Gaussian prior, so that MT-BCS-SE would be more suitable in dealing with Gaussian white noises rather than the non-Gaussian noises.

Further, the Gamma priors could be placed on α and $\boldsymbol{\beta}$ as follows:

$$p(\alpha|a, b) = \Gamma(\alpha|a, b) \tag{5}$$

$$p(\boldsymbol{\beta}|c, d) = \prod_{j=1}^{n} \Gamma(\beta_j|c, d) \tag{6}$$

Here, for α, $a = b = 0$ corresponds to a commonly-used improper prior expressing a priori ignorance about values for the noise variance [16]. Appropriate choice of parameters c and d ensures the sparsity of $\boldsymbol{s}_i$. For computational simplifications and avoiding subjective choice of those parameters, so that we recommend $c = d = 0$ as a default choice.

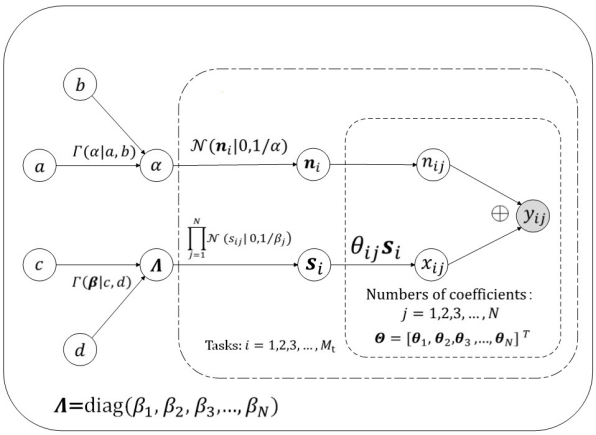


Fig. 1 A hierarchical Bayesian model for MT-BCS-SE.

According to the characteristic that multiple CS measurements are statistically related to each other, we consider the compressive sensing framework from a hierarchical Bayesian perspective [17]. A hierarchical Bayesian model for MT-BCS-SE is illustrated in Fig. 1. We can infer a posterior density function (PDF) on α and $\boldsymbol{\beta}$ from a Bayesian piont of view, then we get

$$p(\alpha, \boldsymbol{\beta}|\boldsymbol{y}, a, b, c, d) = \frac{p(\boldsymbol{y}|\alpha, \boldsymbol{\beta})p(\alpha|a, b)p(\boldsymbol{\beta}|c, d)}{\iint p(\boldsymbol{y}, \alpha, \boldsymbol{\beta})\mathrm{d}\alpha\mathrm{d}\boldsymbol{\beta}} \tag{7}$$

Given that a, b, c and d are very small ($a, b, c, d \to 0$), a point estimate $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$ for the parameters α and $\boldsymbol{\beta}$ can be interpreted as a ML estimate which is written as below:

$$\begin{aligned}(\hat{\alpha}, \hat{\boldsymbol{\beta}}) &= \arg\max \mathcal{L}(\alpha, \boldsymbol{\beta}) = \arg\max \log p(\boldsymbol{y}|\alpha, \boldsymbol{\beta}) \\ &= \arg\max \sum_{i=1}^{M_t} \int \log p(\boldsymbol{y}_i|\boldsymbol{s}_i, \alpha)p(\boldsymbol{s}_i|\boldsymbol{\beta})\mathrm{d}\boldsymbol{s}_i \end{aligned} \tag{8}$$

Using $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$, we can calculate the posterior density function for the sparse representation $\boldsymbol{s}_i$. The posterior of $\boldsymbol{s}_i$ can be expressed as a multivariate Gaussian distribution with mean μ_i and covariance σ_i^2.

$$\begin{aligned}p(\boldsymbol{s}_i|\boldsymbol{y}_i, \hat{\alpha}, \hat{\boldsymbol{\beta}}) &= \frac{p(\boldsymbol{y}_i|\boldsymbol{s}_i, \hat{\alpha})p(\boldsymbol{s}_i|\hat{\boldsymbol{\beta}})}{\int p(\boldsymbol{y}_i|\boldsymbol{s}_i, \hat{\alpha})p(\boldsymbol{s}_i|\hat{\boldsymbol{\beta}})\mathrm{d}\boldsymbol{s}_i} \\ &= \mathcal{N}(\boldsymbol{s}_i|\mu_i, \sigma_i^2) \end{aligned} \tag{9}$$

with

$$\mu_i = \hat{\alpha}\sigma_i^2\boldsymbol{\Theta}^T\boldsymbol{y}_i \quad \text{and} \quad \sigma_i^2 = (\hat{\alpha}\boldsymbol{\Theta}^T\boldsymbol{\Theta} + \boldsymbol{\Lambda})^{-1} \tag{10}$$

where $\boldsymbol{\Lambda} = \mathrm{diag}(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_n)$, and $\{\hat{\beta}_i\}$ $(i = 1, 2, \ldots, n)$ are the components of the vector $\hat{\boldsymbol{\beta}}$.

As implemented in (8) and illustrated in Fig. 1, the point estimate $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$ is based upon all M_t measurements $\{\boldsymbol{y}_i\}$ $(i = 1, 2, \ldots, M_t)$, emphasizing the multi-task nature of the analysis. All measurements are used to α and $\boldsymbol{\beta}$, while to update an approximation to the sparse coefficients of $\boldsymbol{s}_i$, only the measurement $\boldsymbol{y}_i$ of Task-i is employed. This suggests an iterative algorithm that alternates between these global and local solutions.

3.2 Reconstruction via Relevance Vector Machine

To recover the clean speech signals from measurements, we should do the estimation of $\{\boldsymbol{s}_i\}$ $(i = 1, 2, \ldots, M_t)$ according to $\boldsymbol{\beta}$. Though only the associated $\boldsymbol{y}_i$ is used to update approximation of the corresponding $\boldsymbol{s}_i$, the estimate of $\boldsymbol{\beta}$ is based on all M_t measurements. To perform an exact ML estimation of $\boldsymbol{\beta}$ still computes expensively, so that a fast relevance vector machine (RVM) algorithm [18] is applied to do a faster estimation of $\boldsymbol{\beta}$ rather than an exact estimation in this paper.

Compared with other mostly used iterative algorithm like OMP, RVM operates in a constructive manner, i.e., sequentially adding or deleting candidate atoms to model until all m non-zero coefficients have been included. Detailed analysis [18] shows that the complexity of RVM is more related to m (the dimension of measurement $\boldsymbol{y}_i$) rather than N (the dimension of sparse representation $\boldsymbol{s}_i$), so that RVM algorithm is priority especially when the underlying signals are fully sparse ($m \ll N$). In contrast to other algorithm like CoSaMP in which candidate atoms of $\boldsymbol{\Theta}$ once added are never removed, RVM algorithm allows to delete the atoms from model. The deleting operation provides a more efficient sparse representation and in the worst-case scenario, it's proved that RVM algorithm still outperforms the other algorithms mentioned before [18].

4. Overcomplete Dictionary

The performance of MT-BCS-SE is greatly affected by the sparsity of speech signals. Although most natural signals are not sparse themselves, they may be highly compressible and sparse in some transform domain. Many transform methods like Fourier transform (FT), discrete cosine transform (DCT), and wavelet transform (WT) played key roles in early development of sparse representation [16]. Sparse representation methods which are based on overcomplete dictionary become more and more popular recently. A widely used algorithm termed K-SVD algorithm was proposed by Micheal Elad [19] for learning overcomplete dictionaries. To obtain the sparse representations of the speech signals, we employ overcomplete dictionary to decompose the speech signals into sparse coefficients. Various overcomplete dictionaries are learned by K-SVD algorithm for MT-BCS-SE framework.

Considering a $n \times K$ (n denotes the dimension of the speech signals, K denotes the number of the speech samples) training set of speech samples $\boldsymbol{F} = [\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_K]$. According to [19], we have

$$(\boldsymbol{\Psi}, \boldsymbol{W}) = \arg\min_{\boldsymbol{\Psi}, \boldsymbol{W}} \sum_{k=1}^{K} (\rho\|\boldsymbol{w}_k\|_0 + \|\boldsymbol{\Psi}\boldsymbol{w}_k - \boldsymbol{f}_k\|_2^2) \tag{11}$$

where $\boldsymbol{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_K]$ is a set of sparse representations, $\boldsymbol{w}_i$ is of size $N \times 1$ (N denotes the length of dictionary)

and $\boldsymbol{W}$ is of size $N \times K$. $\boldsymbol{\Psi}$ is the dictionary which is updated in the iteration. The above expression (11) seeks to get a sparse representation per each of speech sample in $\boldsymbol{F}$. Parameter ρ controls the relative importance between error ($\|\boldsymbol{\Psi}\boldsymbol{w}_k - \boldsymbol{f}_k\|_2^2$) and sparsity ($\|\boldsymbol{w}_k\|_0$).

K-SVD algorithm adopts an iteration to solve the above problem. The iteration of K-SVD algorithm is consist of two stages: sparse coding stage and dictionary update stage. Sparse coding mainly transforms underlying speech signals into sparse representations which are submitted to $\boldsymbol{F} = \boldsymbol{\Psi}\boldsymbol{W}$. As mentioned before, although this can be implemented with any sparse decomposition, in our case it is implemented with BatchOMP [15]. This algorithm pre-computes the Gram matrix of the dictionary $\boldsymbol{\Psi}$, which allows for faster computation on each speech signal. In the dictionary update stage, K-SVD algorithm updates the dictionary $\boldsymbol{\Psi}$ according to those new sparse representations $\boldsymbol{W}$ obtained at sparse coding stage. Note that K-SVD algorithm only updates one column of dictionary at an iteration by applying a SVD operation on residual data so that this update can be done very optimally and computed only on the speech signals associated with this atom.

A brief and clear procedure of K-SVD algorithm can be described as follows: In sparse coding stage, assuming that dictionary $\boldsymbol{\Psi}^{(t)}$ (t means iteration times) is fixed, and then sparse representations $\boldsymbol{W}^{(t)}$ are calculated; while in dictionary update stage, these sparse representations $\boldsymbol{W}^{(t)}$ are used to update dictionary $\boldsymbol{\Psi}^{(t+1)}$. Along with updated dictionary $\boldsymbol{\Psi}^{(t+1)}$, a new sparse matrix $\boldsymbol{W}^{(t+1)}$ would be recalculated in the $t+1$ iteration. It should be emphasized that a wise selection of $\boldsymbol{\Psi}$ in dictionary initialization may fewer the numbers of training iterations. In our latter experiments, we choose an overcomplete DCT basis as the initial dictionary.

When K-SVD algorithm is adopted to the SE purpose, a crucial step is the selection of training set [19]. It's very hard to find a reasonably good dictionary that fits all signals, so [20] argued that the increase of the length of dictionary N generally improves the performance and an improved results obtained by training a dictionary based on the noisy signals itself. One reasonable explanation of this argument is that K-SVD algorithm has noise averaging built into it, by taking a large set of noisy training data and creating a small, relatively clean representation set. In the latter experiments, we designed various training sets for learning different dictionaries, and evaluated the influence of each dictionary on MT-BCS-SE algorithm.

5. Experiments

In this section, we try to prove the effectiveness of the proposed MT-BCS-SE in terms of various criteria. For all experiments, we select twelve different clean sentences (6 male and 6 female) and 192 corresponding noisy sentences from NOIZEUS corpus [21], which was developed to facilitate comparison of speech enhancement algorithms. All sentences were downsampled at 8 KHz. Twelve clean sentences were selected randomly from corpus, their ID are sp03, sp04, sp07, sp09, sp11, sp12, sp19, sp20, sp24, sp25, sp28 and sp29. All of the 192 noisy sentences are those clean sentences corrupted by four real-world noises including babble, car, train-station and street noise at four SNRs of 0 dB, 5 dB, 10 dB, and 15 dB ($12 \times 4 \times 4 = 192$). Noisy sentences were used to establish training sets (see Sect. 5.1).

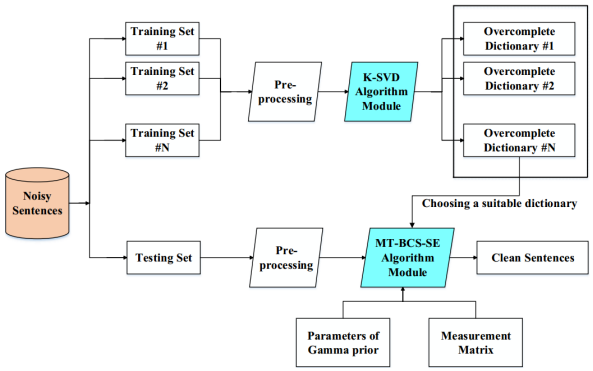


Fig. 2 The flow chart of the experiments.

Table 1 Default configuration for experiments.

Parameters	Default value
frame length n	64 or 128
overlap ratio ϕ	50%
tasks per time M_t	$M_t 3$
Gamma prior α	$a = 10/\sigma^2,\ b = 1$
Gamma prior $\boldsymbol{\beta}$	$c = 0,\ d = 0$
measurement dimension m	$m = 0.5 \times n$
dictionary size ($n \times N$)	(64×256) or (128×512)

The flow chart of our experiments is illustrated in Fig. 2. From Fig. 2, firstly, various overcomplete dictionaries are learned by K-SVD algorithm module. Secondly, when MT-BCS-SE was carried out, a most suitable dictionary was selected from those trained overcomplete dictionaries for transforming time-domain speech signals into sparse representations. At last, the recovered clean sentences were compared with original noisy sentences to evaluate the performance of the proposed algorithm. In this paper, signal-noise ratio (SNR), speech enhancement gain (SEG) and perceptual evaluation of speech quality (PESQ) were adopted as SE criteria. SEG is defined as the difference of SNR between noisy sentences and enhanced sentences, and it's given as:

$$\mathrm{SNR} = 10 \log \frac{\|\boldsymbol{x}\|_2^2}{\|\hat{\boldsymbol{x}}\|_2^2 - \|\boldsymbol{x}\|_2^2} \tag{12}$$

$$\mathrm{SEG} = \mathrm{SNR(enhanced)} - \mathrm{SNR(noisy)} \tag{13}$$

where $\boldsymbol{x}$ is original clean speech signals, and $\hat{\boldsymbol{x}}$ is the test speech signals. Using [14] as a reference for experimental protocol, a default configuration including the values of parameters as summarised in Table 1, was set up globally. It should be noted that Gamma prior parameter α is directly related with additive noise, so that a different prior or different values of a and b may achieve different SE performance.

5.1 Training Sets for Learning Overcomplete Dictionaries

For proving the effectiveness of the proposed MT-BCS-SE algorithm and the influence of overcomplete dictionary on the performance of MT-BCS-SE algorithm, we designed several training sets to learn various dictionaries via K-SVD algorithm. We follow certain rules to select noisy sentences for building the training sets.

- Noisy sentences that are at SNRs of 0 dB and 5 dB are selected to build low SNR training sets. Noisy sentences that are at SNRs of 10 dB and 15 dB are selected to build high SNR training sets.
- Low SNR dictionaries are learned from low SNR training sets and high SNR dictionaries are learned from high SNR training sets.
- We learn two kinds of dictionaries for our experiments, the general dictionary and the specific dictionary.
- For each type of noise, there are 48 ($12 \times 4 = 48$) noisy sentences. Half of them are used to build training sets and the others are used to build testing sets.
- The specific dictionaries are learned from the training sets in which all sentences are of the same type of noise. For example, a high SNR specific dictionary for babble noise is learned from the training set which includes 16 sentences: sp03_babble_sn10, sp04_babble_sn10, sp09_babble_sn10, sp12_babble_sn10, sp19_babble_sn10, sp20_babble_sn10, sp24_babble_sn10, sp29_babble_sn10, sp03_babble_sn15, sp04_babble_sn15, sp09_babble_sn15, sp12_babble_sn15, sp19_babble_sn15, sp20_babble_sn15, sp24_babble_sn15, sp29_babble_sn15.
- The general dictionaries are learned from the training sets which includes noisy sentences for all kinds of noise types.
- Every dictionaries have two different versions. The big size version of dictionary (BD) is of size 128×512 and the small size version of dictionary (SD) is of size 64×256.

For clarity, we named the learned overcomplete dictionaries. For example, we named the big size version of a high SNR specific dictionary for babble noise 'D-ba-high-bd', and the small size version of a high SNR specific dictionary for babble noise 'D-ba-high-sd'. Table 2 lists the names of all dictionaries that would be used in our experiments.

5.2 The Influence of Dictionary Size

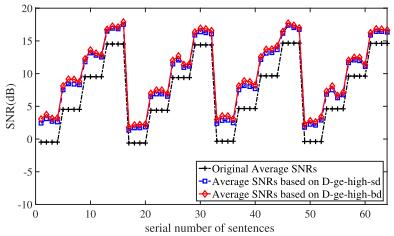
To test the influence of dictionary size on SE performance, two general dictionaries named D-ge-high-sd and D-ge-high-bd were used in the experiments. Figure 3 shows SNRs and SEGs reconstructed by MT-BCS-SE algorithm which employed D-ge-high-sd and D-ge-high-bd, respectively.

From Fig. 3 (a), we can tell that MT-BCS-SE algorithm achieves successful good SNRs with either D-ge-high-sd or D-ge-high-bd. A precise comparison of the SEGs is shown in Fig. 3 (b). The experimental results illustrate that average SEG by using D-ge-high-bd is bigger than that by using D-ge-high-sd. This is because the big size version of dictionary can remain and represent more information of original sentences than the small one, so that the performance of D-ge-high-bd on SE is better than the performance of D-ge-high-sd. Results agree with the conclusion that the increase of the length of dictionary N generally improves the performance in [20].

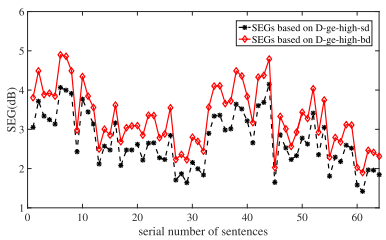
More detailed SEGs are arranged in Table 3. From Table 3, we can find that results differ from noise to noise and MT-BCS-SE algorithm achieves better performance when noisy sentences are at low SNRs, that's because noisy sentences at low SNRs contains much more additive noise than that at high SNRs.

Table 2 The names of overcomplete dictionaries.

Size	Noise \ SNR / Name	low-SNR	high-SNR
BD	BABBLE	D-ba-low-bd	D-ba-high-bd
	CAR	D-ca-low-bd	D-ca-high-bd
	STATION	D-st-low-bd	D-st-low-bd
	STREET	D-tr-low-bd	D-tr-high-bd
	General	D-ge-low-bd	D-ge-high-bd
SD	BABBLE	D-ba-low-sd	D-ba-high-sd
	CAR	D-ca-low-sd	D-ca-high-sd
	STATION	D-st-low-sd	D-st-high-sd
	STREET	D-tr-low-sd	D-tr-high-sd
	General	D-ge-low-sd	D-ge-high-sd



(a) Performance comparison in terms of SNR



(b) Performance comparison between BD and SD

Fig. 3 Performance of MT-BCS-SE.

5.3 The Influence of Dictionary Type

Here we carried out more experiments to test the influence of dictionary type on MT-BCS-SE. By adopting specific dictionaries to deal with the corresponding testing sets, the proposed algorithm can obtain better results. For example, D-ca-high and D-ca-low are used to recover the testing set that

Table 3 SEGs (dB) of enhanced sentences recovered with different dictionaries.

	SNR / Noise	0 dB	5 dB	10 dB	15dB
BD	BABBLE	3.685	3.532	3.297	2.643
	CAR	3.647	3.567	3.186	2.656
	STATION	3.718	3.564	3.238	2.761
	STREET	4.023	3.768	3.327	2.902
SD	BABBLE	3.153	2.950	2.745	2.248
	CAR	3.147	2.922	2.663	2.196
	STATION	3.208	3.042	2.774	2.302
	STREET	3.256	3.098	2.811	2.547

Table 4 SEGs (dB) of enhanced sentences recovered with various dictionaries.

Noise	SNR / Dicts	0 dB	5 dB	10 dB	15 dB
BABBLE	D-ba-low-bd	3.542	3.5404	3.157	2.502
	D-ge-low-bd	3.387	3.396	3.072	2.392
	D-ba-high-bd	**4.002**	**3.922**	**3.345**	**2.739**
	D-ge-high-bd	3.685	3.532	3.297	2.643
CAR	D-ca-low-bd	3.287	3.304	2.854	2.360
	D-ge-low-bd	3.3127	3.198	2.765	2.279
	D-ca-high-bd	**3.985**	**3.884**	**3.243**	**2.678**
	D-ge-high-bd	3.647	3.567	3.186	2.656
STATION	D-st-low-bd	3.592	3.419	3.127	2.604
	D-ge-low-bd	3.476	3.325	3.056	2.538
	D-st-high-bd	**4.082**	**3.896**	**3.350**	**2.856**
	D-ge-high-bd	3.718	3.564	3.238	2.761
STREET	D-tr-low-bd	3.821	3.543	3.175	2.656
	D-ge-low-bd	3.632	3.418	2.915	2.544
	D-tr-high-bd	**4.122**	**3.973**	**3.375**	**3.054**
	D-ge-high-bd	4.023	3.768	3.327	2.902

was consisted of noisy sentences of car noise. 10 dictionaries named D-ge-low, D-ba-low, D-ca-low, D-st-low, D-tr-low, D-ge-high, D-ba-high, D-ca-high, D-st-high and D-tr-high were adopted in this experiment. Experimental results are shown in Table 4.

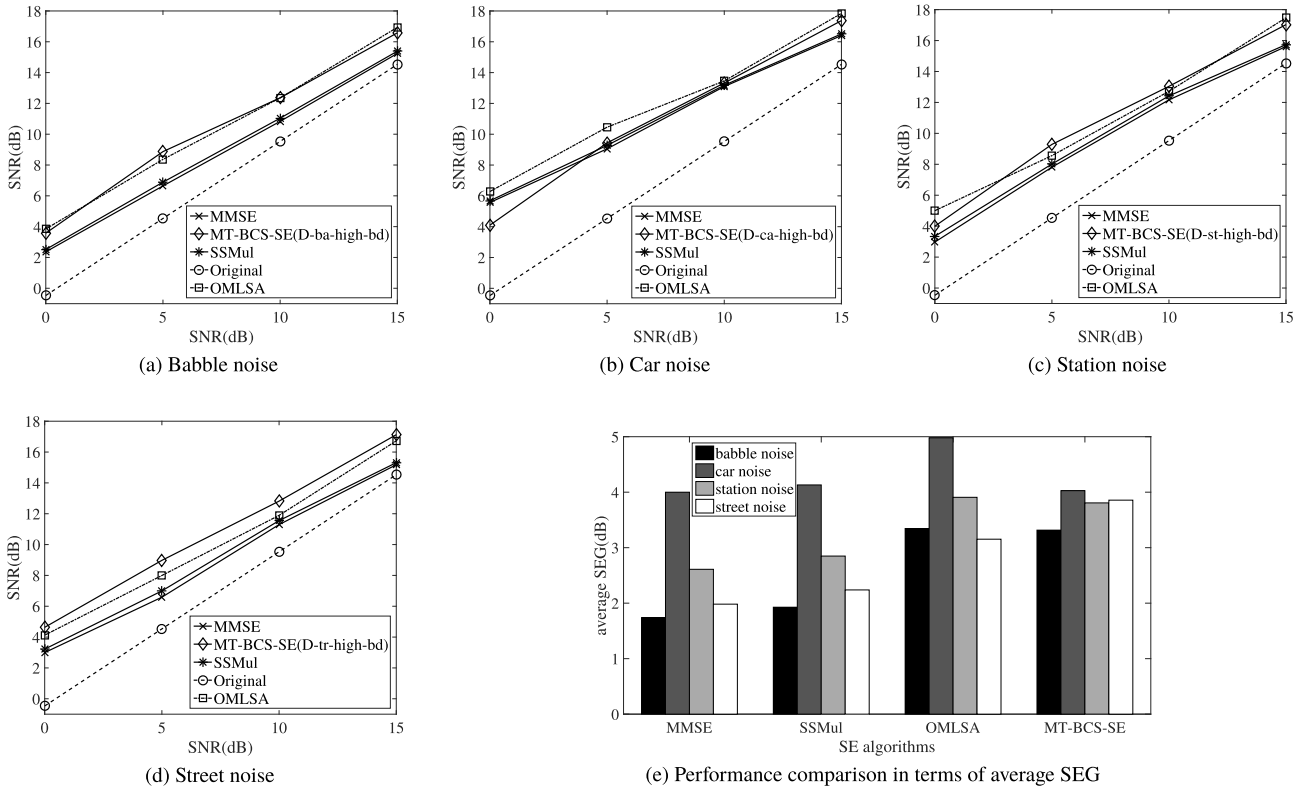
Some observations can be made for Table 4. Firstly, a high SNR dictionary always outperforms the low SNR dictionary. That's because low SNR dictionary remain overmuch information about noise, and those information would reduce the SE performance. Secondly, a high SNR general dictionary has better performance than a low SNR specific dictionary, for example, D-ge-high-bd performs better than D-tr-low-bd in dealing the testing set that was consisted of noisy sentences of street noise. Because D-tr-low-bd was learned from a low SNR training set, so that much interference information was also brought into this low SNR specific dictionary. Thirdly, the best performance is obtained by using the corresponding big size version of high SNR specific dictionary for each testing set (see the bold values in Table 4). When dealing with low SNR noisy sentences, it's interesting that specific dictionaries improved the performance more obviously. An empirical reason is that because specific dictionaries which are learned from high SNR training sets, so that dealing with high SNR testing sets could not have as much improvement as dealing with low SNR testing sets.

5.4 Comparison with Other SE Algorithms

In this section, we further discuss that MT-BCS-SE algorithm also holds certain advantages over some traditional successful SE algorithms such as multi-band spectral subtraction (SSMul) [22], MMSE [23] and optimally-modified log-spectral amplitude (OMLSA) [6]. We use SNR, SEG and PESQ as our criteria to evaluate the SE performance. In order to obtain the best performance, MT-BCS-SE algorithm employed the corresponding big size version of high SNR specific dictionaries as its overcomplete dictionaries. For example, when dealing with noisy sentences of car noise, D-ca-high-bd was adopted by the proposed method.

Experimental results are shown in Fig. 4, several brief summaries can be obtained accordingly:

1. From Fig. 4 (d), MT-BCS-SE outperforms all other algorithms for street noise.
2. As shown in Fig. 4 (a) and Fig. 4 (c), MT-BCS-SE obtains roughly similar SNRs with OMLSA when handling babble and station noise. However, it's able to see that the average SEG of OMLSA varies from the types of noise (see Fig. 4 (e)).
3. As shown in Fig. 4 (e), MMSE and SSMul also fluctuated in different situations. We can see that though both of them do not have the best results, the performance of these algorithms for car noise is much better than that for other noises like babble noise.
4. When comparing with other SE algorithms, MT-BCS-SE does not show much advantages for car noise. That's because car noise is an non-Gaussian noise and MT-BCS-SE is more suitable in dealing with white noises, so that MT-BCS-SE is not mediocre at denoising car noise (see Sect. 3.1).
5. Further more, though MT-BCS-SE did not perform well comparing with other algorithms for car noise, there is little difference between the SEGs of MT-BCS-SE for car noise and that for other types of noise (see Fig. 4 and Table 4).
6. The performance of traditional SE algorithms varies greatly from different noise testing sets, however, MT-BCS-SE always obtained the best or second best results in most situations (see Fig. 4 (e)).
7. From Fig. 4 (e), Table 4 and preceding observations, we are able to say that traditional SE algorithms have fluctuant results when they deal with different types of noisy speech signals while the performance of MT-BCS-SE is relatively stable. That's because MT-BCS-SE can use the most suitable overcomplete dictionary for different types of noise to achieve the stability of performance.
8. Except for car noise, MT-BCS-SE performed very well and were competitive or even superior to traditional SE algorithms. What's more, the performance of MT-BCS-SE does not vary with the type of noise and is more stable than that of these traditional SE algorithms.



(a) Babble noise (b) Car noise (c) Station noise

(d) Street noise (e) Performance comparison in terms of average SEG

Fig. 4 Performance comparison of MT-BCS-SE with other SE algorithms.

Table 5 Performance comparison of MT-BCS-SE with other algorithms in terms of PESQ.

PESQ SE / Noise	MT-BCS-SE	OMLSA	SSMul	MMSE
BABBLE	2.5872	2.6053	2.4726	2.4415
CAR	2.6734	2.6856	2.6770	2.6603
STATION	2.7378	2.7146	2.5310	2.5291
STREET	2.8242	2.7842	2.4626	2.4206

Perceptual Evaluation of Speech Quality (PESQ) which is a worldwide applied industry standard for objective voice quality testing, can be applied to provide an end-to-end (E2E) quality assessment. Besides SNR, we use PESQ to evaluate speech quality from a perceptual perspective. PESQ is a full-reference algorithm and analyzes the speech signal sample-by-sample after a temporal alignment of corresponding excerpts of reference and test signal. The enhanced speech signals at SNR of 10 dB which were enhanced by MT-BCS-SE, OMLSA, SSMul and MMSE are scored by PESQ. PESQ results principally model mean opinion scores (MOS) that cover a scale from 1 (bad) to 5 (excellent). The average score of each SE algorithm is listed in Table 5. Experimental results in Table 5 indicate that MT-BCS-SE algorithm performs better than other algorithms obviously for street noise. MT-BCS-SE and OMLSA almost had little different in scores, that's because OMLSA is based on the statistical modelling SE methods and MT-BCS-SE also assumed some kind of noise prior. The results based on PESQ almost agree with that obtained based on SNR. Hence, a brief summary can be made that MT-BCS-SE has a comparable performance with OMLSA and even has certain advantages on OMLSA when dealing with street noise.

6. Conclusion

In this paper, we have proposed a speech enhancement method termed MT-BCS-SE algorithm. The sparsity of speech signals and the dependence information among measurements are utilized by our proposed method. We used overcomplete dictionaries which are learned via K-SVD algorithm to transform speech signals into sparse representations. Numbers of experiments were conducted on NOIZEUS corpus to evaluate the performance of the proposed algorithm. SNR, SEG and PESQ were adopted as criteria. For overcomplete dictionaries, experimental results show that the increase of the length of dictionary and well selected specific dictionaries generally improve the SE performance of MT-BCS-SE. As compared with other traditional SE algorithms, MT-BCS-SE has better performance in most situations. More importantly, the observations about experimental results indicated MT-BCS-SE is able to have stable performance for multiple types of noisy speech signals. We can conclude that MT-BCS-SE has advantages over traditional SE algorithms in not only speech quality but also stability of performance.

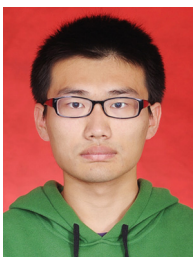
Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) under Grants No. 61271349, 61371147 and 11433002. Thanks to Lianqiang Li and Kevin Anderson for helping with English language editing and checking.

References

[1] J.-C. Junqua and J.-P. Haton, "Robustness in automatic speech recognition: Fundamentals and applications," Engineering and Computer Science, Kluwer Academic Publishers, 1996.
[2] P.C. Loizou, Speech enhancement: Theory and practice, 2nd ed., CRC Press, 2007.
[3] F. Asano, S. Hayamizu, T. Yamada, and S. Nakamura, "Speech enhancement based on the subspace method," IEEE Trans. Speech Audio Process., vol.8, no.5, pp.497–507, 2000.
[4] G.M. Chen, L. Zhao, and C.R. Zou, "Speech enhancement based on subspace method for narrowband noise," J. Applied Sciences, 2007.
[5] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," IEEE Trans. Acoust. Speech Signal Process., vol.32, no.6, pp.1109–1121, 1984.
[6] T.D. Tran, Q.C. Nguyen, and D.K. Nguyen, "Speech enhancement using modified IMCRA and OMLSA methods," 2010 Third International Conference on Communications and Electronics (ICCE), pp.195–200, 2010.
[7] T. Inoue, H. Saruwatari, Y. Takahashi, K. Shikano, and K. Kondo, "Theoretical analysis of musical noise in generalized spectral subtraction based on higher order statistics," J. Japan Society for the Study of Toxemia of Pregnancy, vol.19, no.6, pp.1770–1779, 2011.
[8] D.L. Donoho, "Compressed sensing," IEEE Trans. Inf. Theory, vol.52, no.4, pp.1289–1306, 2006.
[9] E.J. Candes, J.K. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," IEEE Trans. Inf. Theory, vol.52, no.2, pp.489–509, 2006.
[10] K. Hayashi, M. Nagahara, and T. Tanaka, "A user's guide to compressed sensing for communications systems," IEICE Trans. Commun., vol.E96-B, no.3, pp.685–712, March 2013.
[11] J. Jiang, X. Wu, X. He, and P. Karn, "Measuring crowd collectiveness via compressive sensing," IEICE Trans. Fundamentals, vol.E98-A, no.11, pp.2263–2266, Nov. 2015.
[12] D. Wu, W.-P. Zhu, and M.N.S. Swamy, "On sparsity issues in compressive sensing based speech enhancement," 2012 IEEE International Symposium on Circuits and Systems (ISCAS), pp.285–288, 2012.
[13] D. Wu, W.-P. Zhu, and M.N.S. Swamy, "Compressive sensing-based speech enhancement in non-sparse noisy environments," IET Signal Processing, vol.7, no.5, pp.450–457, 2013.
[14] D. Wu, W.-P. Zhu, and M.N.S. Swamy, "The theory of compressive sensing matching pursuit considering time-domain noise with application to speech enhencement," IEEE/ACM Trans. Audio, Speech, Language Process., vol.22, no.3, pp.682–696, March 2014.
[15] R. Rubinstein, M. Zibulevsky, and M. Elad, "Double sparsity: Learning sparse dictionaries for sparse signal approximation," IEEE Trans. Signal Process., vol.58, no.3, pp.1553–1564, 2010.
[16] S. Ji, D. Dunson, and L. Carin, "Multitask compressive sensing," IEEE Trans. Signal Process., vol.57, no.1, pp.92–106, 2008.
[17] S. Ji, Y. Xue, and L. Carin, "Bayesian compressive sensing," IEEE Trans. Signal Process., vol.56, no.6, pp.2346–2356, 2008.
[18] A. Faul and J.J.T. Avenuse, "Fast marginal likelihood maximisation for sparse Bayesian models," Proc. Ninth International Workshop on Artificial Intelligence and Statistics, pp.3–6, 2003.
[19] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," IEEE Trans. Image Process., vol.15, no.12, pp.3736–3745, 2006.
[20] M. Protter and M. Elad, "Image sequence denoising via sparse and redundant representations," IEEE Trans. Image Process., vol.18, no.1, pp.27–35, 2009.
[21] Y. Hu and P.C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," Speech Commun., vol.49, no.7, pp.588–601, 2007.
[22] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp.IV–4164, 2002.
[23] I. Cohen, "Speech enhancement using a noncausal a priori SNR estimator," IEEE Signal Process. Lett., vol.11, no.9, pp.725–728, 2004.

Hanxu You received the M.E. degree in information engineering in 2012 from Tongji University, China. He is a Ph.D. candidate in Shanghai Jiao Tong University, China. He works in Department of Electronic Engineering, Shanghai Jiao Tong University, China. His research interests include compressive sensing, audio and speech signal processing, image processing, etc.



Zhixian Ma received the B.S. degree in telecommunication engineering in 2014 from Xidian University, Xi'an, China. He is a Ph.D. candidate in Shanghai Jiao Tong University, China, he works in Department of Electronic Engineering, Shanghai Jiao Tong University, China. His research interests include image processing,signal processing, etc.

Wei Li received the B.E. degree in information engineering in 2010 from Shanghai Jiao Tong University, China. He is a Ph.D. candidate in Shanghai Jiao Tong University, China, he worked in Department of Electronic Engineering, Shanghai Jiao Tong University, China. His research interests include speaker recognition, speech recognition, computer vision, etc.

Jie Zhu received the Ph.D. degree in information system from Shanghai Jiao Tong University, China. Since 1999, he works as a professor in Department of Electronic Engineering, Shanghai Jiao Tong University, China. He is chairman of IEICE Shanghai section. His research interests include speech signal processing, multimedia system, etc.