Journal of Zhejiang University-SCIENCE C (Computers & Electronics)
ISSN 1869-1951 (Print); ISSN 1869-196X (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Speech enhancement with a GSC-like structure employing sparse coding*

Li-chun YANG[1,2], Yun-tao QIAN[‡1]

([1]*College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China*)

([2]*Intelligent Control Research Institute, Zhejiang Wanli University, Ningbo 315101, China*)

E-mail: lichun_y@126.com; ytqian@zju.edu.cn

Received Mar. 9, 2014; Revision accepted Aug. 5, 2014; Crosschecked Nov. 9, 2014

Abstract: Speech communication is often influenced by various types of interfering signals. To improve the quality of the desired signal, a generalized sidelobe canceller (GSC), which uses a reference signal to estimate the interfering signal, is attracting attention of researchers. However, the interference suppression of GSC is limited since a little residual desired signal leaks into the reference signal. To overcome this problem, we use sparse coding to suppress the residual desired signal while preserving the reference signal. Sparse coding with the learned dictionary is usually used to reconstruct the desired signal. As the training samples of a desired signal for dictionary learning are not observable in the real environment, the reconstructed desired signal may contain a lot of residual interfering signal. In contrast, the training samples of the interfering signal during the absence of the desired signal for interferer dictionary learning can be achieved through voice activity detection (VAD). Since the reference signal of an interfering signal is coherent to the interferer dictionary, it can be well restructured by sparse coding, while the residual desired signal will be removed. The performance of GSC will be improved since the estimate of the interfering signal with the proposed reference signal is more accurate than ever. Simulation and experiments on a real acoustic environment show that our proposed method is effective in suppressing interfering signals.

Key words: Generalized sidelobe canceller, Speech enhancement, Voice activity detection, Dictionary learning, Sparse coding

doi:10.1631/jzus.C1400085 **Document code:** A **CLC number:** TN912.35

1 Introduction

Speech communication applications like mobile phone, teleconferencing, and network communication are often corrupted by an interfering signal, such as music and babble, which will cause severe degradation of the intelligibility and fidelity of the desired signal. The aim of speech enhancement is to suppress the interfering signal while preserving the desired signal. As an interference is usually a non-stationary signal, speech enhancement by using the interfering signal of segments of the desired signal inactivity to estimate the interference of segments of the desired signal activity will be limited.

To deal with the suppression of a non-stationary interfering signal, a microphone arrays based generalized sidelobe canceller (GSC) (Griffiths and Jim, 1982) using an adaptive filter to estimate the interference can work well in theory. The reference signal used in the adaptive filter can be achieved by a blocking matrix. GSC is usually using time delay compensation to block the desired signal. Thus, the position of the desired source should be estimated. Since the error in time difference of arrival (TDOA) exists in the real acoustic environment, a little desired signal

‡ Corresponding author

* Project supported by the National Basic Research Program (973) of China (No. 2012CB316400) and the National Natural Science Foundation of China (No. 61171151)

ORCID: Li-chun YANG, http://orcid.org/0000-0003-1651-798X; Yun-tao QIAN, http://orcid.org/0000-0002-7418-5891

©Zhejiang University and Springer-Verlag Berlin Heidelberg 2014

will leak into the reference signal. To avoid distortion of the desired signal, we should stop updating the weight of the adaptive filter when the desired signal exists. The adaptive blocking matrix (ABM) in the time domain (Hoshuyama *et al.*, 1999) or frequency domain (Herbordt and Kellermann, 2001) has been used to deal with the leakage of the desired signal. However, the adaptive blocking matrix cannot reduce the leakage in the real acoustic environment due to reverberation.

In the reverberation environment, the blocking matrix with acoustic impulse responses (AIRs) can be more effective in blocking the desired signal than that uses delay and attenuation of the desired signal only. Since the desired source AIRs are unknown in practice, the transfer function ratios (TFRs) (Gannot *et al.*, 2001; Talmon *et al.*, 2009; Krueger *et al.*, 2011) were introduced for the blocking matrix. As the impulse response may reach several thousand taps in the reverberant environment, the TFR estimation is not very accurate, which will lead to the leakage of the desired signal into the reference signal.

In recent years, sparse coding with the dictionary of the desired signal has been introduced into the speech enhancement area, which can avoid direct estimation of the interfering signal. The dictionary is a collection of the finite basis functions that are coherent to the structured component of a signal (Rebollo-Neira, 2004; Gribonval and Schnass, 2008). As non-random signals (speech, music, babble, etc.) contain structured components (Plumbley *et al.*, 2010; Sigg *et al.*, 2012) and the signal structure is relatively stable, we can use a dictionary to encode the corresponding signal, while the other signal cannot be represented by the same dictionary. So, the interference is avoided by sparse coding. The dictionary includes a predefined dictionary and the learned dictionary. The predefined dictionary, such as wavelets, Fourier transform, and discrete cosine transform, is a general signal dictionary in which structured components of different signals are difficult to distinguish; thus, it may not work well in sparse coding.

On the other hand, the learned dictionary is coherent to the structured component of a specified signal while incoherent or less coherent to the structured components of other signals. So, the learned dictionary can work well in sparse coding (Elad and Aharon, 2006). The dictionary learning algorithm (Aharon and Elad, 2006; Engan *et al.*, 2007; Mairal *et al.*, 2010; Skretting and Engan, 2010) uses the training samples of a specific signal to obtain the dictionary matrix. Each column of the dictionary matrix is a basis vector (also called 'atom'). A signal can be well approximated by the linear combination of a few atoms of the learned dictionary, while other signals cannot be represented by the same atoms. So, the signal can be sparsely reconstructed effectively, while other signals will be suppressed (Gemmeke and Cranen, 2009; He *et al.*, 2012).

Thus, one of the most important factors for sparse coding is to build a learned dictionary of the desired signal. However, in a real acoustic environment, it is difficult to obtain the training samples of a desired signal for dictionary learning, so interference suppression is limited. As the desired signal has some pauses in practice, the interfering signal used for dictionary learning can be achieved during the segments of the desired signal inactivity and then the learned dictionary of interfering signal can be obtained. As a signal dictionary is relatively stable, the interferer dictionaries of adjacent segments of the desired signal activity are almost the same.

Inspired by this, we present a GSC-like method based on sparse coding for speech enhancement. To obtain the interferer dictionary, the training samples for dictionary learning are achieved by the voice activity detection (VAD) algorithm. At the same time, the blocking matrix based on TFRs is used to block the desired signal. Then the residual desired signal that leaks into the reference signal is further suppressed by sparse coding. Finally, the signal of GSC output is achieved by an adaptive filter algorithm to estimate the original interfering signal.

We use online dictionary learning (ODL) (Mairal *et al.*, 2010) to obtain the interferer dictionary. To ensure that the dictionary meets the varying structured components of a signal, dictionary learning should be peformed at each segment of the desired signal pauses. On the other hand, to achieve the training samples for interferer dictionary learning, we suppose that the first frame signal does not contain the desired signal.

2 Generalized sidelobe canceller

GSC, first proposed by Griffiths and Jim (1982), is an important speech enhancement method. Fig. 1

shows that GSC consists of three blocks. The upper branch is a fixed beamformer (FBF) block, which is usually achieved by the delay-and-sum beamformer (DSB). The aim of the fixed beamformer is to form an undistorted desired direction signal while suppressing other direction signals. The blocking matrix (BM) block and adaptive noise canceller (ANC) block lie in the lower branch. The blocking matrix is used to block the desired signal to form a reference signal, which is used in ANC to estimate the interfering signal. The adaptive noise canceller is an unstrained adaptive algorithm to suppress the remaining interfering signal of the fixed beamformer output.

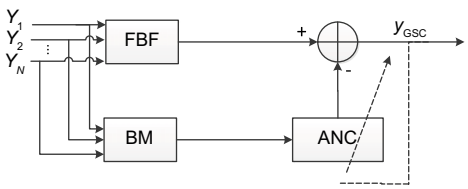

Fig. 1 Structure of the generalized sidelobe canceller (GSC)

2.1 Signal model

Suppose the interfering signal is uncorrelated to the desired signal. The received signal of each microphone should be a convolution of impulse response functions of the array element and the desired signal. The impulse response is formed by the desired source propagation attenuation process, which leads to a large number of echoes due to reflections of the wavefront from walls, ceilings, floors, and other objects in the room. Considering a linear array with M omnidirectional microphones, the received signal of the ith microphone of the array can be represented as

$$y_i(t) = a_i * s(t) + n_i(t), \quad i = 1, 2, \cdots, M, \tag{1}$$

where a_i is a transfer function from the desired speech source to the ith microphone, $s(t)$ is a desired signal, $n_i(t)$ is an interfering signal, and '*' denotes the convolution operator. Applying a short-time Fourier transform (STFT) to both sides, Eq. (1) can be expressed in the frequency domain as

$$\boldsymbol{y}_i(\omega, k) = \boldsymbol{a}_i \boldsymbol{s}(\omega, k) + \boldsymbol{n}_i(\omega, k), \quad i = 1, 2, \cdots, M, \tag{2}$$

where ω is the frequency bin index and k is the frame index.

2.2 Fixed beamformer

The fixed beamformer of GSC is designed to enhance the desired direction gain and suppress the other direction signal. To meet the requirements, the received desired signal of each microphone should be the same at the same time. Both sides of Eq. (2) are multiplied by $\boldsymbol{a}_1/\boldsymbol{a}_i$ $(i = 1, 2, \ldots, M)$:

$$\frac{\boldsymbol{a}_1}{\boldsymbol{a}_i}\boldsymbol{y}_i(\omega, k) = \boldsymbol{a}_1 \boldsymbol{s}(\omega, k) + \frac{\boldsymbol{a}_1}{\boldsymbol{a}_i}\boldsymbol{n}_i(\omega, k), \ i = 1, 2, \ldots, M, \tag{3}$$

where $\boldsymbol{a}_1/\boldsymbol{a}_i$ $(i = 1, 2, \ldots, M)$ is a transfer function ratio of different microphones to the first microphone. We suppose that the statistics of the interfering signal is slowly changing compared with the statistics of the desired signal. The transfer function ratios can be approximated by (Gannot *et al.*, 2001)

$$\begin{aligned}\frac{\boldsymbol{a}_i}{\boldsymbol{a}_1}(\omega, k) \approx & \frac{\langle \boldsymbol{p}_{y_1 y_1}(\omega, k) \boldsymbol{p}_{y_i y_1}(\omega, k)\rangle}{\langle \boldsymbol{p}^2_{y_1 y_1}(\omega, k)\rangle - \langle \boldsymbol{p}_{y_1 y_1}(\omega, k)\rangle^2} \\ & - \frac{\langle \boldsymbol{p}_{y_1 y_1}(\omega, k)\rangle \langle \boldsymbol{p}_{y_m y_1}(\omega, k)\rangle}{\langle \boldsymbol{p}^2_{y_1 y_1}(\omega, k)\rangle - \langle \boldsymbol{p}_{y_1 y_1}(\omega, k)\rangle^2}, \\ & i = 1, 2, \cdots, M,\end{aligned} \tag{4}$$

where $\boldsymbol{p}_{y_m y_n}(\cdot)$ denotes the cross power spectral density (CPSD) function of two signals y_m and y_n, $\boldsymbol{p}_{y_m y_n}(\cdot)$ is the power spectral density (PSD) function of a signal when $m = n$, and $\langle\cdot\rangle$ represents the average operation.

The fixed beamformer output is

$$\boldsymbol{y}_{\mathrm{FBF}}(\omega, k) = \boldsymbol{a}_1 \boldsymbol{s}(\omega, k) + \frac{1}{M}\sum_{i=1}^{M} \frac{\boldsymbol{a}_1}{\boldsymbol{a}_i}\boldsymbol{n}_i(\omega, k). \tag{5}$$

2.3 Blocking matrix

From Eq. (2), we can find that the difference in the desired signal from each microphone is an impulse response. So, the blocking matrix $\boldsymbol{B}$ can be constructed by the transfer function ratio as

$$\boldsymbol{B} = \begin{bmatrix} -\boldsymbol{a}_2/\boldsymbol{a}_1 & \boldsymbol{I} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ -\boldsymbol{a}_3/\boldsymbol{a}_1 & \boldsymbol{0} & \boldsymbol{I} & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \vdots & & \vdots \\ -\boldsymbol{a}_M/\boldsymbol{a}_1 & \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{I} \end{bmatrix}. \tag{6}$$

Thus, the reference signal $\boldsymbol{n}_{\mathrm{ref}}$ in the frequency domain is

$$\boldsymbol{n}_{\mathrm{ref}} = \boldsymbol{y}\boldsymbol{B}^{\mathrm{T}} = \sum_{i=2}^{M} \boldsymbol{n}_i(\omega, k) - \sum_{i=2}^{M} \frac{\boldsymbol{a}_i}{\boldsymbol{a}_1}\boldsymbol{n}_1(\omega, k), \tag{7}$$

where $\boldsymbol{y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_M]$, and $\boldsymbol{a}_i/\boldsymbol{a}_1$ is a transfer function ratio. The amount of the desired signal leaked into $\boldsymbol{n}_{\mathrm{ref}}$ depends on the estimated accuracy of the transfer function ratios. As mentioned before, it is difficult to estimate accuracy since the duration of the impulse response is very long in the reverberant environment.

2.4 Adaptive noise canceller

The adaptive noise canceller uses the reference signal to estimate the interfering signal of the fixed beamformer output. The main concerns of the adaptive noise canceller are computational complexity and convergence. Normalized least mean square (NLMS) based algorithms in the time domain are widely used due to their low computational complexity and fine convergence. However, in the low signal-to-noise ratio (SNR) or reverberation environment, the convergence performance of NLMS is very poor. Thus, we propose an NLMS algorithm in the frequency domain, which has better convergence performance and lower computational complexity than its counterpart (Avargel and Cohen, 2008).

3 GSC with sparse coding

To reduce the residual desired signal component in the reference signal, we use sparse coding for GSC to improve the reference signal. Fig. 2 shows the proposed structure of GSC.

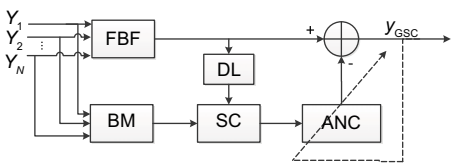

Fig. 2 Structure of the proposed method

Compared with traditional GSC structure, the proposed structure includes the dictionary learning (DL) block to obtain interferer dictionary and a sparse coding (SC) block to suppress the residual desired signal that leaks into the reference signal. Then the weight of an adaptive filter can track the interfering signal changes in segments of speech activity to achieve better speech enhancement than that using the blocking matrix only.

3.1 Dictionary learning

The aim of dictionary learning is to obtain a signal dictionary that is coherent to its structured component, and the dictionary is incoherent or of little coherence to the structured components of other signals. For this purpose, the training samples of dictionary learning should be a part of the signal or coherent in itself. Meanwhile, the training samples do not contain any other signal. In a real communication environment, the desired signal dictionary is difficult to achieve since the training samples of a clean desired signal for dictionary learning are never directly observable. As the interfering signal can be obtained in the segments of the desired signal inactivity, the interferer dictionary that can be used to suppress the leakage of the desired signal into the reference signal by sparse coding is relatively easy to obtain. Obviously, a reference signal with little desired speech leakage is effective in improving speech enhancement.

In addition, in order to use a part of the atoms of a dictionary to code the interfering signal, the dictionary for sparse coding should be an overcomplete dictionary (or called a 'redundant dictionary'). That is to say, the number of atoms of the dictionary is larger than the length of the signal frame. The desired signal that leaks into the reference signal cannot be represented by a few atoms in the interferer dictionary. Then the residual desired signal will be suppressed in the reconstructed signal.

As shown in Fig. 2, the interfering signal vector for dictionary learning, which comes from the segments of the desired signal inactivity of the FBF output, can be expressed as $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]$, where n is the length of a signal frame. If a dictionary is known, the interfering signal can be reconstructed as

$$\boldsymbol{x} \approx \boldsymbol{D}\boldsymbol{w}_l, \tag{8}$$

where $\boldsymbol{w}_l$ is a vector of the dictionary coefficient with m elements, denoting the weights of each atom in sparse coding. Then Eq. (8) should meet the following constraint:

$$\arg\min_{\boldsymbol{D},\boldsymbol{w}} ||\boldsymbol{x} - \boldsymbol{D}\boldsymbol{w}_l||_{\mathrm{F}}^2, \tag{9}$$

where $||\cdot||_{\mathrm{F}}$ denotes the Frobenius norm (usually the ℓ_2 norm), and dictionary $\boldsymbol{D}$ is an $n \times m$ matrix.

There are some optimization methods for constraint (9), such as the method of optimized direc-

tions (MOD), iterative least squares dictionary learning algorithm (ILS-DLA) (Engan *et al.*, 2007), K-SVD (Aharon and Elad, 2006), ODL (Mairal *et al.*, 2010), and recursive least squares dictionary learning algorithm (RLS-DLA) (Skretting and Engan, 2010). Because the interferer dictionary should be dynamically achieved in real time, we need a dictionary learning with low computation complexity to dynamically process a new vector of training samples. As ODL can process the new training vector continuously to realize dictionary update with low computation complexity, in this study we use ODL to obtain the interferer dictionary, in order to meet real-time requirements.

For an overcomplete dictionary, we let $m > n$ in constraint (9) to ensure that the dictionary atoms are redundant. To obtain a sparse solution for dictionary coefficient $\boldsymbol{w}_l$, we need to use a sparse constraint on $\boldsymbol{w}_l$ in constraint (9). As ℓ_1 norm regularization yields a sparse solution, constraint (9) can be further rewritten as

$$\arg\min_{\boldsymbol{D},\boldsymbol{w}_l}\left(||\boldsymbol{x}-\boldsymbol{D}\boldsymbol{w}_l||_2^2+\lambda'||\boldsymbol{w}_l||_1\right), \tag{10}$$

where λ' is the regularization constraint coefficient. For dictionary matrix $\boldsymbol{D}=[\boldsymbol{d}_1,\boldsymbol{d}_2,\cdots,\boldsymbol{d}_m]$ and its coefficient vector $\boldsymbol{w}_l=[w_{l1},w_{l2},\cdots,w_{lm}]$, we can rewrite expression (10) as

$$\min_{\substack{\boldsymbol{D}\in\mathbb{R}^{n\times m}\\ \boldsymbol{w}_l\in\mathbb{R}^{m\times 1}}}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{2}\|\boldsymbol{D}w_{li}-x_i\|_2^2+\lambda'\|w_{li}\|_1\right)$$
$$\text{s.t.}\quad \forall j=1,2,\ldots,k,\quad \boldsymbol{d}_j^{\mathrm{T}}\boldsymbol{d}_j\le 1, \tag{11}$$

where $\boldsymbol{d}_j^{\mathrm{T}}\boldsymbol{d}_j\le 1$ is a constraint to avoid the dictionary coefficient being too small. We can obtain the sparse solution via applying the ℓ_1 norm constraint on w_{li}. $\boldsymbol{w}_l$ is convex when $\boldsymbol{D}$ is fixed, and vice versa. Therefore, the optimization algorithm is an alternating iterative method for the dictionary and its coefficient. In each iteration, we fix the dictionary $\boldsymbol{D}$ to optimize the dictionary coefficient $\boldsymbol{w}_l$, and then fix $\boldsymbol{w}_l$ to update $\boldsymbol{D}$. More details of the ODL method can be found in Mairal *et al.* (2010).

3.2 Sparse coding

Since the dictionary of a desired signal is difficult to achieve directly, we do not use sparse coding for speech enhancement, but use sparse coding with the interferer dictionary to reconstruct the reference signal. As the reconstructed reference signal contains little residual desired signal, the weight of the adaptive filter can track the interfering signal changes in the segments of speech activity to improve speech enhancement.

As the interfering signal component of the FBF output is coherent to the reference signal component, the interferer dictionary is also coherent to the structured component of the reference signal. For an overcomplete interferer dictionary, a few atoms of it can be used to correct the code of the reference signal, and the other signals will be suppressed because they cannot be represented by the same atoms in the dictionary.

Suppose the frame length of a signal is m and define $\boldsymbol{z}$ as a vector with m samples of the reference signal. The corresponding coefficient vector $\boldsymbol{w}$ and the reference signal $\boldsymbol{z}$ in the interferer dictionary satisfy

$$\hat{\boldsymbol{w}}=\arg\min_{\boldsymbol{w}}\left(\frac{1}{2}||\boldsymbol{D}\boldsymbol{w}-\boldsymbol{z}||_2^2+\lambda||\boldsymbol{w}||_1\right), \tag{12}$$

where $\boldsymbol{D}$ is an overcomplete dictionary composed of basis vectors of the interfering signal, and λ is a regularization parameter which controls the degree of sparsity in vector $\boldsymbol{w}$. The second item of Eq. (12) is ℓ_1 norm for sparsity constraints on the coefficient vector $\boldsymbol{w}$. In Eq. (12), as $\boldsymbol{D}$ is an overcomplete dictionary, the optimal solution $\hat{\boldsymbol{w}}$ which uses the ℓ_1 norm constraint can ensure that $\hat{\boldsymbol{w}}$ is sparse and can maximize the recovery of the corresponding signal. The output signal of sparse reconstruction $\tilde{\boldsymbol{z}}$ is

$$\tilde{\boldsymbol{z}}=\boldsymbol{D}\tilde{\boldsymbol{w}}. \tag{13}$$

Eq. (12) is a special case of sparse representation

$$\arg\min_{\boldsymbol{w}}\left(f(\boldsymbol{w})+\lambda\|\boldsymbol{w}\|_1\right), \tag{14}$$

where $f(\cdot)$ is a smooth convex loss function. The optimization problem (14) can be solved by the accelerated proximal gradient method. It is an iterative algorithm and can be summarized as Algorithm 1 (Wright *et al.*, 2009).

3.3 Speech enhancement

In the GSC structure, the adaptive filter uses the reference signal to estimate the interfering signal. The estimated interfering signal is then subtracted

Algorithm 1 Accelerated proximal gradient method for sparse coding

Require:
Loss function $f(\cdot)$,
Regularization parameter λ,
Initial affine combination parameter β^0,
Initial coefficient vector $\boldsymbol{w}^0$,
Convergence threshold τ.

Ensure:
Vector of coefficients $\boldsymbol{w}^*$.

Steps:

1: **Repeat**
2: Calculate the search point via an affine combination method:
$$\boldsymbol{v}^{(k)} = \boldsymbol{w}^{(k)} + \beta^{(k)}(\boldsymbol{w}^{(k)} - \boldsymbol{w}^{(k-1)});$$
3: Calculate the next gradient descent point $\boldsymbol{u}^{(k+1)}$ with an adaptive step size $t^{(k)}$:
$$\boldsymbol{u}^{(k+1)} = \boldsymbol{v}^{(k)} - t^{(k)}\nabla f(\boldsymbol{v}^{(k)});$$
4: Calculate the next vector of coefficients using the proximal operator $\boldsymbol{w}^{(k+1)}$:
$$\boldsymbol{w}^{(k+1)} = \arg\min_{\boldsymbol{w}} \left(\tfrac{1}{2}\|\boldsymbol{w} - \boldsymbol{u}^{(k+1)}\|_2^2 + t^{(k)}\lambda\|\boldsymbol{w}\|_1\right);$$
5: Update $t^{(k+1)}$ and $\beta^{(k+1)}$ for the next iteration;
6: $k \leftarrow k+1$;
7: **Until** $\|\boldsymbol{w}^{(k+1)} - \boldsymbol{w}^{(k)}\|_2 \le \tau$
8: **Return** $\boldsymbol{w}^* = \boldsymbol{w}^{(k+1)}$;

from FBF output for speech enhancement. The reference signal is obtained through using BM to block the desired signal in a noisy signal which is received by microphone arrays. Since the real acoustic environment of communication applications is usually affected by reverberation, the reference signal contains a little desired signal due to leakage. Then the weight of an adaptive filter cannot actively track the interfering signal changes in the segments of the desired signal, and the interference suppression of GSC will be limited.

To reduce the leakage in the reference signal, we use sparse coding to further suppress the residual desired signal that leaks into the reference signal. The training samples for dictionary learning come from the segments of the desired signal inactivity of FBF output. The VAD algorithm (Sohn *et al.*, 1999; Eshaghi and Karami Mollaei, 2010; Tanyer and Ozer, 2000) is employed to obtain the segments of the desired signal inactivity. As the interferer dictionary is also coherent to the structured component of the reference signal of the interference, the reference signal will be preserved, while the desired signal that leaks into the reference signal will be suppressed by sparse coding.

The FBF output is achieved by resolving Eq. (5) and the BM is achieved by resolving Eq. (6). After sparse coding for the reference signal, the NLMS in the frequency domain (Avargel and Cohen, 2008) will be employed to suppress the interference of FBF output.

4 Experiments

The performance of the proposed algorithm has been evaluated in both simulation and the real acoustic environment. The desired speech and the interfering signals come from the TIMIT database and NOISE-92 database respectively, and are downsampled to 16 kHz in all experiments. GSC (Griffiths and Jim, 1982) and TF-GSC (Krueger *et al.*, 2011) methods are used for comparison. The training samples for interferer dictionary learning come from the segments of the desired speech pauses in the fixed beamformer output. The VAD algorithm based on wavelet transform (Eshaghi and Karami Mollaei, 2010) is used to obtain the segments. The analysis window of STFT is a 256-point Hamming window with 50% overlap. The size of the overcomplete dictionary is 512 and each atom is a vector with 256 elements. The regularization parameter λ for the sparse constraint in Eq. (12) is set to 0.1. The microphone array is a uniform linear array composed of four omnidirectional microphones, and the distance between adjacent microphones is set to 4 cm. In addition, we suppose that the noisy signal does not contain the desired speech in the first frame, in order to obtain the interferer dictionary by dictionary learning.

4.1 Simulation environment

The Habets method (Habets, 2010) is used to achieve the simulated acoustic impulse responses in the following. The simulation room is 3 m × 6 m × 2.8 m and the of four microphones are located at (1.44, 2.5, 1.6), (1.48, 2.5, 1.6), (1.52, 2.5, 1.6), and (1.56, 2.5, 1.6), respectively. The desired source is located at (1.49, 3.0, 1.6) and the interference source at (2.5, 3.5, 1.6). The reverberation time (RT_60) is 200 ms. Fig. 3 shows the relative position of the arrays and signal sources.

In the first experiment, the spectrograms in the frequency domain and the waveforms in the time domain are used to demonstrate the ability of nonstationary interference suppression of the proposed

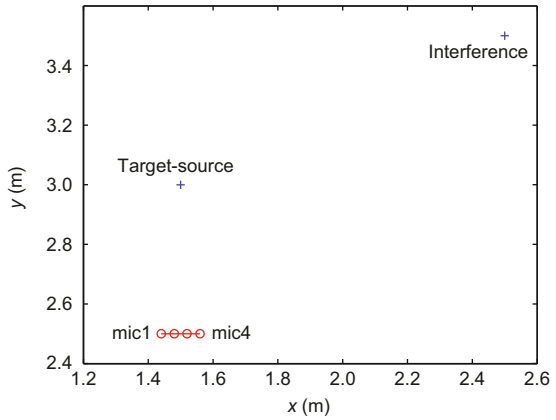

Fig. 3 The positional relationship among the arrays, desired source, and interference source

method. Without loss of generality, we choose music signal as the interference source. Figs. 4a–4i are the spectrograms and waveforms of the clean desired speech, interference, noisy signal, reference signal, and enhanced signal obtained by different methods, respectively. Comparing Figs. 4d, 4e, and 4f, we can easily find that a small component of the desired speech exists in the reference signal in Figs. 4d and 4e, while the proposed reference signal in Fig. 4f has a small component of the desired speech. This demonstrates that by using sparse coding our method can obtain a better reference signal than those using only the blocking matrix. Figs. 4g–4i show that the enhanced signal obtained by the proposed method has fewer interfering components than the enhanced signal obtained by its counterpart.

The results of different enhanced signals illustrate that the less the desired signal that leaks into the reference signal, the more the interference cancellation that will be obtained. In addition, comparison of Figs. 4a and 4i shows that the enhanced signal obtained using the proposed method has no obvious distortion.

In the second experiment, to suppress the non-random signal we use SNR as a metric to test the ability of our method. The results of different algorithms at different SNR levels are shown in Fig. 5. SNR is defined as

$$\mathrm{SNR} = 10\lg\frac{p(x)}{p(n)}, \tag{15}$$

where function $p(\cdot)$ is the PSD of a signal. The PSD of the interfering signal for the output SNR is estimated via the minimum statistics method (Martin, 2001; 2006) and then the PSD of the desired speech can be obtained.

Fig. 5 shows that our proposed algorithm can improve the SNR by about 15 dB on average at different SNR levels and the SNR improvement achieved by our method is higher than that obtained by the other two algorithms.

To evaluate the effect of random signal suppression, we use white noise (Gaussian noise) as the interfering source in the third experiment. The SNR improvements at different SNR levels are shown in Fig. 6. Although white noise is not sparse with respect to any fixed dictionary (Kowalski and Torrésani, 2008; Rauhut *et al.*, 2008), most of reference signal components of white noise will be preserved by sparse coding with the learned dictionary in the reconstructed signal. Meanwhile, the desired speech component that leaks into the reference signal is still incoherent to the learned dictionary and will be suppressed by sparse coding. Then the reconstructed reference signal with a small residual speech component used in an adaptive filter can achieve the estimate of the white noise. Fig. 6 shows that in the white noise environment the SNR improvement achieved by the proposed method is about 2 dB and 5 dB higher than the TF-GSC and GSC algorithms, respectively.

In the last experiment, we use the perceptual evaluation of speech quality mean opinion score (PESQ MOS), a standard of wideband audio (ITU, 2007), to evaluate the ability of different speech enhancement algorithms. The higher the PESQ MOS, the better the quality of the desired speech signal achieved by the speech enhancement algorithm.

We use babble, music, car, factory, and white noise as background interference, respectively. To compare the effect of different interference suppressions, an input SNR level of 1 dB is employed in each interference environment. The results of PESQ MOS are shown in Table 1.

Table 1 shows that the PESQ MOS results of our algorithm for different interfering signals are better than those of the other two speech enhancement algorithms.

4.2 Real acoustic environment

The microphone array is a uniform linear array composed of four silicon micro omnidirectional microphones. We use DAR-2000 digital signal acquisition of Quanzhou Hengtong Technology for audio

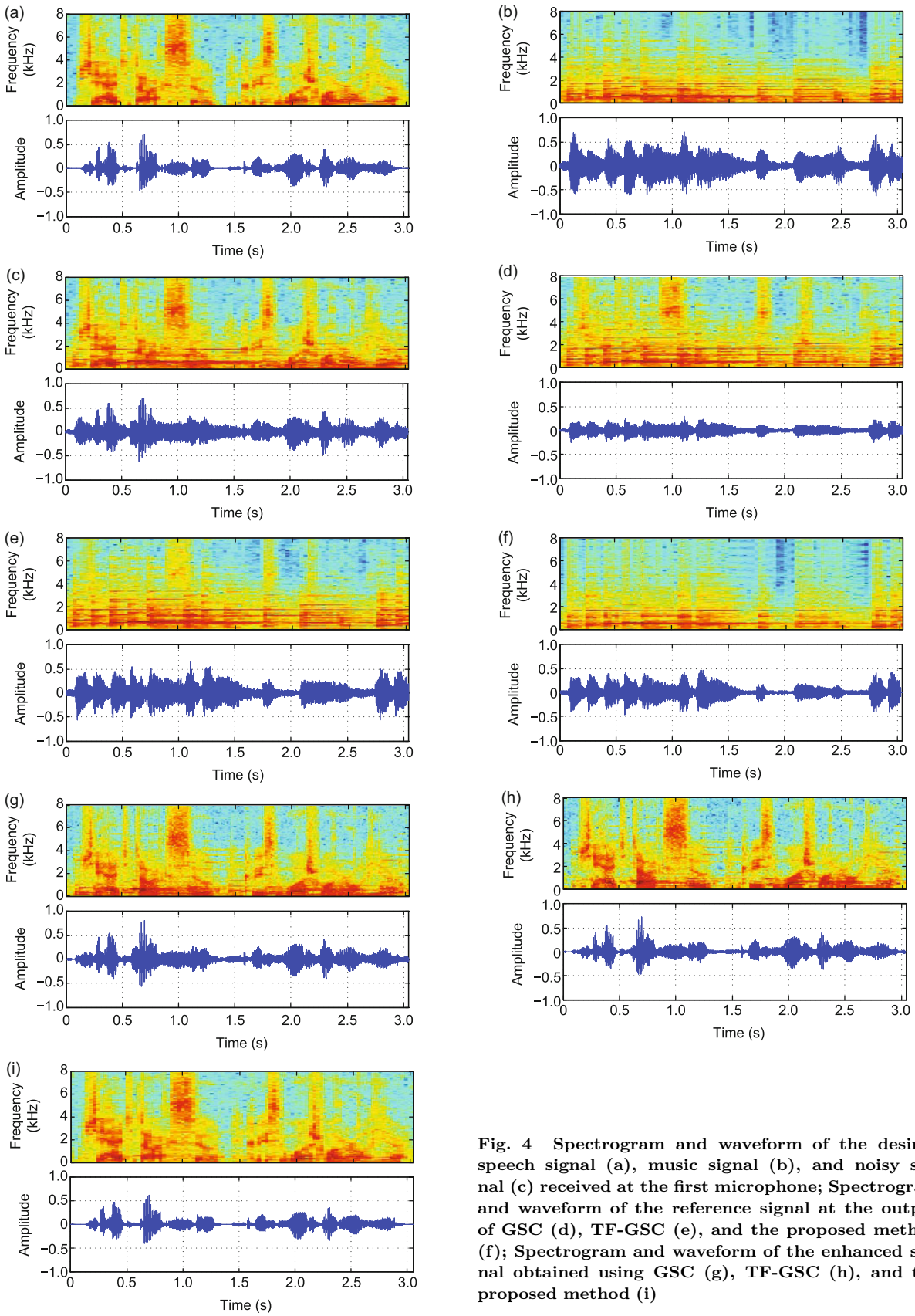

Fig. 4 Spectrogram and waveform of the desired speech signal (a), music signal (b), and noisy signal (c) received at the first microphone; Spectrogram and waveform of the reference signal at the output of GSC (d), TF-GSC (e), and the proposed method (f); Spectrogram and waveform of the enhanced signal obtained using GSC (g), TF-GSC (h), and the proposed method (i)

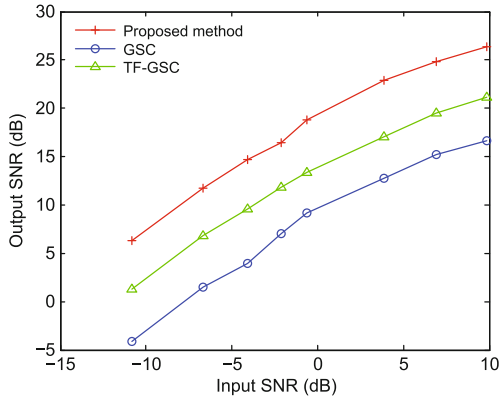

Fig. 5 SNR improvement of the competing algorithms in a music interference environment

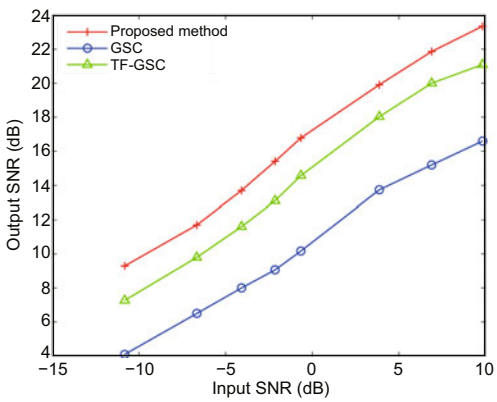

Fig. 6 SNR improvement of the competing algorithms in a white noise environment

Table 1 PESQ MOS results for the three algorithms

Method	PESQ MOS				
	Babble	Car	Factory	Music	White noise
GSC	1.72	2.65	1.83	2.00	2.08
TF-GSC	2.34	3.21	2.73	2.84	3.35
Proposed	3.03	3.81	3.40	3.52	3.76

capturing and the sampling rate is set to 16 kHz. We choose a 6 m × 5 m × 3 m laboratory as the experimental environment. The desired source is located at a distance of about 50 cm to the front of the array, and the interference source is located at a distance of about 1 m to the left front of the array. We use babble, car, factory, music, and white noise as background interfering signals respectively, and the results for the different speech enhancement algorithms are shown in Table 2.

Table 2 shows that the proposed algorithm is better than the other two algorithms for different interference suppressions in the real environment. This further proves that GSC with sparse coding is reliable.

Table 2 SNR results for the three algorithms

Signal	Input	SNR (dB)		
		GSC	TF-GSC	Proposed
Babble	0.50	4.69	9.12	13.93
Car	4.35	13.38	16.06	21.51
Factory	–6.12	1.01	4.36	10.15
Music	–3.83	2.19	6.35	11.56
White noise	–1.79	6.40	11.02	15.91

5 Conclusions

In this paper, speech enhancement based on GSC-like structure with sparse coding is proposed for communication applications. For reference signal, we use sparse coding with the interferer dictionary to reduce the residual desired signal. An adaptive filter with improved reference signal can suppress the interfering signal effectively. The training samples for interferer dictionary learning come from the segments of the desired signal inactivity. Since the interferer dictionary is coherent to the structured component of the reference signal and of little coherence to the structured component of the desired signal, the residual desired signal can be reduced by sparse coding. Simulation and experiments in the real environment demonstrate that our algorithm works well in different interference environments.

References

Aharon, A.M., Elad, M., 2006. *K*-SVD: an algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Trans. Signal Process.*, **54**(11):4311-4322. [doi:10.1109/TSP.2006.881199]

Avargel, Y., Cohen, I., 2008. Adaptive system identification in the short-time fourier transform domain using cross-multiplicative transfer function approximation. *IEEE Trans. Audio Speech Lang. Process.*, **16**(1):162-173. [doi:10.1109/TASL.2007.910789]

Elad, M., Aharon, M., 2006. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Trans. Image Process.*, **15**(12):3736-3745. [doi:10.1109/TIP.2006.881969]

Engan, K., Skretting, K., Husøy, J.H., 2007. Family of iterative LS-based dictionary learning algorithms, ILS-DLA, for sparse signal representation. *Dig. Signal Process.*, **17**(1):32-49. [doi:10.1016/j.dsp.2006.02.002]

Eshaghi, M., Karami Mollaei, M., 2010. Voice activity detection based on using wavelet packet. *Dig. Signal Process.*, **20**(4):1102-1115. [doi:10.1016/j.dsp.2009.11.008]

Gannot, S., Burshtein, D., Weinstein, E., 2001. Signal enhancement using beamforming and nonstationarity with applications to speech. *IEEE Trans. Signal Process.*, **49**(8):1614-1626. [doi:10.1109/78.934132]

Gemmeke, J.F., Cranen, B., 2009. Sparse imputation for noise robust speech recognition using soft masks. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, p.4645-4648. [doi:10.1109/ICASSP.2009.4960666]

Gribonval, R., Schnass, K., 2008. Some recovery conditions for basis learning by ℓ_1-minimization. IEEE 3rd Int. Symp. on Communications, Control and Signal Processing, p.768-773. [doi:10.1109/ISCCSP.2008.4537326]

Griffiths, L., Jim, C., 1982. An alternative approach to linearly constrained adaptive beamforming. *IEEE Trans. Antennas Propag.*, **30**(1):27-34. [doi:10.1109/TAP.1982.1142739]

Habets, E.A.P., 2010. Room Impulse Response Generator for MATLAB. Univeristy of Erlangen-Nuremberg, Bavaria, Germany.

He, Y., Han, J., Deng, S., *et al.*, 2012. A solution to residual noise in speech denoising with sparse representation. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, p.4653-4656. [doi:10.1109/ICASSP.2012.6288956]

Herbordt, W., Kellermann, W., 2001. Efficient frequency-domain realization of robust generalized sidelobe cancellers. IEEE 4th Workshop on Multimedia Signal Processing, p.377-382. [doi:10.1109/MMSP.2001.962763]

Hoshuyama, O., Sugiyama, A., Hirano, A., 1999. A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters. *IEEE Trans. Signal Process.*, **47**(10):2677-2684. [doi:10.1109/78.790650]

ITU, 2007. Wideband Extension to Rec. P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs, P.862.2. International Telecommunication Union, Geneva.

Kowalski, M., Torrésani, B., 2008. Random models for sparse signals expansion on unions of bases with application to audio signals. *IEEE Trans. Signal Process.*, **56**(8):3468-3481. [doi:10.1109/TSP.2008.920144]

Krueger, A., Warsitz, E., Haeb-Umbach, R., 2011. Speech enhancement with a GSC-like structure employing eigenvector-based transfer function ratios estimation. *IEEE Trans. Audio Speech Lang. Process.*, **19**(1):206-219. [doi:10.1109/TASL.2010.2047324]

Mairal, J., Bach, F., Ponce, J., *et al.*, 2010. Online learning for matrix factorization and sparse coding. *J. Mach. Learn. Res.*, **11**:19-60.

Martin, R., 2001. Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Trans. Speech Audio Process.*, **9**(5):504-512. [doi:10.1109/89.928915]

Martin, R., 2006. Bias compensation methods for minimum statistics noise power spectral density estimation. *Signal Process.*, **86**(6):1215-1229. [doi:10.1016/j.sigpro.2005.07.037]

Plumbley, M.D., Blumensath, T., Daudet, L., *et al.*, 2010. Sparse representations in audio and music: from coding to source separation. *Proc. IEEE*, **98**(6):995-1005. [doi:10.1109/JPROC.2009.2030345]

Rauhut, H., Schnass, K., Vandergheynst, P., 2008. Compressed sensing and redundant dictionaries. *IEEE Trans. Inform. Theory*, **54**(5):2210-2219. [doi:10.1109/TIT.2008.920190]

Rebollo-Neira, L., 2004. Dictionary redundancy elimination. *IEEE Proc.-Vis. Image Signal Process.*, **151**(1):31-34. [doi:10.1049/ip-vis:20040294]

Sigg, C.D., Dikk, T., Buhmann, J.M., 2012. Speech enhancement using generative dictionary learning. *IEEE Trans. Audio Speech Lang. Process.*, **20**(6):1698-1712. [doi:10.1109/TASL.2012.2187194]

Skretting, K., Engan, K., 2010. Recursive least squares dictionary learning algorithm. *IEEE Trans. Signal Process.*, **58**(4):2121-2130. [doi:10.1109/TSP.2010.2040671]

Sohn, J., Kim, N.S., Sung, W., 1999. A statistical model-based voice activity detection. *IEEE Signal Process. Lett.*, **6**(1):1-3. [doi:10.1109/97.736233]

Talmon, R., Cohen, I., Gannot, S., 2009. Convolutive transfer function generalized sidelobe canceler. *IEEE Trans. Audio Speech Lang. Process.*, **17**(7):1420-1434. [doi:10.1109/TASL.2009.2020891]

Tanyer, S.G., Ozer, H., 2000. Voice activity detection in nonstationary noise. *IEEE Trans. Speech Audio Process.*, **8**(4):478-482. [doi:10.1109/89.848229]

Wright, S.J., Nowak, R.D., Figueiredo, M.A.T., 2009. Sparse reconstruction by separable approximation. *IEEE Trans. Signal Process.*, **57**(7):2479-2493. [doi:10.1109/TSP.2009.2016892]92]