Available online at www.sciencedirect.com

SciVerse ScienceDirect

Speech Communication 55 (2013) 757–768

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Compressive speech enhancement

Siow Yong Low [a,*], Duc Son Pham [b], Svetha Venkatesh [c]

[a] Curtin University, Sarawak Campus, Miri, Malaysia
[b] Curtin University, Department of Computing, WA, Australia
[c] Deakin University, Centre for Pattern Recognition and Data Analytics, Victoria, Australia

Received 17 August 2012; received in revised form 18 February 2013; accepted 3 March 2013
Available online 26 March 2013

Abstract

This paper presents an alternative approach to speech enhancement by using compressed sensing (CS). CS is a new sampling theory, which states that sparse signals can be reconstructed from far fewer measurements than the Nyquist sampling. As such, CS can be exploited to reconstruct only the sparse components (e.g., speech) from the mixture of sparse and non-sparse components (e.g., noise). This is possible because in a time-frequency representation, speech signal is sparse whilst most noise is non-sparse. Derivation shows that on average the signal to noise ratio (SNR) in the compressed domain is greater or equal than the uncompressed domain. Experimental results concur with the derivation and the proposed CS scheme achieves better or similar perceptual evaluation of speech quality (PESQ) scores and segmental SNR compared to other conventional methods in a wide range of input SNR.

Keywords: Compressed sensing; Speech enhancement; Sparsity

1. Introduction

In many practical applications, speech signal is degraded due to unwanted interference and hence speech enhancement is important to enhance intelligibility and the overall perceptual quality of degraded speech. Speech enhancement can be broadly categorized into multiple channel and single channel approaches (Benesty et al., 2005; Brandstein and Ward, 2001). As the name multi-channel implies, multi-channel solutions require more than one microphone for signal observations. In this case, multi-channel techniques primarily exploit spatial information to separate the signal of interest from other interfering signals. A common method to perform spatial filtering (or beamforming) is to make use of the array geometry to form a beam towards the target signal. This technique has been widely studied and considerable noise suppression is reported (Veen and Buckley, 1988; Davis et al., 2005; Dam et al., 2004). However, beamforming based methods rely on apriori information about the acoustical environments and the target signal localization. This means that beamforming methods may be susceptible to potential modeling errors (e.g., model mismatch error) particularly when the transfer function from source to sensor is difficult to model (Hoshuyama et al., 1999; Low et al., 2002). Some multi-channel approaches, which do not require source localization have been proposed in Lotter et al. (2003) and Low and Nordholm (2005).

Single channel techniques, on the other hand, offer a much more computationally appealing solution since only one microphone is needed (Paliwal et al., 2010; Cohen, 2003; O'Shaughnessy, 2000). Recent advances in single-channel techniques have seen attractive speech enhancement applications such as cochlear implants (Kokkinakis and Loizou, 2008). One popular single channel speech enhancement technique is the spectral subtraction (Boll, 1979). It was originally suggested by Boll and has since

* Corresponding author. Tel.: +60 128708345.
E-mail addresses: siowyong@curtin.edu.my (S.Y. Low), DucSon.Pham@curtin.edu.au (D.S. Pham), svetha.venkatesh@deakin.edu.au (S. Venkatesh).

0167-6393/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2013.03.003

gained huge acceptance due to its simplicity (Paliwal et al., 2010; Yang, 1993; Lu, 2011). Basically, spectral subtraction relies on the assumption that the target signal and noise signal are uncorrelated. Therefore, if the noise spectral component is estimated correctly, the target signal can be enhanced by subtracting the estimated spectral noise from the noisy spectral observations. Typically, a voice activity detector (VAD) is used detect the presence of speech and non-speech periods for the estimation of noise statistics. However, any mis-detection by the VAD will result in erroneous updates, which in turn causes spectral subtraction to become defective. A comprehensive review on single channel techniques can be found in Principi et al. (2010).

Recently, there is a new sampling theory called compressed sensing (CS). CS states that super-resolved signals and images can be reconstructed from far fewer measurements than the Nyquist sampling (Donoho, 2006). Whilst compressed sensing/sparsity learning has been a celebrated theory recently and its applications to images are popular, its applications to speech and audio signals are difficult and limited (Sreenivas and Kleijn, 2009; Jancovic et al., 2012). Most of the recent applications are mainly on linear predictive coding of speech in the residual domain (Giacobello et al., 2012; Griffin et al., 2011) or dictionary design (Christensen et al., 2009) and to the best of the authors' knowledge, none of the applications address speech enhancement.

At the very heart of CS sampling is its sparsity assumption and CS theory shows that sparse signals with a small set of linear measurements can be reconstructed with an overwhelming probability (Candès et al., 2006; Candès and Tao, 2006). Suffice to say, under the presence of both sparse and non-sparse components, only sparse signals will be reconstructed. Interestingly, since speech signals are generally sparse in the time-frequency representation (Pham et al., 2009), while many types of noise is non-sparse, CS may hold the potential as a speech discriminator. This means that CS can be designed to reconstruct only the sparse components (speech) from the mixture of sparse and non-sparse components (noise). Potentially, the speech enhancement process can be made to rely upon the strength of CS to maintain only the sparse components and its weakness in preserving the non-sparse components.

In the spirit of speech enhancement, this paper investigates the feasibility of using CS to perform speech enhancement. The most related work to CS based speech enhancement technique is the wavelet based method by Wu et al. (2011). However, the work in Wu et al. (2011) is mainly empirical and is not supported by a theoretical framework. Most importantly, we have found that the global sparsity model (batch processing) adopted in Wu et al. (2011) is inferior compared to the local sparsity modeling in each subband. In the subsequent section, we also outline the differences between wavelet and STFT transformations. This paper shows that on average, the signal to noise ratio (SNR) of the output of time-frequency CS is greater than the SNR of the original noisy signal. The derivation details that the SNR improvement is directly proportional to the sensing dimension, M and the largest eigenvalue of the observation matrix, $\lambda_{\max}$. Further, the theoretical finding is extended and validated by proposing a CS based algorithmic solution that performs speech enhancement. It is worth mentioning that the CS based method relies only on the sparsity of speech of interest and bypasses the need for noise estimation or a VAD. Whilst the performance of the proposed CS scheme is not exceedingly superior in comparison with other speech enhancement algorithms, it is interesting to note that CS could be readily used for improving the SNR. More importantly, both the perceptual evaluation of speech quality (PESQ) scores and the segmental SNR improvement concur with the theoretical finding. Also, an inherent positive byproduct from this scheme is the reduction of sample measurements as the compression/enhancement is performed on the signals.

The paper is organized as follows. Section 2 provides a general background on CS and the proposed CS speech enhancement scheme is detailed in Section 2. Section 3 presents its performance evaluation and lastly, Section 5 summarizes the findings.

2. Background

2.1. Overview of compressed sensing

CS theory states that if a signal has a sparse representation in one basis then it can be recovered from a small number of projections onto a second basis, which is incoherent with the first (Donoho, 2006; Candès, 2006; Rachlin and Baron, 2008). Therefore, it is possible to reconstruct a signal from far fewer samples or measurements than conventional methods use. In other words, the number of measurements can be much lower than the number of samples needed if the signal is sampled at the Nyquist rate. Such capability brings about the benefits of reduced storage space and transmission bandwidth due to the compression achieved.

At the very heart of this compression capability lie two major assumptions, i.e., sparsity, which pertains to the signals of interest and incoherence, which is related to the sensing modality (Candès and Wakin, 2008). Sparsity refers to the idea that the information rate of a continuous time signal may be much smaller than suggested by its bandwidth. Hence, this assumption can be extended to many natural signals that are sparse or compressible in the sense that they have concise representations when expressed in the proper basis. Incoherence on the other hand expresses the idea that objects having a sparse representation must be spread out in the domain in which they are acquired (Candès and Wakin, 2008; Candès and Tao, 2006).

To define sparsity, let us consider a $N \times N$ matrix $\mathbf{\Psi}$ whose columns form an orthonormal basis. Thus a K-sparse signal, $\mathbf{x}(n) \in \mathbb{R}^N$ can be expressed as

$$\mathbf{x}(n) = \mathbf{\Psi}\theta(n), \tag{1}$$

where $\theta(n) \in \mathbb{R}^N$ has K non-zero entries. A CS vector measurement can then be defined as follows

$$\mathbf{y}(n) = \mathbf{\Phi}\mathbf{x}(n), \tag{2}$$

where $\mathbf{x}(n)$ is a $N \times 1$ vector and $\mathbf{\Phi}$ is a $M \times N$ sensing matrix/linear mapping matrix, i.e., $\mathbf{\Phi} : \mathbb{R}^N \to \mathbb{R}^M$. Note that the sensing matrix, $\mathbf{\Phi}$ has a significantly smaller number of rows than columns, i.e., $M \ll N$. This means that the dimension of $\mathbf{y}(n)$ is considerably smaller than $\mathbf{x}(n)$, hence the term "compressed". Eq. (2) represents an alternative sampling procedure, which samples sparse signals close to their intrinsic information rate rather than their Nyquist rate. It has been shown that the tractable recovery of K-sparse signal, $\mathbf{x}(n)$ from the measurements, $\mathbf{y}(n)$ requires the sensing matrix, $\mathbf{\Phi}$ to obey the restricted isometry property (RIP) (Candès and Tao, 2006). Here, a sensing matrix, $\mathbf{\Phi}$ is said to satisfy RIP of order K for all K-sparse signal, $\mathbf{x}(n)$, if there exists a constant, $\delta_K \in (0, 1)$ such that

$$(1 - \delta_K)\|\mathbf{x}(n)\|_2^2 \leqslant ||\mathbf{\Phi}\mathbf{x}(n)||_2^2 \leqslant (1 + \delta_K)\|\mathbf{x}(n)\|_2^2, \tag{3}$$

where $\|\cdot\|_2$ denotes ℓ_2 norm.

2.2. Compressed sensing reconstruction

There are broadly two reconstruction algorithms for CS, namely basis pursuit (BP) (Chen et al., 1998) and orthogonal matching pursuit (MP) (Tropp and Gilbert, 2007). Briefly, BP seeks to find a solution to the following problem

$$\hat{\mathbf{x}}(n) = \arg\min_{\mathbf{x}(n)} \|\mathbf{x}(n)\|_1 \quad \text{s.t.} \quad \mathbf{y}(n) = \mathbf{\Phi}\mathbf{x}(n), \tag{4}$$

where $\|\cdot\|_1$ is the ℓ_1 norm. Orthogonal MP on the other hand, is a greedy algorithm, which iteratively approximates $\mathbf{x}(n)$ by incorporating the component from the measurement set in the reconstructed set that forms a residual function, $\mathbf{r} = \mathbf{y}(n) - \mathbf{\Phi}\mathbf{x}(n)$ (Tropp and Gilbert, 2007; Boufounos et al., 2007). The algorithm is terminated once the residual satisfies a certain threshold. By orthogonal, it means that the residual is made orthogonal to all measurement vectors in the previous iterations. This is to ensure that the reconstruction process is guaranteed to a sparse solution (Boufounos et al., 2007).

Importantly, there is a wide range of choices for the CS matrix $\mathbf{\Phi}$. One of the popular choice is a random Gaussian matrix which can be shown to satisfy the CS requirement with high probability, so that the acquisition part is very practical. For the recovery part, there are different approaches to CS recovery, including greedy methods (Boufounos et al., 2007; Chen et al., 1998; Tropp and Gilbert, 2007), convex optimization methods, iterative thresholding methods, or Bayesian sparse learning methods, all of which finds sparse $\mathbf{x}(n)$ that is consistent with the measurements as much as possible. Due to space restriction, we omit the detailed background on CS and for this the reader is referred to the CS repository[1]. Hereinafter, we assume the reader is already familiar with basic CS terminology and only discuss how to exploit the CS ideas for solving the speech enhancement problem.

2.3. Compressive speech enhancement

This paper first examines the ability or the feasibility of compressed sensing in enhancing speech via the removal of noise. In each branch of the diagram shown in Fig. 1, the CS 'black box' acts like a denoiser. This means that for each noisy speech input $\mathbf{x}$, which may contain clean speech plus some noise, the CS box return a version of the input, denoted as $\mathbf{x}'$. The denoising process is controlled via the regularization parameter λ in the CS formulation mentioned previously. Note that a large λ makes the output more sparse. Clearly a good choice of λ should provide a reasonable trade-off between smoothness of the reconstructed signal and similarity to the original signal (Kim et al., 2007).

The CS theory tells us that if the input signal $\mathbf{x}$ is sparse, then the recovery $\mathbf{x}'$ from the CS-projected data $\mathbf{y} = \mathbf{\Phi}\mathbf{x}$ is also sparse and is close to $\mathbf{x}$ under certain conditions. Interestingly, what will happen if the input signal $\mathbf{x}$ has some offset or DC constant? The following toy example illustrates that such an offset is automatically suppressed when passing through CS.

In the aforementioned example, we consider a sparse signal with some DC offset of about 1 unit. The sparse signal has a length of $N = 128$ and has two peaks at indices 10 and 20 respectively. Note that in this case, sparse refers only to the two peaks, whereas the DC offset is not sparse as the offset is present throughout. Here, CS attempts only to recover the sparse components, which in this case are the two peaks. The "not sparse" components, i.e., the DC offset will naturally not be recovered by CS. This is shown on the top plot of Fig. 1. The middle plot shows the CS-projected data $\mathbf{y} = \mathbf{\Phi}\mathbf{x}$ whilst the bottom plot shows that CS recovery from $\mathbf{y}$ with suitable λ. Whilst the original noisy input signal is not strictly sparse due to the presence of the DC component, the recovery is observed to be able to do two important things:

- It recovers the two major peaks of the true signal successfully.
- The DC noise floor almost disappears from the recovered signal, implying the amount of noise energy in the output signal is much reduced.

In other words, this illustrates a very surprising property of CS: *DC suppression*. This property has not been mentioned in previous work on CS. Intuitively, it can be explained that due to the i.i.d. random property of the CS matrix $\mathbf{\Phi}$ e.g., random Gaussian matrix, its product with any constant vector yields almost a zero vector with high probability, leading to DC suppression. For speech

[1] http://dsp.rice.edu/cs.

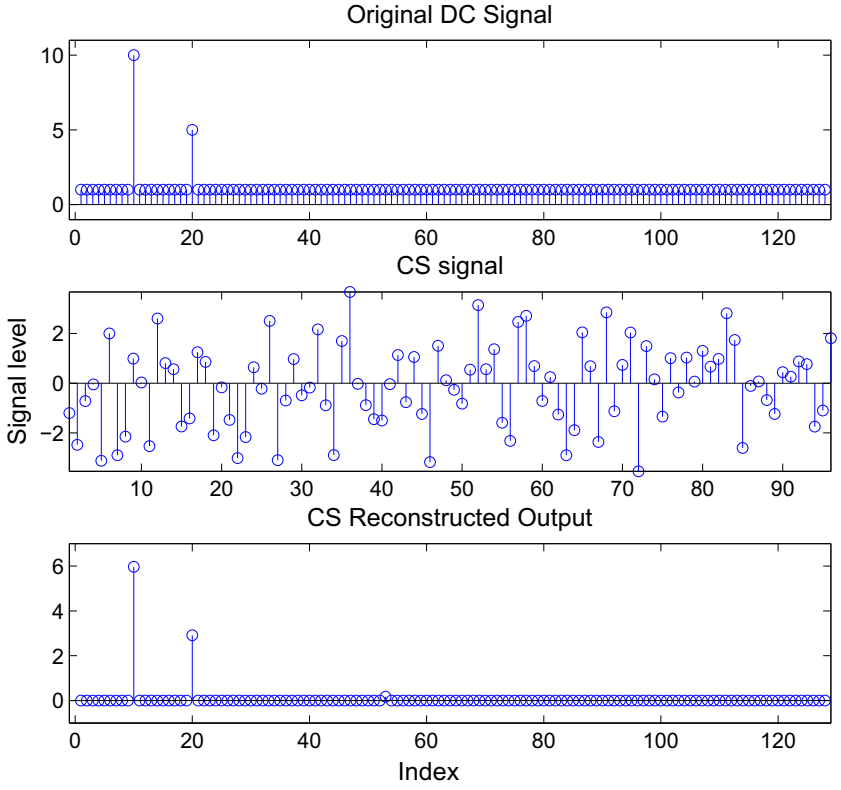


Fig. 1. A toy example illustrating the DC supression capability of the CS filter

enhancement, this property is very useful if we view the DC offset as the noise floor in the frequency domain and the sparse signal is the time-frequency distribution of speech. The above example suggests that if the noise floor can be suppressed when passed through CS then we can obtain a sparsely enhanced output signal. In practice, the real signals might be much more complicated and the noise is not an ideal DC but rather a noisy floor. However, the advantage of the CS is that we do not need to set multiple thresholds at each subband separately, but control the smoothness of the filtered signal through only one parameter λ which designates the prior 'sparsity' of speech signals. In the following, we further quantify the improvement in the enhanced speech signal more precisely.

3. Compressed sensing based speech enhancement

3.1. Proposed method

As mentioned previously, speech signals are generally sparse in the time-frequency representation (Pham et al., 2009; Yu and Hansen, 2009). The sparsity nature of speech can be understood from the fact that speech is highly non-stationary and there are lapses of time-frequency periods where the speech power is negligible compared to the average power. Moreover, humans rarely excite all frequencies at any one time. For instance, average speech signal consists of approximately ten to fifteen phonemes per second and each of these phonemes has varying spectral components (Ghosh et al., 2011). In this paper, the denoising process is proposed to be performed on the spectral envelope of the short-time Fourier transform (STFT) of the noisy signal. The spectral envelope is known to be a good representation of both the speech and noise information.

Fig. 2 illustrates the block diagram of the proposed CS based speech enhancement scheme. The figure shows the denoising process is performed in parallel for all frequency bin. First, the noisy speech signal is decomposed into time-frequency presentation via STFT. Denote as $\mathbf{x}$ the signal in a subband. Then the signal is CS-transformed via $\mathbf{y} = \mathbf{\Phi x}$ where $\mathbf{\Phi}$ can be selected from a number of CS matrices such as partial DCT or Gaussian matrices. Due to the nature of embedding from high- to low-dimensional spaces, all the salient features of the speech are approximately preserved whilst the random noise can be effectively suppressed. Following that, CS recovery is performed to seek the sparse solution $\hat{\mathbf{x}}$ from $\mathbf{y}$. In our implementation, the following basis pursuit denoising (BPDN) formulation is adopted

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}\in\mathbb{R}^N} \|\mathbf{y} - \mathbf{\Phi x}\|_2^2 + \lambda\|\mathbf{x}\|_1. \tag{5}$$

In this framework, only a single parameter λ controls the sparsity of the reconstructed (denoised) signal $\hat{\mathbf{x}}$. Note that in the proposed method, no noise estimation or the use of a VAD is needed. The CS method automatically preserves only the sparse components in the observed signal. Naturally, an optimal choice of λ should provide a reason-

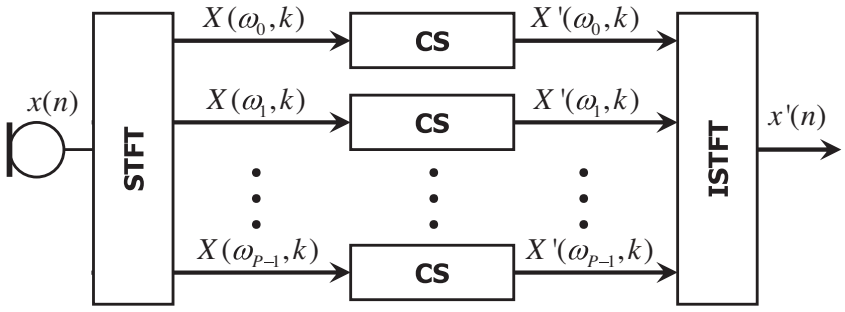


Fig. 2. The proposed CS based speech enhancement scheme.

able trade-off between smoothness of the reconstructed signal and similarity to the original signal (see Section 4). In this implementation, the `l1_ls` algorithm (Kim et al., 2007) is used to solve this reconstruction problem. The CS-reconstructed signals from all frequency bins are then synthesized by using the inverse STFT to obtain the full-band enhanced speech signal.

3.2. Denoising principle

In what follows, we give an analysis of the denoising aspects of our proposed method. Let the noisy signal be

$$x(n) = s(n) + v(n), \tag{6}$$

where $s(n)$ and $v(n)$ are the speech and noise signals, respectively. Its corresponding P-point STFT is given as

$$X(\omega,k) = \sum_{n=0}^{P-1} x(n)w(n-kR)e^{-j\omega n} = S(\omega,k) + V(\omega,k), \tag{7}$$

where $w(n-kR)$ is a time-limited window function with a hop size of R and length P, $\omega \in \omega_0,\ldots,\omega_{P-1}$ and k is the time index. The kth instant data envelope of (7) is $|X(\omega,k)|$, where $|\cdot|$ denotes the absolute value operator. Further, it is assumed that the data matrix $\mathbf{X}(\omega,k) \in \mathbb{R}^{N\times L}$ consists of L realizations of the noisy envelope signal of length N. Mathematically, it is defined as

$$\mathbf{X}(\omega,k) = [\mathbf{x}_1(\omega,k), \mathbf{x}_2(\omega,k), \ldots, \mathbf{x}_L(\omega,k)], \tag{8}$$

where the lth realization is given as

$$\mathbf{x}_l(\omega,k) = [|X_l(\omega,k)|, |X_l(\omega,k-1)|, \cdots, |X_l(\omega,k-N+2)|, |X_l(\omega,k-N+1)|]^T. \tag{9}$$

The symbol $[\cdot]^T$ is the transposition operator. For notational simplicity, from here on, the indices (ω,k) are omitted, e.g., $\mathbf{X}(\omega,k)$ will be expressed as $\mathbf{X}$. Also, denote the covariance matrix as $\mathbf{\Sigma_X} = \mathbf{X}\mathbf{X}^T$ and let $\mathcal{K}$ be the number of principal (largest) eigenvalues of $\mathbf{\Sigma_X}$ where $\mathcal{K} \ll N$. Then the SNR of $\mathbf{X}$ can be expressed in terms of its eigenvalues as

$$\mathsf{SNR}_{\mathbf{X}} = \frac{\sum_{i=1}^{\mathcal{K}} \lambda_i(\mathbf{\Sigma_X})}{\sum_{i=\mathcal{K}+1}^{N} \lambda_i(\mathbf{\Sigma_X})}, \tag{10}$$

where $\lambda_i[\mathbf{\Sigma_X}]$ is the ith eigenvalue of the covariance matrix, $\mathbf{\Sigma_X}$ and the index i is ordered according to the descending value of the eigenvalues. Suppose that (8) is CS transformed by a linear measurement matrix $\mathbf{\Phi} \in \mathbb{R}^{M\times N}$ resulting in the compressed data matrix $\mathbf{Y} = \mathbf{\Phi X}$. Similarly, let the CS covariance matrix be $\mathbf{\Sigma_Y} = \mathbf{Y}\mathbf{Y}^T$, then the SNR for the compressed domain can be equivalently expressed as

$$\mathsf{SNR}_{\mathbf{Y}} = \frac{\sum_{i=1}^{\mathcal{K}} \lambda_i(\mathbf{\Sigma_Y})}{\sum_{i=\mathcal{K}+1}^{M} \lambda_i(\mathbf{\Sigma_Y})}, \tag{11}$$

where $\lambda_i[\mathbf{\Sigma_Y}]$ is the ith eigenvalue of the covariance matrix, $\mathbf{\Sigma_Y}$.

Theorem 1. *Let $\mathbf{\Phi}$ be an i.i.d. Gaussian random matrix whose entries follow the $\mathcal{N}(0, 1/\sqrt{M})$ distribution. Then, as $\mathbf{\Phi}$ varies, the SNR in the compressed domain is larger than that in the original domain on average and is given as*

$$\overline{\mathsf{SNR}}_{\mathbf{Y}} = \mathsf{SNR}_{\mathbf{X}} + \mathcal{O}\left\{\lambda_{\max}(\mathbf{\Sigma_X})\sqrt{\mathsf{M/N}}\right\}, \tag{12}$$

where $\lambda_{\max}$ is the largest eigenvalue of the correlation matrix, $\mathbf{\Sigma_X}$.

Firstly, it is proven that when the CS matrix is properly normalized by, for example, the choice of the Gaussian ensemble mentioned above, the total power is unchanged.

Lemma 1. *The traces of the sample covariance matrix in both original and compressed domains are the same, or equivalently*

$$\sum_i \lambda_i(\mathbf{\Sigma_X}) = \sum_i \lambda_i(\mathbf{\Sigma_Y}). \tag{13}$$

Proof. It remains to be shown that

$$\mathsf{Tr}[\mathbf{\Sigma_X}] = \mathsf{Tr}[\mathbf{\Sigma_Y}] = \mathsf{Tr}[\mathbf{\Phi}\mathbf{\Sigma_X}\mathbf{\Phi}^T], \tag{14}$$

where $\mathsf{Tr}[\cdot]$ is the trace operator. Suppose that the eigenvalue decomposition of $\mathbf{\Sigma_X}$ gives

$$\mathbf{\Sigma_X} = \mathbf{U}\mathbf{\Lambda_X}\mathbf{U}^T, \tag{15}$$

where $\mathbf{U}$ is a matrix that contains the eigenvectors and $\mathbf{\Lambda_X}$ is a diagonal matrix with eigenvalues in its diagonal. Then $\mathsf{Tr}[\mathbf{\Sigma_Y}]$ can be expanded as

$$\mathsf{Tr}[\mathbf{\Sigma_Y}] = \mathsf{Tr}[\mathbf{\Phi\Sigma_X\Phi}^T] = \mathsf{Tr}[\mathbf{\Phi'\Lambda_X\Phi'}^T], \tag{16}$$

where $\mathbf{\Phi'} = \mathbf{\Phi U}$. Since $\mathbf{U}$ is an unitary matrix, it follows that $\mathbf{\Phi'}$ also has the same statistical properties as $\mathbf{\Phi}$. This implies that the columns of $\mathbf{\Phi'}$ are also normalized to unit norm. Therefore,

$$\begin{aligned}\mathsf{Tr}[\mathbf{\Phi'\Lambda_X\Phi'}^T] &= \sum_{i=1}^{M}\sum_{j=1}^{N}\phi_{ij}'^2\lambda_j = \sum_{j=1}^{N}\left(\lambda_j\sum_{i=1}^{M}\phi_{ij}'^2\right)\\ &= \sum_{j=1}^{N}\lambda_j = \mathsf{Tr}[\mathbf{\Lambda_X}].\end{aligned} \tag{17}$$

Denote the residual power of a covariance matrix, $\mathbf{\Sigma}$ as

$$\mathcal{R}(\mathbf{\Sigma}) = \sum_{i=K+1}^{N}\lambda_i(\mathbf{\Sigma}). \tag{18}$$

Then Lemma 1 shows that to prove the main theorem, it is sufficient to concentrate on the *residual* power, $\mathcal{R}$ since the total power, $\mathcal{P}$ is unchanged in both domains and

$$\mathsf{SNR} = \frac{\mathcal{P}-\mathcal{R}}{\mathcal{R}} = \frac{\mathcal{P}}{\mathcal{R}} - 1. \tag{19}$$

□

Lemma 2. *The residual power, $\mathcal{R}$ is a concave function over the positive semi-definite cone.*

Proof. The Courant–Fischer minimax theorem (Shawe-Taylor et al., 2002) implies that the residual power is given as

$$\mathcal{R} = \min_{\dim(V)=K}\|\mathbf{P}_V^{\perp}\mathbf{X}\|_F^2, \tag{20}$$

where $\mathbf{P}_V^{\perp}\mathbf{X}$ is the projection of $\mathbf{X}$ by projection matrix $\mathbf{P}_V^{\perp}$ onto space that is perpendicular to V. By using the identity $\|\mathbf{A}\|_F^2 = \mathsf{Tr}[\mathbf{AA}^T]$, (20) can be rewritten as

$$\mathcal{R} = \min_{\dim(V)=K}\mathsf{Tr}[\mathbf{P}_V^{\perp}\mathbf{\Sigma_X}\mathbf{P}_V^{\perp T}]. \tag{21}$$

Let us now consider two covariance matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$. Suppose that

$$V_1 = \arg\min_{\dim(V)=K}\mathsf{Tr}[\mathbf{P}_V^{\perp}\mathbf{\Sigma}_1\mathbf{P}_V^{\perp T}], \tag{22}$$

$$V_2 = \arg\min_{\dim(V)=K}\mathsf{Tr}[\mathbf{P}_V^{\perp}\mathbf{\Sigma}_2\mathbf{P}_V^{\perp T}]. \tag{23}$$

By definition $\forall V : \dim(V) = K$

$$\mathsf{Tr}[\mathbf{P}_{V_1}^{\perp}\mathbf{\Sigma}_1\mathbf{P}_{V_1}^{\perp\ T}] \leqslant \mathsf{Tr}[\mathbf{P}_V^{\perp}\mathbf{\Sigma}_1\mathbf{P}_V^{\perp T}], \tag{24}$$

$$\mathsf{Tr}[\mathbf{P}_{V_2}^{\perp}\mathbf{\Sigma}_2\mathbf{P}_{V_2}^{\perp\ T}] \leqslant \mathsf{Tr}[\mathbf{P}_V^{\perp}\mathbf{\Sigma}_2\mathbf{P}_V^{\perp T}]. \tag{25}$$

Multiplying both inequalities with α and $(1-\alpha)$ respectively yields

$$\begin{aligned}&\alpha\mathsf{Tr}[\mathbf{P}_{V_1}^{\perp}\mathbf{\Sigma}_1\mathbf{P}_{V_1}^{\perp\ T}] + (1-\alpha)\mathsf{Tr}[\mathbf{P}_{V_2}^{\perp}\mathbf{\Sigma}_2\mathbf{P}_{V_2}^{\perp\ T}]\\ &\quad\leqslant \mathsf{Tr}[\mathbf{P}_V^{\perp}(\alpha\mathbf{\Sigma}_1 + (1-\alpha)\mathbf{\Sigma}_2)\mathbf{P}_V^{\perp T}]\end{aligned} \tag{26}$$

and the proof follows.

By using the concavity property of the residual power as shown in Lemma 2 yields the following

$$\begin{aligned}\mathsf{E}[\mathcal{R}(\mathbf{\Sigma_Y})] &\leqslant \mathcal{R}(\mathsf{E}[\mathbf{\Sigma_Y}]) = \mathcal{R}(\mathsf{E}[\mathbf{YY}^T]) = \mathcal{R}(\mathsf{E}[\mathbf{Y}^T\mathbf{Y}])\\ &= \mathcal{R}(\mathsf{E}[\mathbf{X}^T\mathbf{\Phi}^T\mathbf{\Phi X}^T]) = \mathcal{R}(\mathbf{X}^T\mathbf{X})\\ &= \mathcal{R}(\mathbf{XX}^T),\end{aligned} \tag{27}$$

where the $\mathsf{E}[\cdot]$ is the statistical expectation operator taken with respect to $\mathbf{\Phi}$ and $\mathsf{E}[\mathbf{\Phi}^T\mathbf{\Phi}] = \mathbf{I}$. This proves the first part of the theorem which asserts that the residual power is likely to be less in the compressed data. The second part of the theorem is concerned with the actual reduction in the residual power of the compressed data relative to the original data. It is noted that the absolute reduction is dependent on the actual realization of the CS matrix $\mathbf{\Phi}$ and it has to be computed numerically. However, for a particular family of the CS matrix, the order of magnitude of the reduction can be calculated, which translates the operating factors that influence the CS-based speech enhancement. This is what claimed by the second part of the theorem. To prove this, the following is decomposed as

$$\mathbf{Y}^T\mathbf{Y} = \mathbf{X}^T\mathbf{\Phi}^T\mathbf{\Phi X} = \mathbf{X}^T\mathbf{X} + \mathbf{X}^T\mathbf{EX} \tag{28}$$

where $\mathbf{E}$ is a matrix with zero diagonal entries and the off-diagonal entries being the cross-correlation between the columns of $\mathbf{\Phi}$. The immediate consequence of the Weyl theorem (Golub and Loan, 1996) is that

$$|\mathcal{R}(\mathbf{X}^T\mathbf{X}) - \mathcal{R}(\mathbf{Y}^T\mathbf{Y})| = \mathcal{O}(\|\mathbf{X}^T\mathbf{EX}\|_2), \tag{29}$$

where $\|\cdot\|_2$ denotes the spectral norm of a matrix. To bound the left hand side, the inequality on matrix norm is used

$$\|\mathbf{X}^T\mathbf{EX}\|_2 \leqslant \|\mathbf{X}^T\mathbf{X}\|_2\|\mathbf{E}\|_2. \tag{30}$$

Since $\|\mathbf{X}\|_2 = \sqrt{\lambda_{\max}(\mathbf{\Sigma_X})}$, then the first term on the right side of (30) is given as $\|\mathbf{X}^T\mathbf{X}\|_2 = \lambda_{\max}(\mathbf{\Sigma_X})$. Thus to complete the proof, the norm of $\mathbf{E}$ needs to be bounded. This follows from the result on concentration of Gaussian measure that with probability of at least $1-\delta$ yields

$$\|\mathbf{\Phi}^T\mathbf{\Phi}\|_2 \leqslant 1 + 2(\sqrt{M/N} + \sqrt{2\ln(1/\delta)/N}). \tag{31}$$

Since $\mathbf{E} = \mathbf{\Phi}^T\mathbf{\Phi} - \mathbf{I}$, it follows that

$$\|\mathbf{E}\|_2 = \mathcal{O}\left(\sqrt{M/N}\right). \tag{32}$$

Theorem 1 states that on average the SNR of the compressed sensing observation is greater than the original observation, i.e., $\overline{\mathsf{SNR}}_{\mathbf{Y}} \geqslant \mathsf{SNR}_{\mathbf{X}}$. It is worth mentioning that the SNR improvement is directly proportional to $\lambda_{\max}$ and M. Since the desired speech power is largely represented by $\lambda_{\max}$, CS acts as a salient feature detector, which automatically discriminates and separates speech signal from noise signal. Naturally, the higher the SNR, the better the discrimination becomes. Moreover, Theorem 1 suggests that the SNR improvement is likely to be significant for a large value of M rather than a small value. We also note that in the above analysis, the role of K is only to show possible average improvement in the SNR, but its exact value is not required in the proposed method. □

4. Experiments and discussions

4.1. Experimental settings

The settings for all the experiments were as follows:

- input sequences were six English utterances from the TIMIT database Garofolo (1988), consisted of three males and three females each with a five seconds duration. Note that all results presented are based on the average performance of the six utterances,
- three different types of noise from NOISEX-92, namely babble noise, destroyer operations room background noise and white noise,
- sampling frequency of 8 kHz,
- varying number of STFT points, L, with an oversampling factor of 4,
- CS reconstruction algorithm is based on $\ell 1$-regularized least squares problems (Kim et al., 2007).

4.2. Performance measures

Two objective measures are used to evaluate the performance of the proposed CS based algorithm, namely, the PESQ (ITU, 2001) and the average segmental SNR (Loizou, 2007). As reported in Hu (2008), the PESQ measure is more accurate in predicting speech distortion of the processed speech whereas the average segmental SNR reflects noise suppression more accurately. Both measures give a good indication on noise suppression and speech distortion. The PESQ measure however appears to be well rounded as it aims to predict the results of subjective listening tests. Also, the PESQ has been proven to be more reliable and closely correlated with the Mean Opinion Score (MOS). Extensive evaluation against actual MOS shows that PESQ yields an accurate evaluation both on speech distortion and noise distortion and it is a consistent measure for overall quality of the enhanced speech (Loizou, 2007).

4.3. Experimental results and discussions

4.3.1. Varying M/N and STFT points

Fig. 3 shows the PESQ scores for the proposed CS scheme as a function of compression ratio, M/N and the number of STFT points for three different types of noise. The noise profile has been chosen to include non-stationary

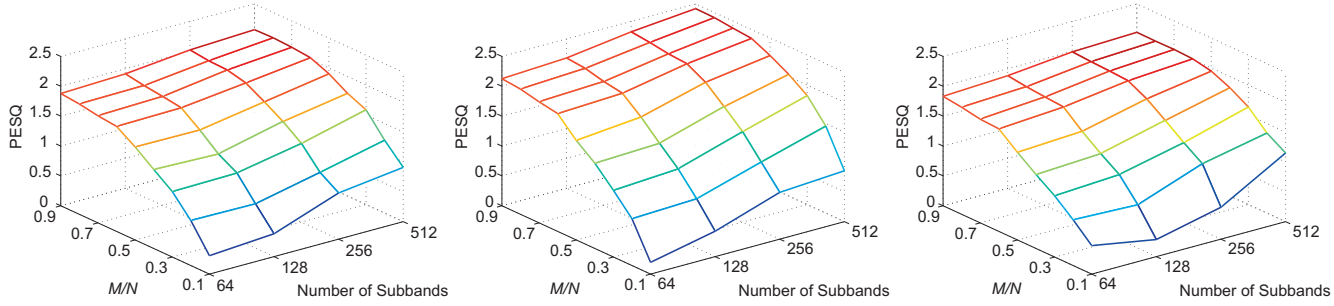


Fig. 3. The PESQ scores for the output of the proposed scheme for speech corrupted with three types of noise at SNR= 0 dB as a function of the compression ratio M/N and the number of subband.

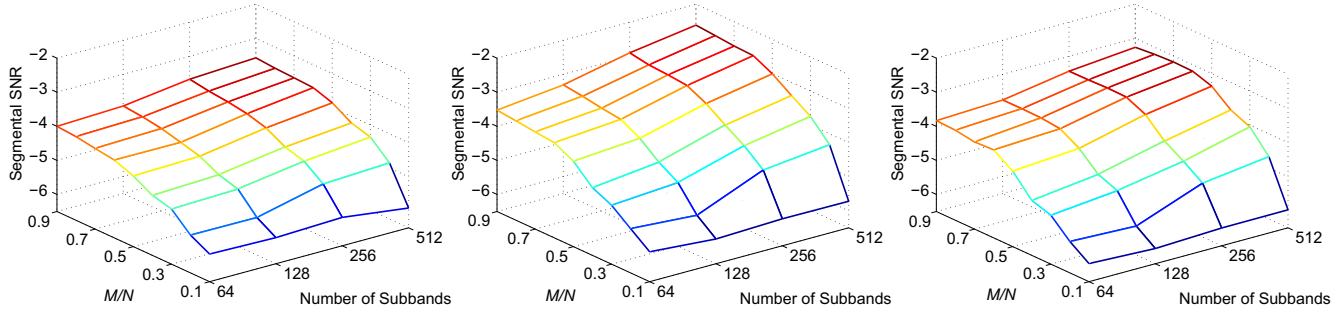
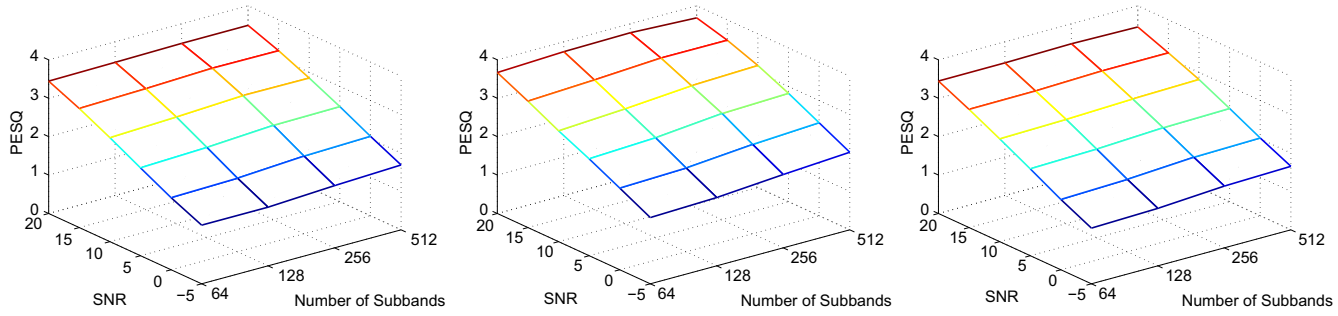


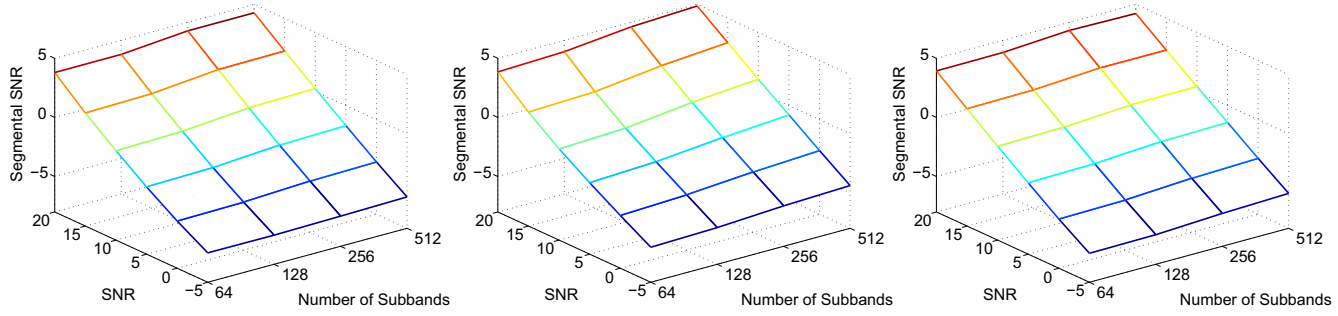
Fig. 4. The segmental SNR for the output of the proposed scheme for speech corrupted with three types of noise at SNR = 0 dB as a function of the compression ratio M/N and the number of subband.

noise such as babble, semi non-stationary noise namely, destroyer noise and the stationary white noise. The input SNR in this case has been set to 0 dB. Fig. 4 shows the corresponding average segmental SNR for varying M/N and the number of STFT points. From the results, it is clear that both the PESQ scores and average segmental SNR share the trend of increasing values as the M/N increases from 0.1 to 0.9 for all STFT points. This is in agreement with Theorem 1, which states that the SNR improvement is likely to be the case for large values of M. However, as the results show, greater number of STFT points results in slightly better PESQ scores and average segmental SNR. The experiment also shows that whilst there is improvement in SNR as M/N increases, the same is observed for the overall perceptual quality. Empirically, both figures indicate that the value of M/N needs to be at least 0.7 onwards for a reasonable improvement in SNR and PESQ.



(a)Babble noise (b)Destroyer noise (c)White noise

Fig. 5. The PESQ scores for the output of the proposed scheme for speech corrupted with three types of noise at $M/N = 0.9$ as a function of the SNR and the number of subband.



(a)Babble noise (b)Destroyer noise (c)White noise

Fig. 6. The segmental SNR for the output of the proposed scheme for speech corrupted with three types of noise at $M/N = 0.9$ as a function of the SNR and the number of subband.

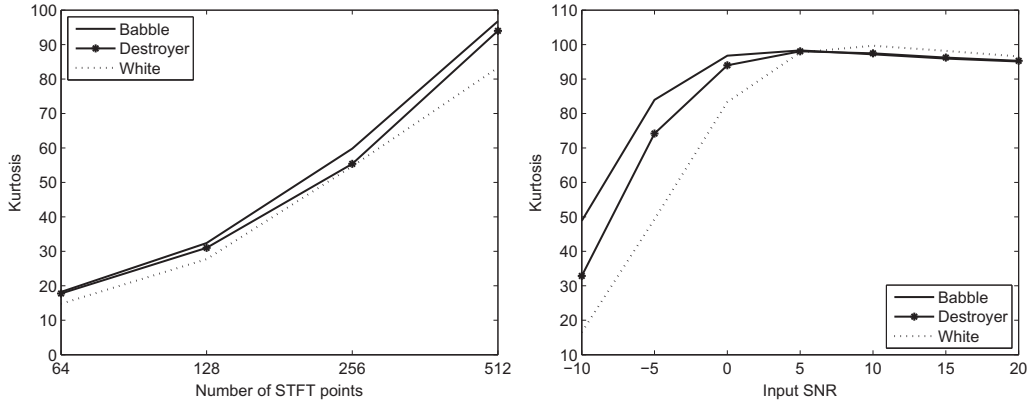


Fig. 7. The kurtosis values for the three types of noise computed by using spectral envelope at 1 kHz.

4.3.2. Varying Input SNR and STFT points

In the subsequent experiments, the value of the compression ratio is fixed M/N at 0.9 against a varying input SNR. Figs. 5 and 6 show the behavior of the proposed scheme as a function of input SNR and the number of STFT points. As expected, the results show that both the PESQ scores and the average segmental SNR have increasing values as the input SNR improves. Similar to the previous experiment, the plots reveal a higher number of STFT points results in a slight increase in both the PESQ scores and average segmental SNR. This is attributed to the increased sparsity of the speech observations as the number of STFT points increases (Pham et al., 2009; Gardner and Magnasco, 2006). To illustrate this, the kurtosis is used to indicate the level of sparseness in the noisy signal. The kurtosis is an approximate measure of sparsity and in this case, a small value of kurtosis indicates a highly non-sparse signal and conversely, a larger kurtosis value represents a sparser signal (Karvanen and Cichocki, 2003).

Fig. 7 plots the kurtosis for the spectral envelope of the three types of noisy speech with STFT points and different input SNR at 1 kHz. The numerical result clearly demonstrates that as the number of STFT increases, the kurtosis increases. This is in line with the previous observation where there is improvement in both the PESQ and the average segmental SNR as the number of STFT point is increased. Interestingly, Fig. 7(b) shows that as the input SNR improves, the kurtosis grows. The kurtosis remains in an equilibrium state upon reaching 5 dB before decaying slightly as the input SNR reaches 20 dB. Clearly, this suggests that there is a good level of sparsity between 0 dB and 15 dB. The kurtosis mildly explains the improvement in both the PESQ and segmental SNR as the level of sparsity changes.

4.3.3. A comparison against other conventional speech enhancement methods

The performance of the proposed CS based speech enhancement is compared against another compressive sensing based method (Wu et al., 2011), the Ephraim–Malah logarithm based minimum mean square error (logMMSE) (Ephraim and Malah, 1985) and the Wiener based method (Benesty et al., 2005). The noise estimates in Ephraim and Malah (1985) and Benesty et al. (2005) are estimated by using the minimum statistics approach (Martin, 2001). The implementation in Wu et al. (2011) has been modified to by using the theory of alternating direction method of multipliers (ADMM) for better convergence properties (Wahlberg et al., 2012). Figs. 8 and 9 give the comparison between the proposed method and other algorithms in terms of PESQ and the average segmental SNR, respectively.

In terms of PESQ, the proposed method outperforms the other algorithms. Fig. 9 shows that the CS approach achieves a higher segmental SNR improvement against the other methods for the input SNR range of −5–10 dB. However, the performance of the proposed method is comparable to the performance of the Wiener solution for SNR greater than 10 dB and above. Interestingly, it is observed that the average segmental SNR improvement for the proposed method is almost linear with the input SNR. In particular, the suggested method consistently obtains better performance for both PESQ and segmental SNR for input SNR 0db up to 10 dB, which is the most commonly observed SNR range for general noisy speech applications. Significantly, the results show that the proposed scheme

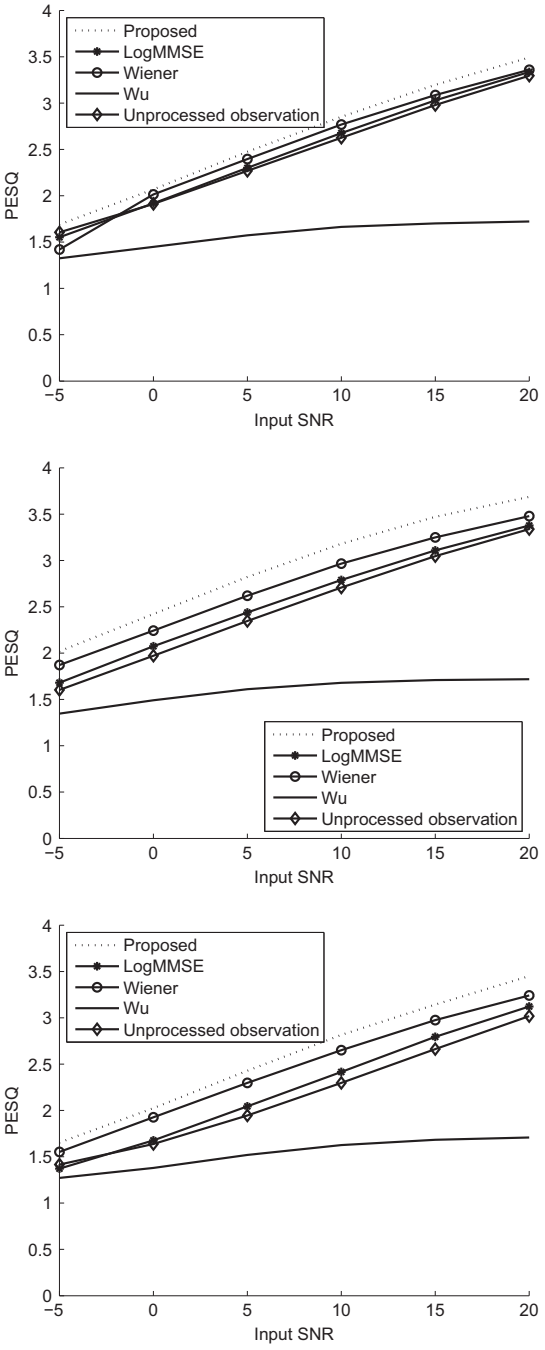


Fig. 8. The PESQ scores for the unprocessed observation and the output of the proposed scheme for speech corrupted with babble noise (top), destroyer noise (middle) and white noise (bottom) for different input SNR for $M/N = 0.9$ and 512 STFT points.

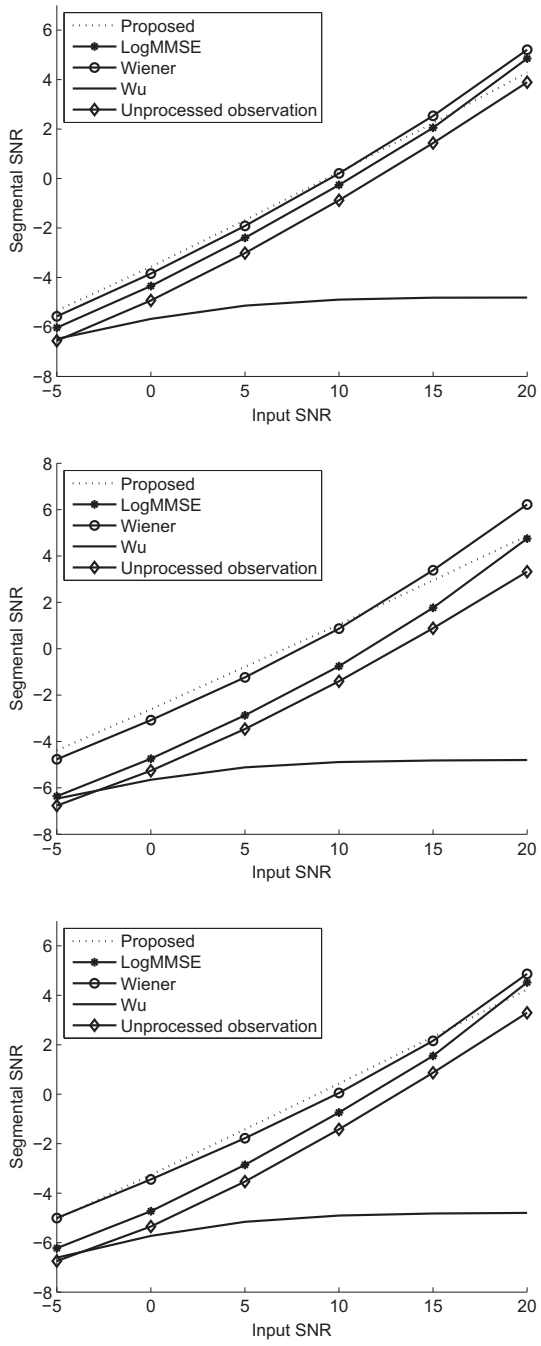


Fig. 9. The segmental SNR for the unprocessed observation and the output of the proposed scheme for speech corrupted with babble noise (top), destroyer noise (middle) and white noise (bottom) for different input SNR for $M/N = 0.9$ and 512 STFT points.

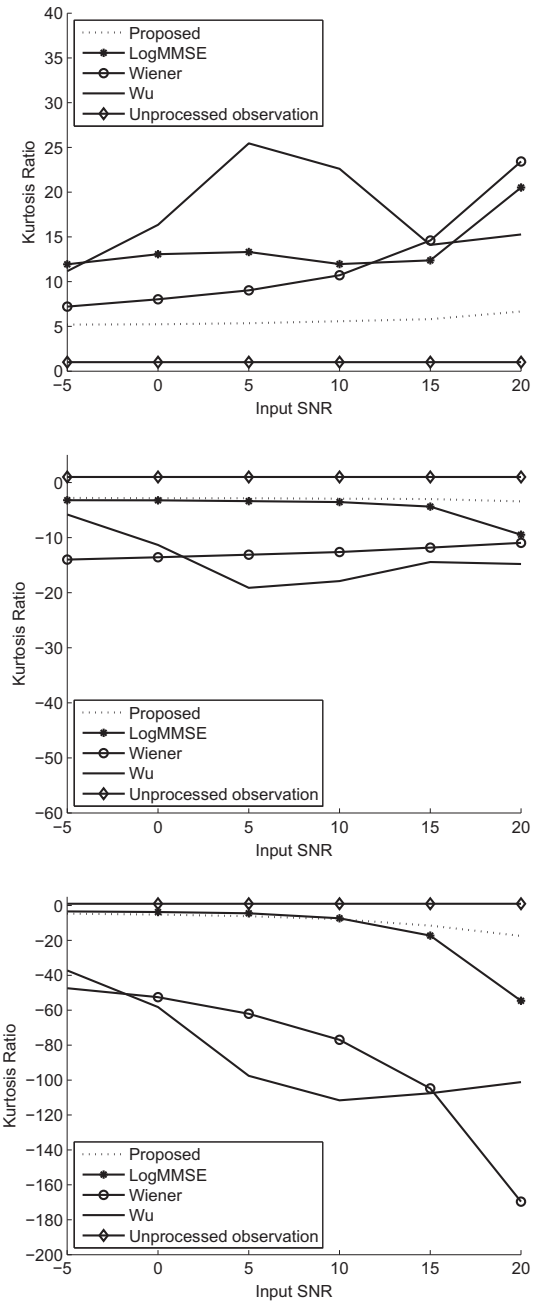


Fig. 10. The kurtosis ratio for the unprocessed observation and the output of the proposed scheme for speech corrupted with babble noise (top), destroyer noise (middle) and white noise (bottom) for different input SNR for $M/N = 0.9$ and 512 STFT points.

improves the PESQ of the input signal across the input SNR independent of noise type.

In terms of noise suppression, the proposed scheme gives improvement in its segmental SNR as the input SNR reaches > 15 dB. This shows that under favorable input SNR conditions the CS will perform very little or no noise attenuation. This is because as the input SNR improves, there exists very little non-sparse components in the input signal. In such situations, the majority is the sparse speech components (see Fig. 7). By virtue of CS sparse reconstruction property, the overall signal is more or less remains unattenuated. To better illustrate the point, the kurtosis ratio was used to compare the amount of artifacts present in the processed signal. Kurtosis ratio is an indicative metric to measure the amount of musical artifacts in the processed signal (Uemura et al., 2008). The measure has been shown to be strongly correlated with the human perception of musical noise (Miyazaki et al., 2012). Fig. 10 shows the kurtosis ratio for the various

methods over a range of input SNR and different types of noise. Ideally, if there is no artifacts introduced in the output, the kurtosis ratio should yield a unity value. Results show that the proposed method has consistently low values of kurtosis ratio under the various settings. More importantly, the kurtosis ratio of the proposed method does not fluctuate as the SNR improves.

As discussed, the proposed CS based speech extraction is only dependent on the sparse and non-sparse components in the signal. Under poor SNR situations, sparse speech components may be heavily masked by the noise. As such, the discrimination between noise (non-sparse) and speech (sparse) become less obvious for CS. On the other extreme of SNR = 20 dB, there is hardly any noise for the CS to suppress. Again this results in very little distinction between the noise and speech. The very nature of this characteristic helps to reduce any potential speech artifacts caused by "over processing" as observed by other methods (see the kurtosis ratio comparison in Fig. 10). Notwithstanding, the proposed method will fail to perform under sparse noise such as impact noise e.g., hammering etc. Such sparse noise will be passed through the system as CS will actively reconstruct the sparse components.

It is worth pointing out that the proposed CS scheme is very different compared to wavelet based method in Wu et al. (2011). Both wavelet transform and STFT are closely related in the sense that both transformations capture information in terms of time and frequency. The major difference between the two is that wavelet transform allows the flexibility of using a size adjustable window as opposed a fixed window in STFT. Generally, wavelet transform is useful from a psychoacoustic perspective as it has close relation to many perceptual scales such as pitch and loudness (Pinter, 1996). In this case however, the transformation is not used for analysis rather as a transformation platform for CS to identify the sparse and non-sparse information. Thus, the results so far indicate no immediate benefit in using wavelet transformation over the STFT transformation (see Figs. 8–10). Regardless, this opens room for future investigation on the impact of various transformations for CS.

5. Conclusion

A new approach to speech enhancement based on CS is presented. The proposed scheme makes use of the strength of CS in preserving only the sparse components to perform denoising. This is possible since speech signal is sparse in its time-frequency representation and CS can be readily exploited to reconstruct only speech components with an overwhelming probability. The proposed CS based method bypasses the need for a VAD and acts as an intrinsic salient speech extractor. Its denoising capability is shown in terms of PESQ and segmental SNR improvement for a wide range of input SNR. Whilst the performance of the proposed method is not all encompassing, the paper provides an incremental contribution and insight as to how the popular CS can be used to perform speech enhancement. Future work includes theoretical studies on the influence of various transformations on CS speech enhancement capability.

References

Benesty, J., Makino, S., Chen, J., 2005. Speech Enhancement Signals and Communication Technology. Springer-Verlag, Berlin.

Boll, S.F., 1979. Supression of acoustic noise in speech using spectral subtraction. IEEE Transactions on Acoustics, Speech and Signal Processing ASSP 27 (2), 113–120.

Boufounos, P., Duarte, M.F., Baraniuk, R.G., 2007. Sparse signal reconstruction from noisy compressive measurements using cross validation. IEEE Workshop on Statistical Signal Processing, 299–303.

Brandstein, M., Ward, D. (Eds.), 2001. Microphone Arrays: Signal Processing Techniques and Applications. Digital Signal Processing. Springer-Verlag, Berlin.

Candès, E.J., 2006. Compressive sampling, Proceedings of the International Congress of Mathematicians, Madrid, Spains.

Candès, E.J., Tao, T., 2006. Near-optimal signal recovery from random projections: universal encoding strategies? IEEE Transactions on Information Theory 52 (12), 5406–5425.

Candès, E.J., Wakin, M.B., 2008. An introduction to compressive sampling. IEEE Signal Processing Magazine, 21–30.

Candès, E., Romberg, J., Tan, T., 2006. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. IEEE Transactions on Information Theory 52 (2), 489–509.

Chen, S., Donoho, D., Saunders, M., 1998. Atomic decomposition by basis pursuit. SIAM Journal on Scientific Computing 20 (1), 33–61.

Christensen, M.G., Stergaard, J., Jensen, S.H., 2009. On compressed sensing and its applications to speech and audio signals. In: Proceedings of Asilomar Conference on Signals, Systems and Computers, pp. 356–360.

Cohen, I., 2003. Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging. IEEE Transactions on Speech and Audio Processing 11 (5), 466–475.

Dam, H.Q., Low, S.Y., Nordholm, S., Dam, H.H., 2004. Adaptive microphone array with noise statistics updates. IEEE International Symposium on Circuits and Systems 3, 433–436.

Davis, A., Low, S.Y., Nordholm, S., Grbic, N., 2005. A subband space constrained beamformer incorporating voice activity detection. IEEE International Conference on Acoustics, Speech and Signal Processing 3, 65–68.

Donoho, D.L., 2006. Compressed sensing. IEEE Transactions on Information Theory 52 (4), 1289–1306.

Ephraim, Y., Malah, D., 1985. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Transactions on Acoustics, Speech and Signal Processing ASSP-33, 443–445.

Gardner, T.J., Magnasco, M.O., 2006. Sparse time-frequency representations. Proceedings National Academy of Science 103, 6094–6099.

Garofolo, J.S., 1988. Getting started with the DARPA TIMIT CD-ROM: an acoustic phonetic continuous speech database. National Institute of Standards Technology (NIST).

Ghosh, P.K., Tsiartas, A., Narayanan, S., 2011. Robust voice activity detection using long-term signal variability. IEEE Transactions on Audio, Speech and Language Processing 19 (3), 600–613.

Giacobello, D., Christensen, M.G., Murthi, M.N., Jensen, S.J., Moonen, M., 2012. Sparse linear prediction and its applications to speech processing. IEEE Transactions on Audio, Speech, and Language Processing 20 (5), 1644–1657.

Golub, G.H., Loan, C.F.V. (Eds.), 1996. Matrix Computations, third ed. JHU Press, MD, USA.

Griffin, A., Hirvonen, T., Tzagkarakis, C., Mouchtaris, A., Tsakalides, P., 2011. Single-channel and multi-channel sinusoidal audio coding using compressed sensing. IEEE Transactions on Audio, Speech, and Language Processing 19 (5), 1382–1395.

Hoshuyama, O., Sugiyama, A., Hirano, A., 1999. A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters. IEEE Transactions on Signal Processing 47 (10), 2677–2684.

Hu, Y., 2008. Evaluation of objective quality measures for speech enhancement. IEEE Transactions on Audio, Speech and Language Processing 16 (1), 229–238.

ITU, 2001. Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs. ITU Recommendation, 862.

Jancovic, P., Zou, X., Kokuer, M., 2012. Speech enhancement based on sparse code shrinkage employing multiple speech models. Speech Communication 54 (1), 108–118.

Karvanen, J. Cichocki, A., 2003. Measuring sparseness of noisy signals. In: Proceedings of the Symposium of Independent Component Analysis and Blind Signal Separation, pp. 125–128.

Kim, S.J., Koh, K., Lustig, M., Boyd, S., Gorinevsky, D., 2007. A method for large-scale l1-regularized least squares. IEEE Journal on Selected Topics in Signal Processing 4 (1), 606–617.

Kokkinakis, K., Loizou, P., 2008. Using blind source separation techniques to improve speech recognition in bilateral cochlear implant patients. Journal of the Acoustical Society of America 123 (4), 2379–2390.

Loizou, P., 2007. Speech Enhancement: Theory and Practice. CRC Press, Taylor and Francis, Boca Raton, FL, USA.

Lotter, T., Benien, C., Vary, P., 2003. Multichannel direction-independent speech enhancement using spectral amplitude estimation. EURASIP Journal on Applied Signal Processing 2003, 1147–1156.

Low, S.Y., Nordholm, S., 2005. A blind approach to joint echo and noise cancellation. IEEE International Conference on Acoustics, Speech and Signal Processing 3, 69–72.

Low, S.Y., Grbic, N., Nordholm, S., 2002. Robust microphone array using subband adaptive beamformer and spectral subtraction. IEEE International Conference on Communication Systems 2, 1020–1024.

Lu, C.T., 2011. Enhancement of single channel speech using perceptual-decision-directed approach. Speech Communication 53 (4), 495–507.

Martin, R., 2001. Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Transactions on Speech and Audio Processing 9 (5), 504–512.

Miyazaki, R., Saruwatari, H., Inoue, T., Takahashi, Y., Shikano, K., Kondo, K., 2012. Musical-noise-free speech enhancement based on optimized iterative spectral subtraction. IEEE Transactions on Audio, Speech, and Language Processing 20 (7), 2080–2094.

O'Shaughnessy, D., 2000. Speech Communications: Human and Machine. IEEE Press, NJ, USA.

Paliwal, K., Wojcicki, K., Schwerin, B., 2010. Single-channel speech enhancement using spectral subtraction in the short-time modulation domain. Speech Communication 52 (5), 450–475.

Pham, D.T., El-Chami, Z., Guérin, A., Servière, C., 2009. Modeling the Short Time Fourier Transform Ratio and Application to Underdetermined Audio Source Separation. In: Lecture Notes in Computer Science, vol. 5441/2009. Springer, Berlin.

Pinter, I., 1996. Perceptual wavelet-representation of speech signals and its application to speech enhancement. Computer Speech and Language 10 (1), 1–22.

Principi, E., Cifani, S., Rotili, R., Squartini, S., Piazza, F., 2010. Comparative evaluation of single-channel MMSE-based noise reduction schemes for speech recognition. Journal of Electrical and Computer Engineering 2010, 1–6 (Article ID 962103).

Rachlin, Y., Baron, D., 2008. The secrecy of compressive sensing measurements. In: 46th Allerton Conference on Communication, Control, and Computing.

Shawe-Taylor, J., Cristianini, N., Kandola, J.., 2002. On the concentration of spectral properties. Advances in Neural Information Processing Systems 14, 511–517.

Sreenivas, T.V., Kleijn, W.B., 2009. Compressive sensing for sparsely excited speech signals. Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 4125–4128.

Tropp, J., Gilbert, A., 2007. Signal recovery from random measurements via orthogonal matching pursuit. IEEE Transactions on Information Theory 53 (12), 4655–4666.

Uemura, Y., Takahashi, Y., Saruwatari, H., Shikano, K., Kondo, K., 2008. Automatic optimization scheme of spectral subtraction based on musical noise assessment via higher-order statistics. In: International Workshop on Acoustics, Echo and Noise Control.

Veen, B.D.V., Buckley, K.M., 1988. Beamforming: a versatile approach to spatial filtering. IEEE Acoustics, Speech and Signal Processing Magazine 5, 4–24.

Wahlberg, B., Boyd, S., Annergren, M., Wang, Y., 2012. An ADMM algorithm for a class of total variation regularized estimation problems. In: IFAC Symposium on System Identification, vol. 1, p. 16.

Wu, D., Zhu, W.P., Swamy, M.N.S., 2011. A compressive sensing method for noise reduction of speech and audio signals. In: IEEE 54th International Midwest Symposium on Circuits and Systems (MWSCAS), pp. 1–4.

Yang, J., 1993. Frequency domain noise suppression approaches in mobile telephone systems. IEEE International Conference on Acoustics, Speech and Signal Processing 2, 363–366.

Yu, T., Hansen, J.H.L., 2009. A speech presence microphone array beamformer using model based speech presence probability estimation. IEEE International Conference on Acoustics, Speech, and Signal Processing, 213–216.