EURASIP Journal on Advances in Signal Processing
a SpringerOpen Journal

RESEARCH Open Access

Study on adaptive compressed sensing & reconstruction of quantized speech signals

Ji Yunyun[1*] and Yang Zhen[2]

Abstract

Compressed sensing (CS) is a rising focus in recent years for its simultaneous sampling and compression of sparse signals. Speech signals can be considered approximately sparse or compressible in some domains for natural characteristics. Thus, it has great prospect to apply compressed sensing to speech signals. This paper is involved in three aspects. Firstly, the sparsity and sparsifying matrix for speech signals are analyzed. Simultaneously, a kind of adaptive sparsifying matrix based on the long-term prediction of voiced speech signals is constructed. Secondly, a CS matrix called two-block diagonal (TBD) matrix is constructed for speech signals based on the existing block diagonal matrix theory to find out that its performance is empirically superior to that of the dense Gaussian random matrix when the sparsifying matrix is the DCT basis. Finally, we consider the quantization effect on the projections. Two corollaries about the impact of the adaptive quantization and nonadaptive quantization on reconstruction performance with two different matrices, the TBD matrix and the dense Gaussian random matrix, are derived. We find that the adaptive quantization and the TBD matrix are two effective ways to mitigate the quantization effect on reconstruction of speech signals in the framework of CS.

Keywords: Compressed sensing, Speech signals, Sparsity, Sparsifying matrix, Sensing matrix, Quantization

1 Introduction

In recent years, compressed sensing (CS) [1-4] has been a new and popular paradigm of signal acquisition and compression in applied science and engineering such as image processing, wireless communication, magnetic resonance imaging (MRI) and so on. In contrast with the conventional Nyquist sampling theorem, CS theory demonstrates that a sparse signal can be exactly recovered through far fewer projections, providing that the sensing matrix is highly incoherent with the sparsifying matrix.

As an important branch of signal processing, speech signal processing has achieved a considerable development in past decades. In addition, the application of CS theory to the field of speech signal processing is becoming a rising research focus. In [5,6], the sparsity of the residual excitation is utilized to construct sparsifying matrices for voiced speech signals. However, in the aforementioned two literatures, the sparsifying matrix constructed using the impulse response for voiced speech is impractical for its dependence on the currently reconstructed signal itself. Therefore, a codebook of impulse response vectors generated from the training speech data is proposed as the sparsifying matrix in [5].

This work also constructs an adaptive sparsifying matrix for voiced speech based on the quasi-periodicity during voiced segments. And this adaptive sparsifying matrix is a kind of symmetric cyclic matrix which is generated on the basis of the long term prediction. Therefore, this adaptive sparsifying matrix is dependent on the previously reconstructed signal instead of the current signal.

Then, a kind of CS matrix called two-block diagonal (TBD) matrix is constructed for voiced speech signals. The concentration inequality of the TBD matrix is simply demonstrated in Section 4. Subsequently, we can find that the TBD matrix satisfies the restricted isometry property (RIP) [7] according to a theorem in [8].

The third key point of this work to be discussed is quantization. It is well known that analog signals should be sampled, quantized and then encoded before transmission.

* Correspondence: jiyunyun1988@126.com
[1]College of Communication and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing, Jiangsu 210003, China
Full list of author information is available at the end of the article

Thus, quantization of CS projections is of great importance. The distortion rate performance and some measures to mitigate the impact of quantization noise on reconstruction have been considered in [9-15]. In this paper, we apply uniform scalar quantization to the measurements of the speech signal and quantitatively show that how adaptive quantization affects the reconstruction quality compared with the nonadaptive quantization. In addition, we find that the TBD matrix is more robust to the quantization noise than the dense Gaussian matrix based on the fact that the TBD matrix can effectively restricted the impact of quantization noise on reconstruction of speech signals.

The rest of the paper is organized as follows. In Section 2, we briefly review the principle of CS. Section 3 presents the construction of an adaptive sparsifying matrix for voiced speech signals. In Section 4, a sensing matrix is constructed for voiced speech signals. And in Section 5, the effect of quantization of projections on reconstruction is discussed. Section 6 then concludes our work.

2 Compressed sensing background

Supposed that a vector $x = [\, x(1) \quad x(2) \quad \cdots \quad x(N)\,]^T$ can be represented as a linear combination of some basis vectors $\{\varphi_1 \; \varphi_2 \; \cdots \; \varphi_N\}$, we have

$$x = \Psi\theta = \sum\nolimits_{i=1}^{N} \varphi_i \theta(i) \tag{1}$$

where $\Psi = [\varphi_1 \varphi_2 \cdots \varphi_N]$ and $\theta = [\, \theta(1) \;\; \theta(2) \;\; \cdots \quad \theta(N)\,]^T$. If the number of nonzero entries of θ which can be represented as $\|\theta\|_{l_0}$ satisfies

$$\|\theta\|_{l_0} \leq K \tag{2}$$

x is considered to be K-sparse with respect to Ψ. Then Ψ is called a sparsifying matrix.

And a matrix $\Phi \in R^{M\times N}$ can be employed to project a N-dimensional vector onto a M-dimensional subspace. Then, we can acquire a low-dimensional vector y and we have

$$y = \Phi x = \Phi\Psi\theta = A\theta \tag{3}$$

where Φ is called the sensing matrix and A is named the CS matrix. It is required that the CS matrix must satisfy certain conditions for effective reconstruction of the coefficient vector θ. And RIP is a sufficient condition for effective reconstruction. In the following, we firstly recall the definition of restricted isometry constant.

Definition 1(Restricted isometry constant) ([7,16]). The restricted isometry constant δ_K of matrix A is defined as the smallest quantity such that

$$(1-\delta_K)\|\theta\|_{l_2}^2 \leq \|A\theta\|_{l_2}^2 \leq (1+\delta_K)\|\theta\|_{l_2}^2 \tag{4}$$

holds for all K-sparse vectors. And the matrix A is said to satisfy K-order RIP with prescribed constant δ_K.

Although Eq. (3) is ill-conditioned, it is demonstrated in [16] that as

$$\delta_{2K} < \sqrt{2} - 1 \tag{5}$$

we can find the exact solution for K-sparse vector θ from

$$min\|\theta\|_{l_1} \;\text{ s.t. }\; y = A\theta \tag{6}$$

which is called BP algorithm [17].

When the measurement vector is corrupted by bounded noise and can be represented as

$$y = A\theta + t \tag{7}$$

we can employ the basis pursuit denoising (BPDN) algorithm [17]

$$min\|\theta\|_{l_1} \,..s.t.\|y - A\theta\|_{l_2} \leq \varepsilon \tag{8}$$

to achieve effective reconstruction, where ε is an upper bound of l_2-norm of the noise vector t. A theorem introducing the reconstruction performance of BPDN algorithm in detail is presented in Section 5 which is firstly formulated in [16].

Another kind of reconstruction algorithms are named greedy pursuit algorithms including orthogonal matching pursuit (OMP) [18], subspace pursuit (SP) [19], stagewise orthogonal matching pursuit (StOMP) [20], regularized orthogonal matching pursuit (ROMP) [21] and sparsity adaptive matching pursuit (SAMP) [22].

3 Sparsity and sparsifying matrix of speech signals

Speech signals, because of their natural characteristics such as the rich frequency components, cannot meet the definition of exact sparsity in a strict sense. And speech signals can only be regarded as compressible with a lot of nonzero but small coefficients in some basis like DCT. It is known that sparsity of signals is the precondition of CS. Thus, in the following, we firstly construct an adaptive sparsifying matrix for voiced segments.

3.1 Sparsifying matrix and sparsity of voiced speech

The sparsity of voiced speech has some bearing on its quasi-periodicity. In conventional speech signal coding system, the long-term prediction is always used to minimize the mean-square error between the predicted and the true values of voiced speech signals [23]. Supposing that a voiced segment includes several pitch periods (the reciprocal of vibration frequency of vocal cords) and x_i and x_{i+1} denote the vectors of the i^{th} period and the $(i+1)^{th}$ period respectively, according to the principle of long-term prediction, we have

$$\begin{aligned} x_{i+1}(n) &\approx \beta(-1)x_i(n-\mathrm{T}+1)+\beta(0)x_i(n-\mathrm{T}) \\ &+\beta(1)x_i(n-\mathrm{T}-1)(n=i\mathrm{T}, i\mathrm{T}+1, \cdots(i+1)\mathrm{T}-1) \end{aligned} \tag{9}$$

where T denotes the number of samples in a pitch period, namely, pitch period. In terms of the quasi-periodicity of voiced speech, some assumptions are made below.

As for the first point and the last point in the $(i+1)^{th}$ period, we have $x_{i+1}(i\mathrm{T}) \approx \beta(-1)x_i((i-1)\mathrm{T}+1)+\beta(0)x_i((i-1)\mathrm{T})+\beta(1)x_i((i-1)\mathrm{T}-1)$ and

$$\begin{aligned} x_{i+1}((i+1)\mathrm{T}-1) &\approx \beta(-1)x_i(i\mathrm{T})+\beta(0)x_i(i\mathrm{T}-1) \\ &+\beta(1)x_i(i\mathrm{T}-2). \end{aligned}$$

However, the time-domain range of x_i is from $(i-1)\mathrm{T}$ to $i\mathrm{T}-1$. Therefore, in the duration of x_i, we make artificially $x_i((i-1)\mathrm{T}-1)$ and $x_i(i\mathrm{T})$ in Eq. (9) equal to $x_i(i\mathrm{T}-1)$ and $x_i((i-1)\mathrm{T})$ and then we have

$$\begin{bmatrix} x_{i+1}(i\mathrm{T}) \\ x_{i+1}(i\mathrm{T}+1) \\ x_{i+1}(i\mathrm{T}+2) \\ \vdots \\ x_{i+1}((i+1)\mathrm{T}-1) \end{bmatrix} \approx \begin{bmatrix} x_i((i-1)\mathrm{T}) & x_i((i-1)\mathrm{T}+1) & \cdots & x_i(i\mathrm{T}-1) \\ x_i((i-1)\mathrm{T}+1) & x_i((i-1)\mathrm{T}+2) & \cdots & x_i((i-1)\mathrm{T}) \\ x_i((i-1)\mathrm{T}+2) & x_i((i-1)\mathrm{T}+3) & \cdots & x_i((i-1)\mathrm{T}+1) \\ \vdots & \vdots & \vdots & \vdots \\ x_i(i\mathrm{T}-1) & x_i((i-1)\mathrm{T}) & \cdots & x_i(i\mathrm{T}-2) \end{bmatrix} \begin{bmatrix} \beta(0) \\ \beta(-1) \\ 0 \\ \vdots \\ \beta(1) \end{bmatrix} \tag{10}$$

Furthermore, in terms of Eq. (10), we establish that

$$\Psi = \begin{bmatrix} x_i((i-1)\mathrm{T}) & x_i((i-1)\mathrm{T}+1) & x_i((i-1)\mathrm{T}+2) & \cdots & x_i(i\mathrm{T}-1) \\ x_i((i-1)\mathrm{T}+1) & x_i((i-1)\mathrm{T}+2) & x_i((i-1)\mathrm{T}+3) & \cdots & x_i((i-1)\mathrm{T}) \\ x_i((i-1)\mathrm{T}+2) & x_i((i-1)\mathrm{T}+3) & x_i((i-1)\mathrm{T}+4) & \cdots & x_i((i-1)\mathrm{T}+1) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_i(i\mathrm{T}-1) & x_i((i-1)\mathrm{T}) & x_i((i-1)\mathrm{T}+1) & \cdots & x_i(i\mathrm{T}-2) \end{bmatrix} \tag{11}$$

$$\beta = [\,\beta(0) \quad \beta(-1) \quad 0 \quad \cdots \quad \beta(1)\,]^T \tag{12}$$

and

$$x_{i+1} \approx \Psi\beta \tag{13}$$

Thus, the vector β is called the coefficient vector of x_{i+1} with respect to the adaptive sparsifying matrix Ψ and we have

$$\|\beta\|_{l_0} = 3. \tag{14}$$

It is obvious that x_{i+1} is approximately sparse with respect to the matrix Ψ defined in Eq. (11) which is composed of components of x_i. Thus, at the decoder, the recovered signal of the current pitch period can be used to constitute a sparsifying matrix for the signal of next pitch period.

As the adaptive sparsifying matrix Ψ is a real symmetric cyclic matrix, we can get its eigenvalues [24] which are denoted by $\lambda_m(m = 0, 1 \cdots \mathrm{T}-1)$. We define

$$f(z) = \sum\nolimits_{l=0}^{\mathrm{T}-1} x_i((i-1)\mathrm{T}+l)z^l \tag{15}$$

and

$$\omega = e^{j\frac{2\pi}{\mathrm{T}}} \tag{16}$$

Supposed that T is even, we have

$$\lambda_0 = f(1) \tag{17}$$

$$\lambda_m = |f(\omega^m)| \quad \left(m = 1, 2 \cdots \frac{\mathrm{T}}{2}-1\right) \tag{18}$$

$$\lambda_m = -\left| f\left(\omega^{m-\frac{\mathrm{T}}{2}+1}\right)\right| \quad \left(m = \frac{\mathrm{T}}{2}, \frac{\mathrm{T}}{2}+1, \cdots \mathrm{T}-2\right) \tag{19}$$

and

$$\lambda_{\mathrm{T}-1} = f(-1) \tag{20}$$

Otherwise, when T is odd, we have

$$\lambda_0 = f(1) \tag{21}$$

$$\lambda_m = |f(\omega^m)| \left(m = 1, 2 \cdots \frac{\mathrm{T}-1}{2}\right) \tag{22}$$

and

$$\lambda_m = -\left| f\left(\omega^{m-\frac{\mathrm{T}-1}{2}}\right)\right| \left(m = \frac{\mathrm{T}-1}{2}+1, \frac{\mathrm{T}-1}{2}+2, \cdots \mathrm{T}-1\right). \tag{23}$$

Moreover, we can recall the DFT transform of x_i which can be expressed as $X_i(k) = \sum_{l=0}^{\mathrm{T}-1} x_i((i-1)\mathrm{T}+l)\omega^{-lk}$. Then we can obtain the relation between the eigenvalues of the adaptive sparsifying matrix Ψ and the spectrum of the signal x_i. When T is even, we have

$$\lambda_m = |X_i(m)| \quad \left(m = 1, 2 \cdots \frac{\mathrm{T}}{2}-1\right) \tag{24}$$

and

$$\lambda_m = -\left| X_i\left(m - \frac{\mathrm{T}}{2}+1\right)\right| \left(m = \frac{\mathrm{T}}{2}, \frac{\mathrm{T}}{2}+1, \cdots \mathrm{T}-2\right) \tag{25}$$

And when T is odd, we have

$$\lambda_m = |X_i(m)| \quad \left(m = 1, 2 \cdots \frac{\mathrm{T}-1}{2}\right) \tag{26}$$

and

$$\lambda_m = -\left| X_i\left(m - \frac{\mathrm{T}-1}{2}\right)\right| \times \left(m = \frac{\mathrm{T}-1}{2}+1, \frac{\mathrm{T}-1}{2}+2, \cdots \mathrm{T}-1\right) \tag{27}$$

And we define

$$g = \lambda_0 \lambda_{\mathrm{T}-1} \prod_{m=1}^{\mathrm{T}-2} \lambda_m \tag{28}$$

Moreover, if $g \neq 0$, the adaptive sparsifying matrix Ψ defined in Eq. (11) is invertible.

Although this adaptive sparsifying matrix is not a canonical basis in a conventional sense, it has two advantages. On the one hand, as an adaptive sparsifying matrix which is constructed by the recovered signal, the decoder doesn't need additional storage space and at the encoder it is not necessary to spend time attaining the training data to construct the codebook and to transmit it to the decoder such as the approach proposed in [5]. On the other hand, the approximate sparsity of speech signals with respect to this adaptive sparsifying matrix is superior to the DCT basis, which can be verified by the comparison of reconstruction performance between the adaptive sparsifying matrix and the DCT basis in the subsection 3.3.

3.2 Sparsity of unvoiced speech signals

The transform coefficients based on the spectral characteristics of unvoiced speech signals are nearly uniformly distributed in the frequency domain with no obvious decay. Consequently, the sparsity of unvoiced speech signal with respect to the DCT basis is undesirable. Furthermore, we have not found a satisfactory sparsifying matrix for unvoiced speech signals. Therefore, the usual practice in the framework of CS is to apply the scheme to entire speech signals and not to distinguish voiced speech signals and unvoiced speech signals in advance. Moreover, we find that the overall performance has not been greatly influenced, which can be verified by the simulation results in the following subsection. The reason is that the proportion of voiced speech is more than seventy percent and voiced speech bears dominating information of speech. Certainly, it is of great significance for us to seek to construct a basis or a redundant dictionary for unvoiced speech signals, which is the focus of our future work.

3.3 Simulation

Some simulation results are illustrated in this subsection to show the performance of the adaptive sparsifying matrix. The testing speech signals are sampled at 16*K*Hz with the length of a frame N=320. There are 152 frames including 135 frames of voiced speech and 17 frames of unvoiced speech. And the sensing matrix used in this section is the dense Gaussian random matrix whose entries are i.i.d Gaussian random variables with mean zero and variance $\frac{1}{M}$. And BP algorithm is used in this subsection to achieve reconstruction of speech signals.

It should be pointed out that the first pitch period in each frame is recovered with respect to the DCT basis. Moreover, the following pitch periods are compressed with the same compression rate and at the decoder we achieve reconstruction with respect to the adaptive sparsifying matrix. The compression rate is defined as

$$u = \frac{M}{N} \tag{29}$$

Moreover, it is necessary for us to distinguish the compression rate denoted as u_f for the first pitch period and the compression rate denoted as u_s for the following periods. Thus, we have

$$u_f = \frac{M_f}{N} \tag{30}$$

and

$$u_s = \frac{M_s}{N} \tag{31}$$

where M_f and M_s represent the number of measurements for the first period and the following ones respectively. Moreover, it is required that

$$u_f \geq u_s \tag{32}$$

for mitigating error propagation.

The measure used to evaluate the reconstruction performance is signal to noise ratio (SNR) which is defined as

$$\mathrm{SNR} = 10log_{10}\frac{\|x\|_{l_2}^2}{\|x - x^*\|_{l_2}^2} \tag{33}$$

where x^* is the reconstructed signal vector.

As the adaptive sparsifying matrix is constructed according to the quasi-periodicity of voiced speech, it is necessary for us to analyze the reconstruction performance of the different types of voiced speech signals. We make an analysis of the testing speech signals and identify three types of voiced speech signals which are shown in Figure 1a, Figure 2a and Figure 3a. There are 41 frames, 21frames and 18 frames of voiced speech signals similar to the first type, the second type and the third type of voiced speech respectively in the testing speech signals. Figure 1b, Figure 2b and Figure 3b show the average compression rate for the above three types of voiced speech signals.

Moreover, it is illustrated in Figure 4, Figure 5 and Figure 6 the comparison of reconstruction qualities for the above three different types of voiced speech signals between the adaptive sparsifying matrix and the DCT basis. Figure 4a, Figure 5a and Figure 6a show average SNR of each pitch period with different compression rates with respect to the adaptive sparsifying matrix and the DCT basis. And Figure 4b, Figure 5b and Figure 6b show average SNR of each frame.

Regardless of the types of pitch periods, when $u_s \leq 0.5$, the reconstruction performance of the adaptive sparsifying matrix is far better than that of DCT. But when $u_s >$ 0.5, the adaptive sparsifying matrix and the DCT basis have similar performance for the first type and third type of voiced speech. However, for the second type of voiced speech, the reconstruction performance of the adaptive sparsifying matrix is slightly worse than that of the DCT basis. The reason is that with the great attenuation of

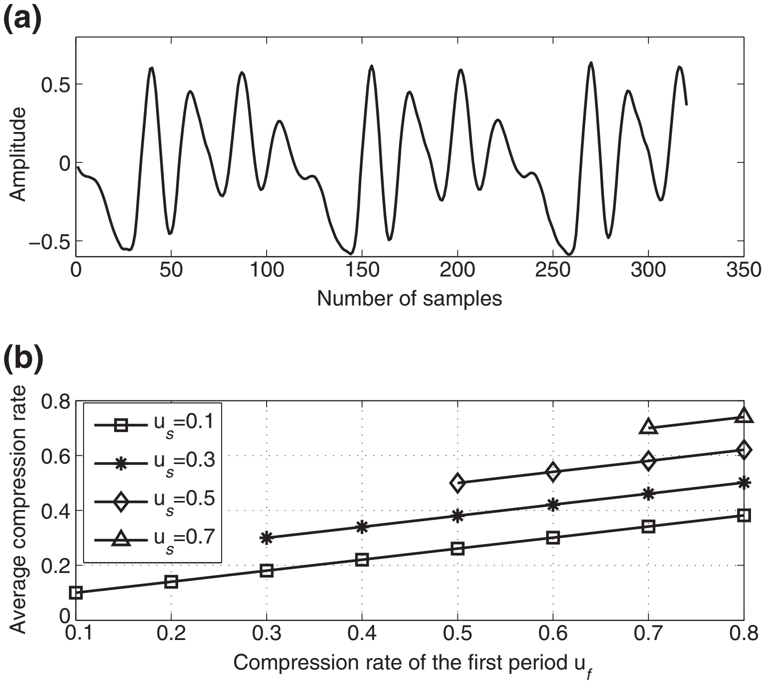

Figure 1 Waveform of the first type of voiced speech signals and average compression rate: (a) Waveform of the first type of voiced speech signals. (**b**) Average compression rate with different values of u_f and u_s.

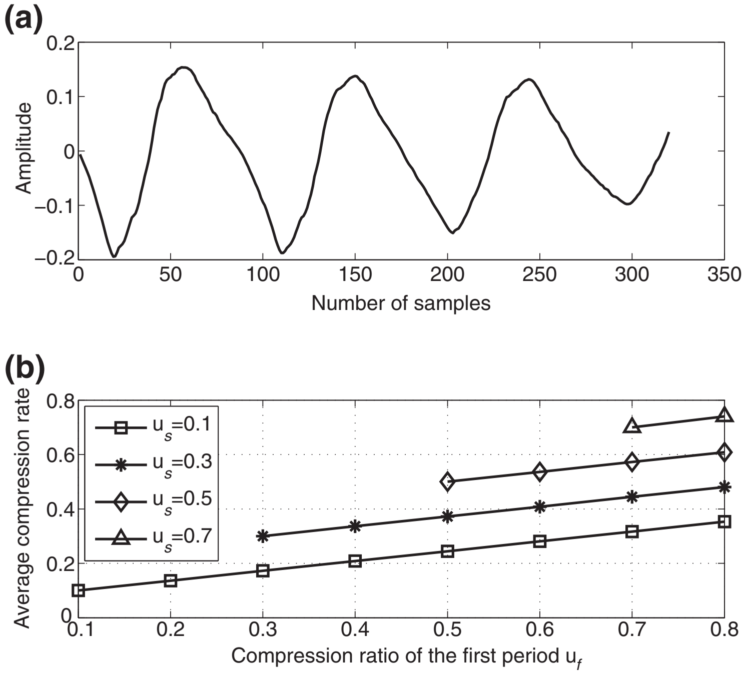

Figure 2 Waveform of the second type of voiced speech signals and average compression rate: (a) Waveform of the second type of voiced speech signals. (**b**) Average compression rate with different values of u_f and u_s.

the amplitude, the quasi-periodicity of the second type of voiced speech is undesirable.

Figure 7 illustrates the average reconstruction performance of all the voiced speech signals in the testing speech signals. It is obvious that the adaptive sparsifying matrix can achieve better reconstruction performance for voiced speech than the DCT basis with $u_s \le 0.5$. However, it is obvious in Figure 7 that the reconstruction performance of voiced speech signals with respect to the adaptive sparsifying matrix is slightly worse than that of the DCT basis with $u_s = 0.7$. The reason is that the approximate sparsity of the adaptive sparsifying matrix is far better than that of the DCT basis but the whole approximation accuracy of the adaptive sparsifying matrix is slightly worse than that of the DCT basis.

Finally, we apply the adaptive sparsifying matrix to the entire speech signals including voiced speech and unvoiced speech and illustrate the reconstruction performance in Figure 8. Compared with Figure 7, we found out the performance in this case just degrades slightly.

4 Sensing matrix for speech signals

4.1 Two-block diagonal matrix

A sufficient condition for successful reconstruction of a sparse vector from undersampled measurements is that the CS matrix satisfies RIP with a required constant. It has been shown in some literatures that a dense Gaussian random matrix whose entries are i.i.d. random variables drawn according to normal distribution with mean zero and variance $\frac{1}{M}$ [1, 2, 8] satisfies RIP with high probability.

In this section, a sensing matrix is constructed according to the characteristics of voiced speech signals. In [25–29], a kind of structured random matrix called block diagonal matrix is applied to achieve CS in wireless communication and image processing. In [25,26], a lot of identical blocks are used to construct a block diagonal matrix as a sensing matrix for image processing with no proof of its property to meet RIP. From a view of information theory, [27] proposes the block diagonal matrix for natural images also with no proof of its property to meet RIP. In addition, [28,29] present RIP for block diagonal matrices.

However, in this work, a specific block diagonal matrix with just two different blocks called two-block diagonal (TBD) matrix is constructed for voiced speech signals and a simple proof of its RIP is given although some proofs of RIP for block diagonal matrices have been given in [28,29].

As we know, the spectral energy of voiced speech signals is concentrated in low-frequency domain and decays rapidly. Thus, the high-frequency coefficients of a voiced speech signal in DCT domain are much sparser than the low-frequency coefficients. In the following, the definition of the TBD matrix is stated.

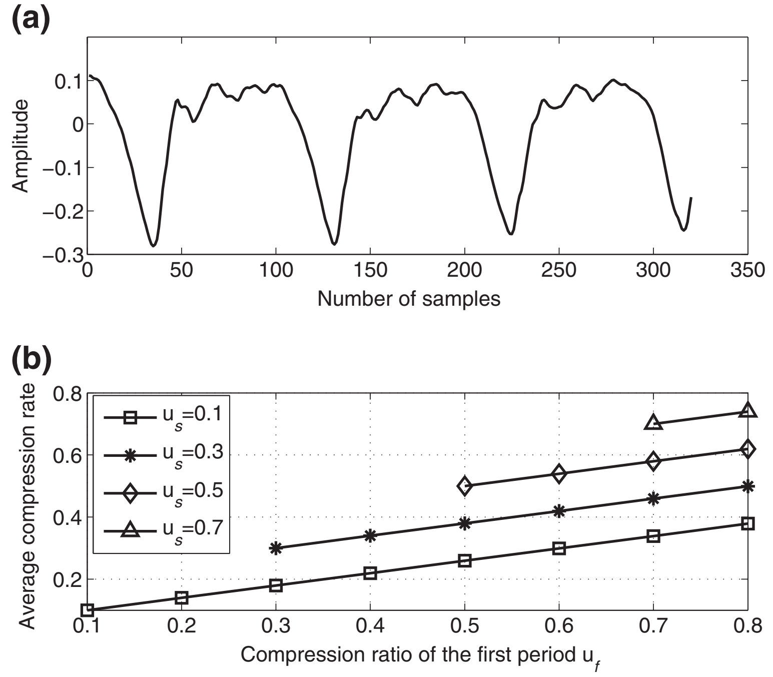

Figure 3 Waveform of the third type of voiced speech signals and average compression rate: (a) Waveform of the third type of voiced speech signals. (**b**) Average compression rate with different values of u_f and u_s.

Definition 2 (TBD matrix) A matrix $A \in R^{M \times N}$ is defined as the TBD matrix endowed with the following structure

$$A = \begin{bmatrix} \Phi_1 & 0 \\ 0 & \Phi_2 \end{bmatrix} \tag{34}$$

where $\Phi_1 \in R^{M_1 \times N_1}$ is a Gaussian random matrix whose entries are i.i.d. random variables drawn according to normal distribution with mean zero and variance $\frac{1}{M_1}$ and $\Phi_2 \in R^{M_2 \times N_2}$ is also a Gaussian random matrix whose entries are i.i.d. random variables drawn according to normal distribution with mean zero and variance $\frac{1}{M_2}$.

In line with this characteristic, a matrix $\Phi = \begin{bmatrix} \Phi_1 & 0 \\ 0 & \Phi_2 \end{bmatrix} \Psi^T$ is constructed as a sensing matrix for voiced speech signals, where Ψ^T is the transpose of an orthonormal basis. In addition, it is required that

$$M_1 \geq M_2 \tag{35}$$

$$M_1 + M_2 = M \tag{36}$$

and

$$N_1 + N_2 = N \tag{37}$$

And then we have

$$y = \Phi x = \begin{bmatrix} \Phi_1 & 0 \\ 0 & \Phi_2 \end{bmatrix} \Psi^T x = \begin{bmatrix} \Phi_1 & 0 \\ 0 & \Phi_2 \end{bmatrix} \theta = A\theta \tag{38}$$

where $\theta = \Psi^T x$. We just need to prove that A satisfies RIP.

Lemma 1 (Concentration inequality of TBD matrix) Suppose that the matrix A is a TBD matrix defined in definition 2. Then, the matrix obeys the concentration inequality with the prescribed constant δ

$$P\left(\left| \|A\theta\|_{l_2}^2 - \|\theta\|_{l_2}^2 \right| \geq \delta \|\theta\|_{l_2}^2 \right) \leq 2e^{-MC(\delta)} \tag{39}$$

where $C(\delta)$ is a constant depending on δ.

The proof of Lemma 1 can be found in Appendix. In order to prove that the TBD matrix satisfies RIP, a theorem in literature [8] is first recalled.

Theorem 1 ([8]) Suppose that a CS matrix A satisfies the concentration inequality. If

$$K \leq c_1 M / log(N/K) \tag{40}$$

the matrix A satisfies the K-order RIP with the prescribed constant δ with probability $\geq 1 - 2e^{-c_2 M}$, where c_1 and c_2 are constants depending on δ.

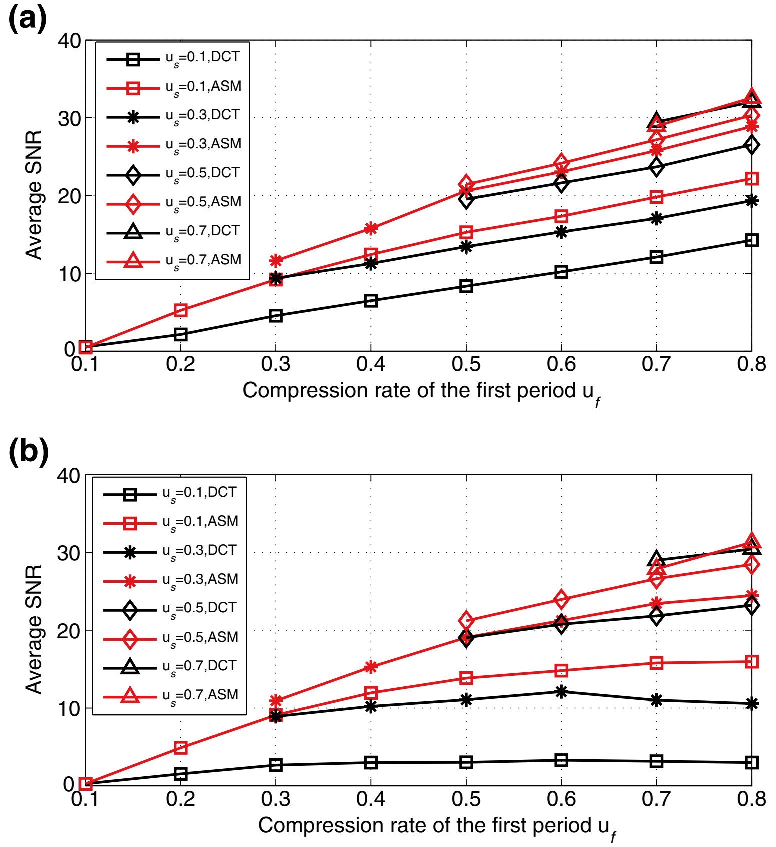

Figure 4 Average SNR of voiced speech signals whose waveforms are similar to that in Figure 1 (a): (a) Average reconstruction SNR of pitch periods with different values of u_f and u_s. (**b**) Average reconstruction SNR of frames with different values of u_f and u_s. ASM in the figure stands for adaptive sparsifying matrix.

Therefore, in light of Lemma 1 and theorem 1, it suffices to show that the TBD matrix A satisfies RIP. In fact, the TBD matrix can also be employed as the CS matrix when the sparsifying matrix is the adaptive sparsifying matrix in Section 3. The reason is that the coefficient vector β with respect to the adaptive sparsifying matrix in Eq. (12) also exhibits similar concentration characteristic to the DCT coefficients. However, it is inappropriate to employ the adaptive sparsifying matrix and the TBD matrix simultaneously in CS system. Firstly, the adaptive sparsifying matrix must be orthonormalized in this case, which undoubtedly increase the computational complexity of the CS system. Secondly, more parameters need to be adjusted. The last but not the least, the TBD matrix cannot considerably improve the reconstruction performance with respect to the adaptive sparsifying matrix for the extremely compressible coefficient vector β and limited approximation accuracy. Thus, we employ the DCT basis as the sparsifying matrix for speech signals in Section 4 and Section 5.

4.2 Simulation

The testing speech signals used in the experiments of this subsection are the same as in Section 3. The BP algorithm is also employed in this subsection to achieve reconstruction. At first, we define

$$u_l = \frac{M_1}{N} \, and \, u_h = \frac{M_2}{N} \tag{41}$$

and then we have

$$u = u_l + u_h \, and \, u_l \geq u_h \tag{42}$$

In this subsection, we firstly compare the reconstruction performance between the TBD matrix and the dense Gaussian random matrix with respect to the adaptive sparsifying matrix. Figure 9a and Figure 9b show the average SNR of pitch periods and frames respectively for the TBD matrix and the dense Gaussian random matrix in the case of the adaptive sparsifying matrix. It is obvious in Figure 9 that the TBD matrix cannot bring about desirable improvement on the reconstruction performance with respect to

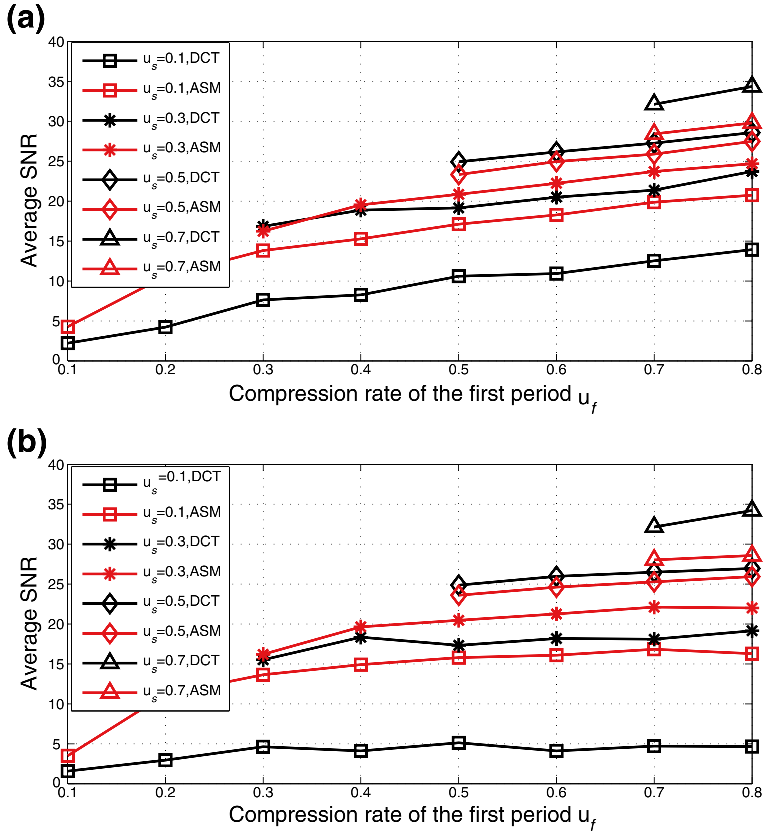

Figure 5 Average SNR of voiced speech signals whose waveforms are similar to that in Figure 2a. (**a**) Average reconstruction SNR of pitch periods with different values of *f u* and *s u*. (**b**) Average reconstruction SNR of frames with different values of *f u* and *s u*.

the adaptive sparsifying matrix. Therefore, we focus on the reconstruction performance when TBD matrix is used as the CS matrix with respect to the DCT basis.

Figure 10a shows the comparison of average SNR of 135 frames of voiced speech signals between the TBD matrix and dense Gaussian random matrix when the sparsifying matrix is the DCT basis. It is obvious that the performance of the TBD matrix with the right values of u_l and u_h is much better than that of the dense Gaussian random matrix especially when the value of overall compression rate u is relatively small.

Figure 10b demonstrates the comparison of average SNR of the entire testing speech signals between the TBD matrix and the dense Gaussian random matrix. Although the overall reconstruction performance degrades slightly, the TBD matrix with right values of u_l and u_h still performs much better than the dense Gaussian random matrix.

More importantly, as the TBD matrix can restrict the impact of quantization noise on reconstruction of speech signals, it can attain better reconstruction performance than the dense Gaussian matrix when the measurements are quantized, which is described in details in the next section.

5 Quantization effect on speech signals with compressed sensing

5.1 Quantization of speech signals in the framework of CS

In this paper, we apply CS to speech signals to achieve efficient compression. However, we still need to quantify the projections before transmission. At first, we should analyze the distribution of the projections. When the sensing matrix is the dense Gaussian random matrix, we have

$$y = \Phi x = \sum_{i=1}^{N} x(i)\phi_i \qquad (43)$$

where $= [\phi_1\ \phi_2 \cdots \phi_N] = \begin{bmatrix} \phi_{1,1} & \phi_{1,2} & \cdots & \phi_{1,N} \\ \phi_{2,1} & \phi_{2,2} & \cdots & \phi_{2,N} \\ \vdots & \vdots & \vdots & \vdots \\ \phi_{M,1} & \phi_{M,2} & \cdots & \phi_{M,N} \end{bmatrix}.$

And then, we can obtain

$$y(k) = \sum_{i=1}^{N} x(i)\phi_{k,i}\,(k = 1, 2\cdots M) \qquad (44)$$

where $\phi_{k,i}$ is the i.i.d. Gaussian random variable with mean zero and variance $\frac{1}{M}$. Thus, $y(k)$ is a random

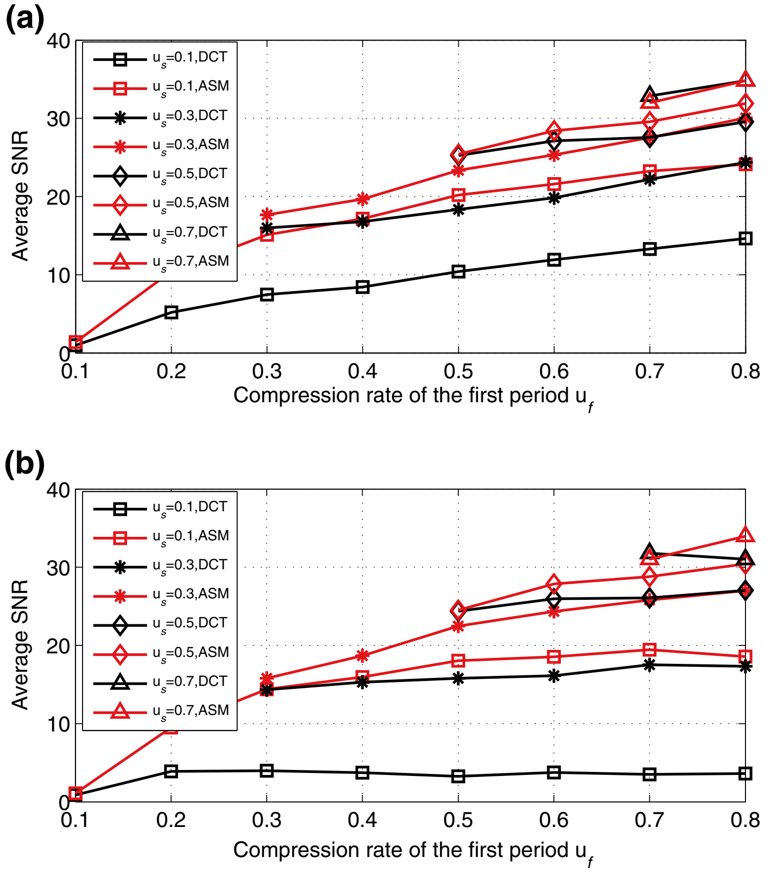

Figure 6 Average SNR of voiced speech signals whose waveforms are similar to that in Figure 3a. (**a**) Average reconstruction SNR of pitch periods with different values of *f u* and *s u*. (**b**) Average reconstruction SNR of frames with different values of *f u* and *s u*.

variable independently drawn by the normal distribution with

$$E(y(k)) = 0 \tag{45}$$

$$D(y(k)) = \frac{1}{M}\sum\nolimits_{i=1}^{N}(x(k))^2 = \frac{1}{M}\|x\|_{l_2}^2 \tag{46}$$

However, when the CS matrix is the TBD matrix, we have

$$y = A\theta = \begin{bmatrix} \Phi_1 & 0 \\ 0 & \Phi_2 \end{bmatrix}\begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} \tag{47}$$

where $\Phi_1 = \begin{bmatrix} \varphi_1^1 & \varphi_2^1 & \cdots & \varphi_{N_1}^1 \end{bmatrix}$ and $\Phi_2 = \begin{bmatrix} \varphi_1^2 & \varphi_2^2 & \cdots & \varphi_{N_2}^2 \end{bmatrix}$ are both dense Gaussian random matrices. Thus, we have

$$y(k) = \sum\nolimits_{i=1}^{N_1}\varphi_{k,i}^1\theta_1(i) \quad (k = 1, 2\cdots M_1) \tag{48}$$

and

$$y(k) = \sum\nolimits_{i=1}^{N_2}\varphi_{k,i}^2\theta_2(i) \; (k = M_1 + 1, M_1 + 2, \cdots M_1 + M_2) \tag{49}$$

Then $y(k)$ is also an independent Gaussian random variable with

$$E(y(k)) = 0 \;\; (k = 1, 2\cdots M_1 + M_2) \tag{50}$$

$$D(y(k)) = \frac{1}{M_1}\|\theta_1\|_{l_2}^2 \;\; (k = 1, 2\cdots M_1) \tag{51}$$

and

$$D(y(k)) = \frac{1}{M_2}\|\theta_2\|_{l_2}^2 \;\; (k = M_1{+}1, M_2{+}1, \cdots M_1{+}\, M_2) \tag{52}$$

We apply uniform scalar quantization to the projections. In [30], an analysis of the noise power generated by the uniform scalar quantization when the input signal meets the Gaussian distribution has been carried out and a table for the optimal values of finite quantization range for different quantization levels is provided, which contributes to our following analysis on adaptive quantization.

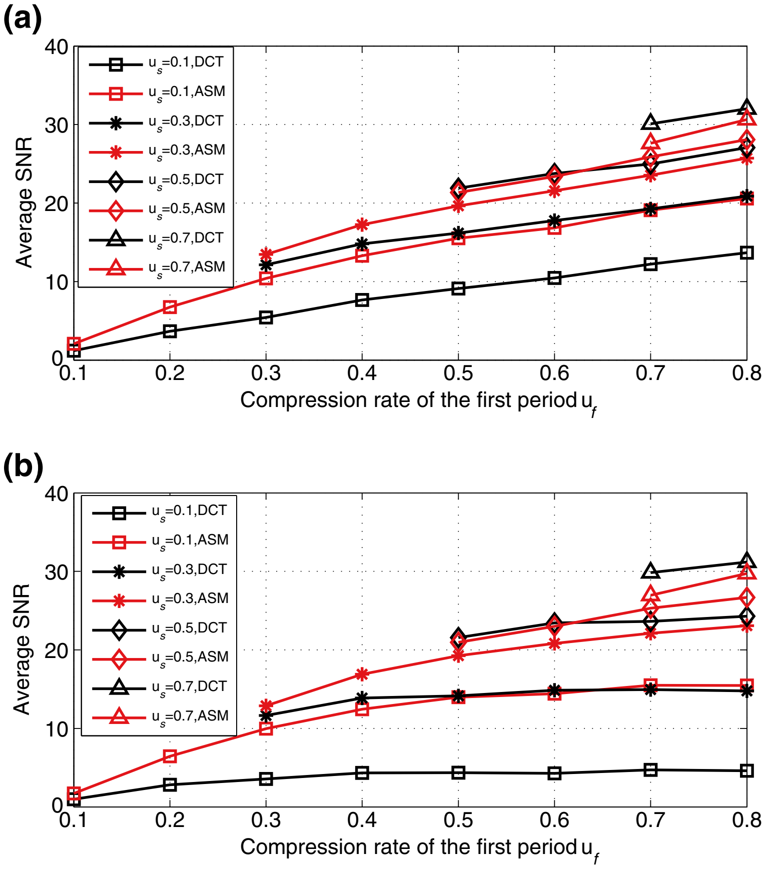

Figure 7 Average SNR of all the voiced speech signals in the testing speech signals: (a) Average reconstruction SNR of pitch periods with different values of u_f and u_s. (**b**) Average reconstruction SNR of frames with different values of u_f and u_s. ASM in the figure stands for adaptive sparsifying matrix.

Speech signals are a kind of time-variant signals and it is possible for the energy of different segments to show great changes. Furthermore, in an expectation sense, the energy of measurement vector is equal to that of the signal vector. Therefore, it is necessary to implement adaptive quantization to the projections. In the following, the effect of adaptive quantization on reconstruction performance is discussed in the framework of CS.

As we know, the noise of uniform scalar quantizer is induced by quantization and saturation. Let Δ denote the quantization interval, Q denote the number of quantization intervals and σ_i denote the standard deviation of the projection of the i^{th} frame of voiced speech. $[-m\sigma_i, m\sigma_i]$ is the quantization range for the i^{th} frame. And when the quantization is adaptive, $[-m\sigma_{i+1}, m\sigma_{i+1}]$ is the quantization range for the $(i+1)^{th}$ frame. Otherwise, when the quantizer is fixed, in the convenience of analysis, $[-m\sigma_i, m\sigma_i]$ is used as the quantization range for the $(i+1)^{th}$ frame. In other words, the nonadaptive quantization is used in the fixed quantizer. And EN_a denotes the noise power for the adaptive quantizer and EN_f denotes the noise power for the fixed quantizer. Hence, for adaptive quantizer, we have

$$\Delta = \frac{2m\sigma_{i+1}}{Q} \tag{53}$$

and

$$EN_a = \frac{\Delta^2}{12}\left(1-2\int_m^{+\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt\right) +2\left(\left(m^2+1\right)\sigma_{i+1}^2\int_m^{+\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt-\frac{m\sigma_{i+1}^2}{\sqrt{2\pi}}e^{-\frac{m^2}{2}}\right) \tag{54}$$

For a fixed quantizer, we have

$$\Delta = \frac{2m\sigma_i}{Q} \tag{55}$$

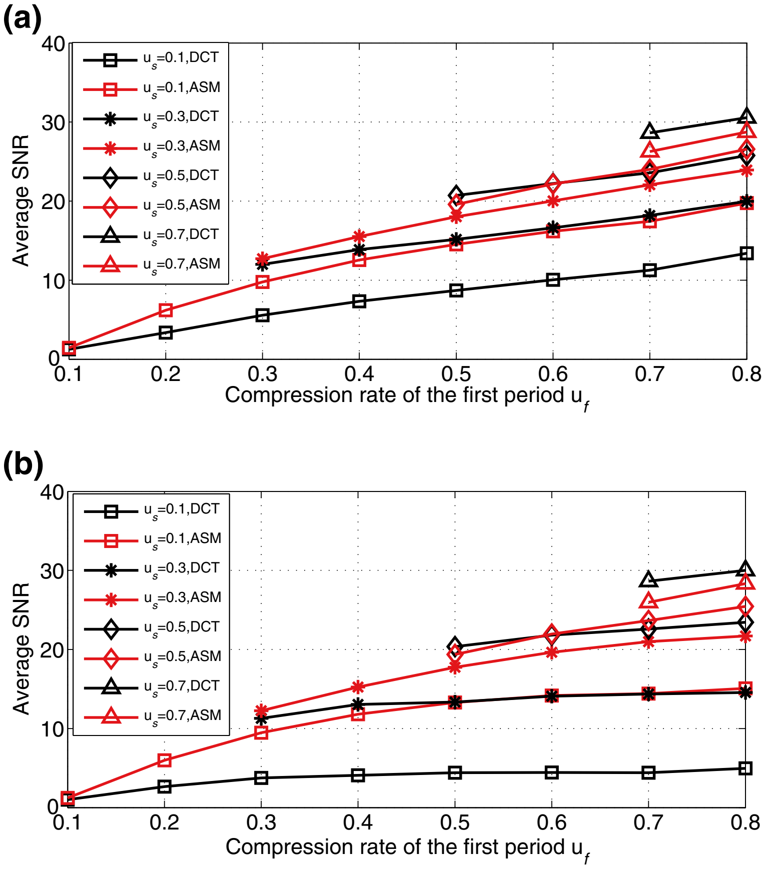

Figure 8 Average SNR of the entire speech signals: (a) Average reconstruction SNR of pitch periods with different values of u_f and u_s. (**b**) Average reconstruction SNR of frames with different values of u_f and u_s. ASM in the figure stands for adaptive sparsifying matrix.

and

$$EN_f = \frac{\Delta^2}{12}\left(1 - 2\int_{m\frac{\sigma_i}{\sigma_{i+1}}}^{+\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt\right) + 2\left(\sigma_{i+1}^2\left(1+\frac{m^2\sigma_i^2}{\sigma_{i+1}^2}\right)\int_{m\frac{\sigma_i}{\sigma_{i+1}}}^{+\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt - \frac{m\sigma_i\sigma_{i+1}}{\sqrt{2\pi}}e^{-\frac{m^2\sigma_i^2}{2\sigma_{i+1}^2}}\right) \tag{56}$$

From Eq. (56), it is clear that the noise power of a fixed quantizer depends not only on the variance of the current frame but also depends on the ratio of the variances of the successive two frames.

Theorem 2 ([16]): Suppose that θ is an approximately sparse vector in R^N. Assuming that the $2K$-order restricted isometry constant of the CS matrix satisfies

$$\delta_{2K} < \sqrt{2} - 1 \tag{57}$$

the solution θ^* to Eq. (8) obeys

$$\|\theta^* - \theta\|_{l_2} \le C_1\varepsilon + C_2\frac{\|\theta - \theta_K\|_{l_1}}{\sqrt{K}} \tag{58}$$

where C_1 and C_2 are constants depending on δ_{2K}.

For an adaptive quantizer, the reconstruction SNR is written as SNR_a, and for a fixed quantizer, the reconstruction SNR is written as SNR_f. In the following, two corollaries about the impact of the adaptive quantization on reconstruction performance are derived in this paper. In this paper, we focus on the effect of quantization noise. Therefore, in the two corollaries below, we assume that $\|\theta - \theta_K\|_{l_1}$ extends to zero.

Corollary 1: Suppose that x is a voiced speech signal vector and the sparsifying matrix is an orthonormal basis Ψ. Provided that the sensing matrix is the dense Gaussian random matrix whose entries are i.i.d. Gaussian variables with mean 0 and variance $\frac{1}{M}$, there exist a constant C_q so that the reconstruction SNR for an adaptive quantizer with the value of quantization level Q to be 32 obeys

$$\mathrm{SNR_a} \ge 24.792 - 10log_{10}C_1^2C_q \tag{59}$$

Assuming that $\frac{\sigma_i}{\sigma_{i+1}} = 1.25$, then the reconstruction SNR for a fixed quantizer with the value of Q to be 32 obeys

$$\mathrm{SNR}_f \ge 23.656 - 10log_{10}C_1^2C_q \tag{60}$$

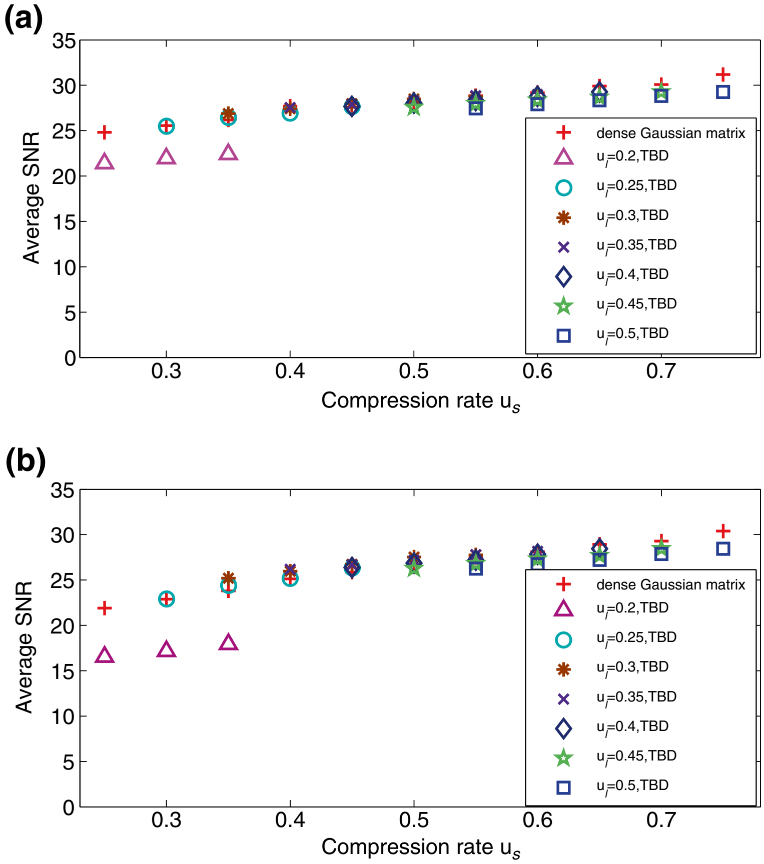

Figure 9 Comparsion of average SNR between the TBD matrix and the dense Gaussian matrix: (a) Average reconstruction SNR of pitch periods of voiced speech signals with respect to the adaptive sparsifying matrix. (**b**) Average reconstruction SNR of frames of voiced speech signals with respect to the adaptive sparsifying matrix.

Assuming that $\frac{\sigma_i}{\sigma_{i+1}} = 0.75$, then the reconstruction SNR for a fixed quantizer with the value of Q to be 32 obeys

$$\mathrm{SNR}_f \geq 20.8067 - 10log_{10}C_1^2C_q \tag{61}$$

Corollary 2*:* Suppose that x is a voiced speech signal vector and the sparsifying matrix is the DCT basis. Provided that the CS matrix is the TBD matrix and u_l and u_h are defined as in Eq. (41), there exist a constant C_p so that the reconstruction SNR for an adaptive quantizer with the value of Q to be 32 obeys

$$\mathrm{SNR}_a \geq 10log_{10}\frac{u_l}{C_1^2C_p(3.317\times10^{-3}u_l + 2.738\times10^{-3}u_h)} \tag{62}$$

Assuming that $\frac{\sigma_i}{\sigma_{i+1}} = 1.25$, then the reconstruction SNR for a fixed quantizer with the value of Q to be 32 obeys

$$\mathrm{SNR}_f \geq 10log_{10}\frac{u_l}{C_1^2C_p(4.39\times10^{-3}u_l + 4.2775\times10^{-3}u_h)} \tag{63}$$

Assuming that $\frac{\sigma_i}{\sigma_{i+1}} = 0.75$, then the reconstruction SNR for a fixed quantizer with value of Q to be 32 obeys

$$\mathrm{SNR}_f \geq 10log_{10}\frac{u_l}{C_1^2C_p(8.305\times10^{-3}u_l + 1.534\times10^{-3}u_h)} \tag{64}$$

5.2 Simulation

The testing speech signals used in experiments of this subsection are also the same as that in Section 3. The sparsifying matrix used in this section is the DCT basis. And we employ the BPDN algorithm to achieve reconstruction in this subsection. The measure of performance evaluation is also the average SNR.

At first, we analyze the performance of adaptive quantization compared with the nonadaptive quantization for both the TBD matrix and the dense Gaussian random matrix in the framework of CS. We fixed the value of Q to be 32. Figure 11a illustrates the quantization effect on voiced speech signals of the testing speech signals. And Figure 11b illustrates the quantization effect on the entire

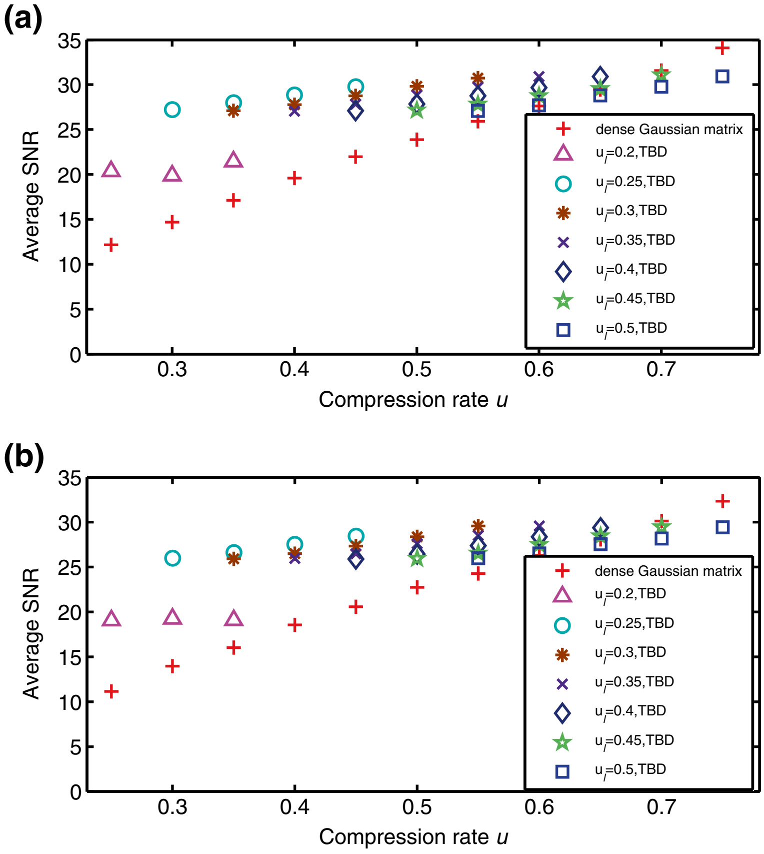

Figure 10 Comparison of average SNR between the TBD matrix and the dense Gaussian matrix: (a) Average reconstruction SNR of voiced speech signals when the sparsifying matrix is the DCT basis. (**b**) Average reconstruction SNR of the entire speech signals when the sparsifying matrix is the DCT basis.

testing speech signals. It is obvious that the adaptive quantization can greatly improve the reconstruction performance compared with the nonadaptive quantization. Moreover, we can find out from Figure 11 that the performance of TBD matrix with $u_h = 0.05$ is superior to the dense Gaussian random matrix for both the adaptive quantization and nonadaptive quantization. The reason is that the TBD matrix is more robust to quantization noise based on the fact the TBD matrix can effectively restrict the impact of quantization on speech signals.

In the following, we focus on the adaptive quantization effect on reconstruction of speech signals with different quantization levels. Figure 12a, Figure 12b, Figure 13a and Figure 13b show the average reconstruction SNR of voiced speech signals with the quantization level Q to be 8, 16, 32 and 64 respectively when the adaptive quantization is applied to the projections in the case of TBD matrices and the dense Gaussian matrix. On the one hand, the reconstruction performance in the case of adaptive quantization improves with the increase of the quantization level. On the other hand, with right values of u_l and u_h, TBD matrix performs much better than the dense Gaussian random matrix confronted with the quantization noise regardless of the quantization level. In addition, Figure 14a, Figure 14b, Figure 15a and Figure 15b show the average reconstruction SNR of entire speech signals with the quantization level Q to be 8, 16, 32 and 64 respectively. And the above findings also hold for the entire speech signals including voiced and unvoiced speech signals. Thus, we can conclude that the adaptive quantization and the TBD matrix can effectively mitigate the impact of quantization noise on reconstruction in the framework of CS.

6 Conclusions

This paper demonstrates the potential of applying CS to speech signals especially voiced speech signals. From the viewpoint of long-term prediction, we analyze the sparsity of voiced speech signals and construct an adaptive sparsifying matrix. Moreover, a CS matrix called TBD matrix is constructed in terms of the spectral characteristics of voiced speech signals. Finally, the distribution of the projections is analyzed to carry out quantization. And the reconstruction performance of the adaptive

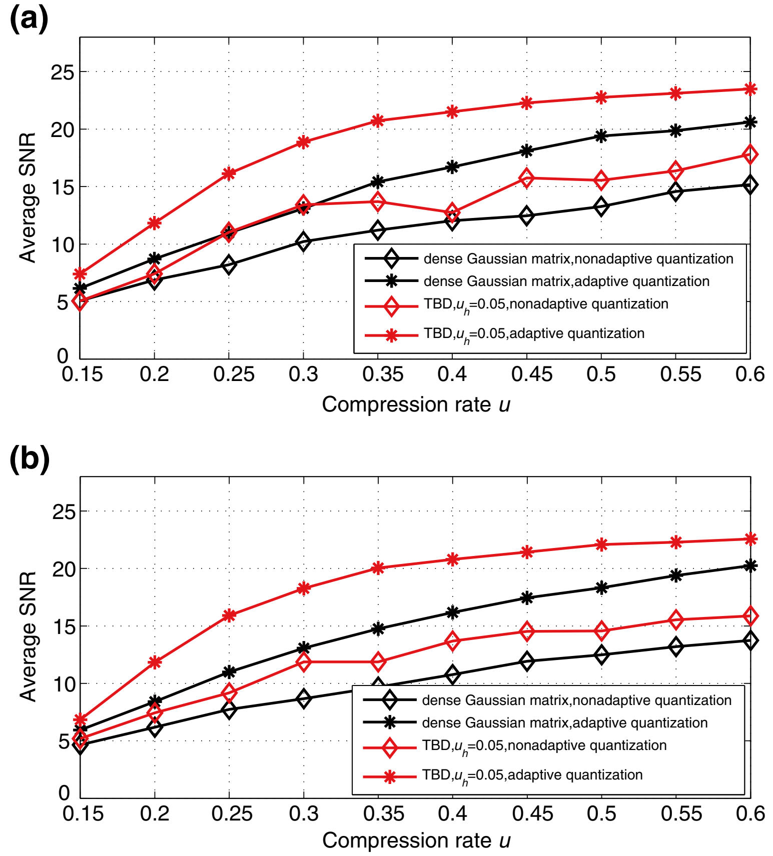

Figure 11 Average reconstruction SNR of adaptive quantization and nonadaptive quantization: (a) Average reconstruction SNR of voiced speech signals with respect to the TBD matrix and dense Gaussian matrix in the case of adaptive quantization and nonadaptive quantization. (**b**) Average reconstruction SNR of the entire testing speech signals with respect to the TBD matrix and dense Gaussian matrix in the case of adaptive quantization and nonadaptive quantization.

quantization and nonadaptive quantization is studied. In addition, under the adaptive quantization, the reconstruction qualities of TBD matrix and the dense Gaussian matrix are empirically compared with different quantization bits. Therefore, we find that the TBD matrix and the adaptive quantization can effectively mitigate the quantization effect on reconstruction of speech signals in the framework of CS.

Appendix

Proof of Lemma 1 Let $\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}$ where θ_1 and θ_2 are also column vectors. Then, we have

$$A\theta = \begin{bmatrix} \Phi_1 & 0 \\ 0 & \Phi_2 \end{bmatrix} \theta = \begin{bmatrix} \Phi_1 & 0 \\ 0 & \Phi_2 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} \Phi_1\theta_1 \\ \Phi_2\theta_2 \end{bmatrix} \tag{65}$$

and

$$\|A\theta\|_{l_2}^2 = \|\Phi_1\theta_1\|_{l_2}^2 + \|\Phi_2\theta_2\|_{l_2}^2 \tag{66}$$

As Φ_1 is an $M_1 \times N_1$ Gaussian matrix whose entries are i.i.d. random variables drawn according to normal distribution with mean zero and variance $\frac{1}{M_1}$ and Φ_2 is an $M_2 \times N_2$ Gaussian matrix whose entries are i.i.d. random variables drawn according to normal distribution with mean zero and variance $\frac{1}{M_2}$, we establish

$$E\left(\|\Phi_1\theta_1\|_{l_2}^2\right) = \|\theta_1\|_{l_2}^2 \tag{67}$$

and

$$E\left(\|\Phi_2\theta_2\|_{l_2}^2\right) = \|\theta_2\|_{l_2}^2 \tag{68}$$

Hence, we have

$$E\left(\|A\theta\|_{l_2}^2\right) = \|\theta_1\|_{l_2}^2 + \|\theta_2\|_{l_2}^2 = \|\theta\|_{l_2}^2 \tag{69}$$

Moreover, it is proved in [31] and [32] that

$$P\left(\left|\|\Phi_1\theta_1\|_{l_2}^2 - \|\theta_1\|_{l_2}^2\right| \geq \delta\|\theta_1\|_{l_2}^2\right) \leq 2e^{-\frac{M_1\delta^2}{8}} \tag{70}$$

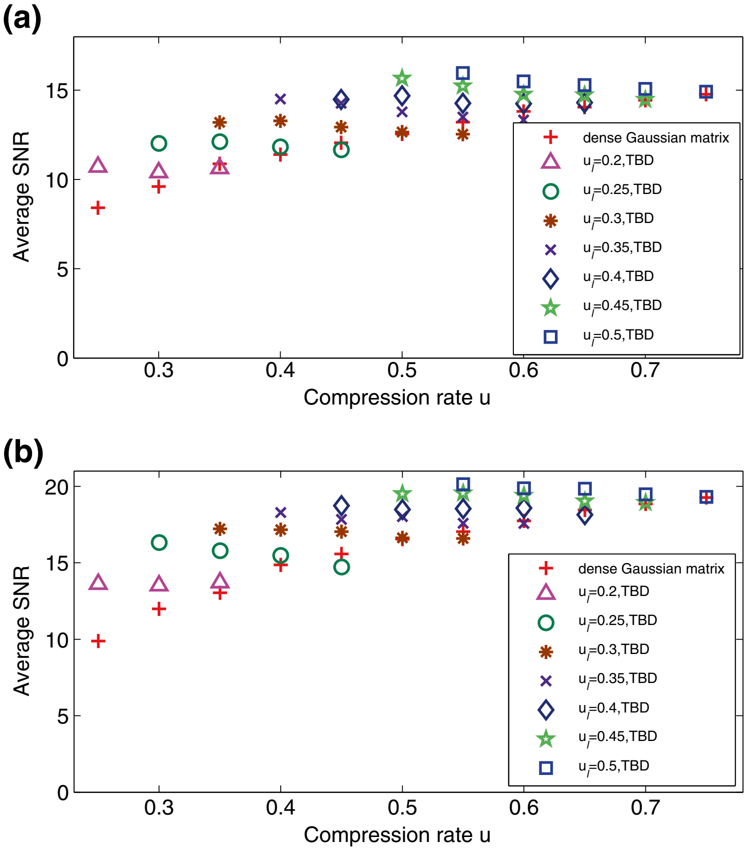

Figure 12 Average SNR of adaptive quantization of voiced speech signals with different quantization levels: (a) *Q*=8. (b) *Q*=16.

and

$$P\left(\left|\|\Phi_2\theta_2\|_{l_2}^2 - \|\theta_2\|_{l_2}^2\right| \geq \delta\|\theta_2\|_{l_2}^2\right) \leq 2e^{-\frac{M_2\delta^2}{8}} \tag{71}$$

Therefore, we have

$$P\left(-\delta\|\theta_1\|_{l_2}^2 \leq \|\Phi_1\theta_1\|_{l_2}^2 - \|\theta_1\|_{l_2}^2 \leq \delta\|\theta_1\|_{l_2}^2\right) \geq 1 - 2e^{-\frac{M_1\delta^2}{8}} \tag{72}$$

and

$$P\left(-\delta\|\theta_2\|_{l_2}^2 \leq \|\Phi_2\theta_2\|_{l_2}^2 - \|\theta_2\|_{l_2}^2 \leq \delta\|\theta_2\|_{l_2}^2\right) \geq 1 - 2e^{-\frac{M_2\delta^2}{8}} \tag{73}$$

Then, it suffice to show that

$$P\left(\left\{\left|\|\Phi_1\theta_1\|_{l_2}^2 - \|\theta_1\|_{l_2}^2\right| \leq \delta\|\theta_1\|_{l_2}^2\right\}\right.$$
$$\left.\cap\left\{\left|\|\Phi_2\theta_2\|_{l_2}^2 - \|\theta_2\|_{l_2}^2\right| \leq \delta\|\theta_2\|_{l_2}^2\right\}\right) \geq 1 - 2e^{-\frac{M_1\delta^2}{8}} - 2e^{-\frac{M_2\delta^2}{8}} \tag{74}$$

We can use the union bound to show that

$$\begin{aligned} P\left(\left|\|A\theta\|_{l_2}^2 - \|\theta\|_{l_2}^2\right| \geq \delta\|\theta\|_{l_2}^2\right) &\leq P\left(\left\{\left|\|\Phi_1\theta_1\|_{l_2}^2 - \|\theta_1\|_{l_2}^2\right|\right.\right. \\ &\geq \delta\|\theta_1\|_{l_2}^2\Big\} \cup \left.\left\{\left|\|\Phi_2\theta_2\|_{l_2}^2 - \|\theta_2\|_{l_2}^2\right| \geq \delta\|\theta_2\|_{l_2}^2\right\}\right) \\ &\leq P\left(\left|\|\Phi_1\theta_1\|_{l_2}^2 - \|\theta_1\|_{l_2}^2\right| \geq \delta\|\theta_1\|_{l_2}^2\right) \\ &+ P\left(\left|\|\Phi_2\theta_2\|_{l_2}^2 - \|\theta_2\|_{l_2}^2\right| \geq \delta\|\theta_2\|_{l_2}^2\right) \leq 2e^{-\frac{M_1\delta^2}{8}} + 2e^{-\frac{M_2\delta^2}{8}} \end{aligned} \tag{75}$$

There is certainly a constant $C(\delta) > 0$ for $\delta \in (0, 1)$ so that

$$e^{-\frac{M_1\delta^2}{8}} + e^{-\frac{M_2\delta^2}{8}} = e^{-MC(\delta)} \tag{76}$$

which yields that

$$C(\delta) = -\frac{log\left(e^{-\frac{M_1\delta^2}{8}} + e^{-\frac{M_2\delta^2}{8}}\right)}{M} \tag{77}$$

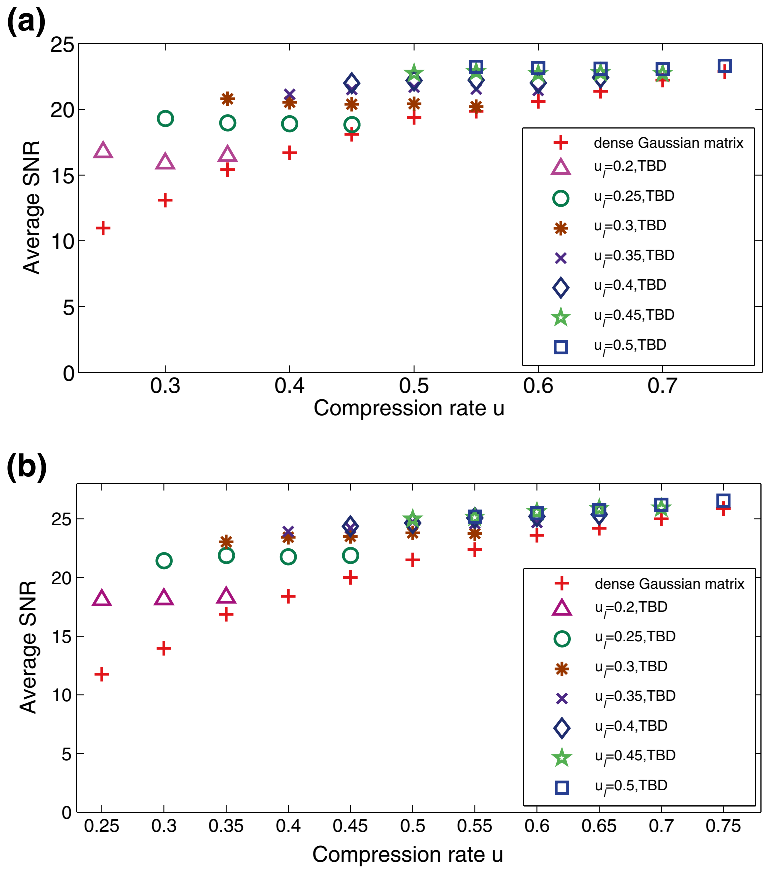

Figure 13 Average SNR of adaptive quantization of voiced speech signals with different quantization levels: (a) *Q*=32. (b) *Q*=64.

Thus, we can conclude that

$$P\left(\left|\|A\theta\|_{l_2}^2 - \|\theta\|_{l_2}^2\right| \geq \delta\|\theta\|_{l_2}^2\right) \leq 2e^{-MC(\delta)} \tag{78}$$

Proof of Corollary 1 The class X of interest is a finite set of objects x which are voiced segments. Denote then

$$X=\left\{x_k : x_k \textit{ is the } k^{th} \textit{ frame of voiced speech signals}\right\}. \tag{79}$$

When the sensing matrix is the dense Gaussian random matrix, the projection vector of the $(i+1)^{th}$ frame of voiced speech signal x_{i+1} is denoted by y_{i+1} and then

$$y_{i+1} = \Phi x_{i+1}. \tag{80}$$

In terms of Eq. (45) and Eq. (46), the entries of y_{i+1} are i.i.d. Gaussian random variables with mean 0 and variance $\frac{1}{M}\|x_{i+1}\|_{l_2}^2$. And the quantization vector of y_{i+1} is denoted by

$$\hat{y}_{i+1} = y_{i+1} + e_{i+1} = \Phi x_{i+1} + e_{i+1} \tag{81}$$

where $e_{i+1} = \left[e_{i+1}(1) \quad e_{i+1}(2) \quad \cdots \quad e_{i+1}(M)\right]^T$ is the quantization error vector of the $(i+1)^{th}$ frame. The quantization error vectors for all the voiced segments in X can be represented by a matrix —$e = \left[e_1 \quad e_2 \quad \cdots \quad e_{|X|}\right]$ where $|X|$ denotes the cardinality of the set X. When $Q = 32$, according to the results in [30], $m = 2.9$. Then for an adaptive quantizaer, in light with Eq. (54), we have,

$$E\left((e_{i+1}(k))^2\right) = 3.317 \times 10^{-3}\sigma_{i+1}^2 \quad (k = 1, 2\cdots M) \tag{82}$$

and

$$E\left(\|e_{i+1}\|_{l_2}^2\right) = ME\left((e_{i+1}(k))^2\right) = 3.317 \times 10^{-3}M\sigma_{i+1}^2. \tag{83}$$

We can find a subset in X denoted by V that can be represented as

$$V = \left\{k : \|x_k\|_{l_2}^2 = \|x_{i+1}\|_{l_2}^2, x_k \in X\right\}. \tag{84}$$

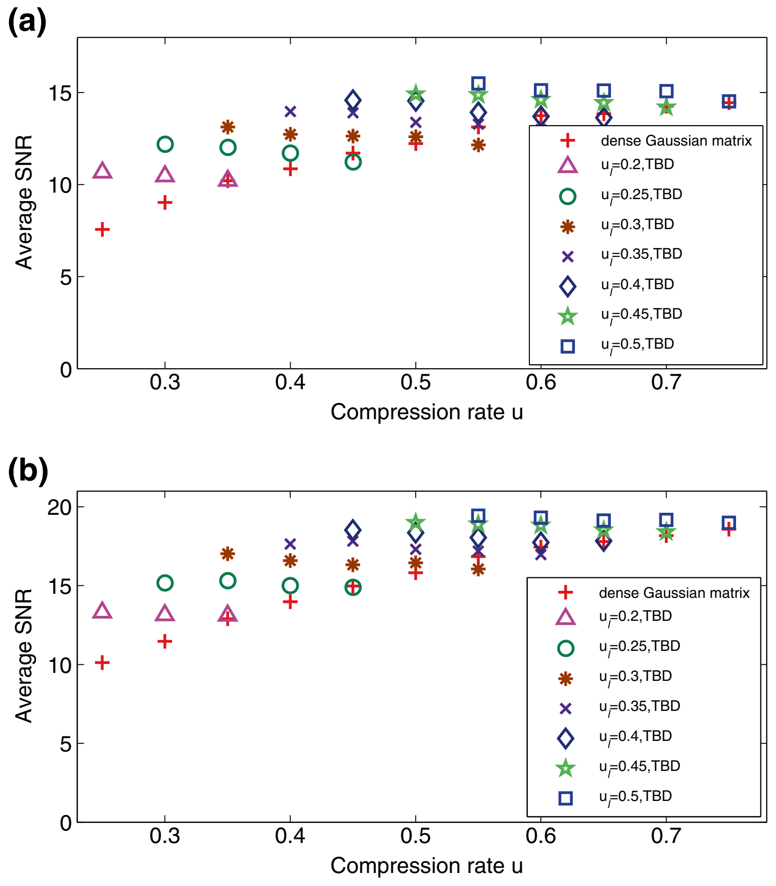

Figure 14 Average SNR of adaptive quantization of entire speech signals with different quantization levels: (a) *Q*=8. (b) *Q*=16.

Let $\varepsilon > 0$ and we have

$$\varepsilon^2 = \sup_{j \in V} \|e_j\|_{l_2}^2. \qquad (85)$$

There exist a constant C_a such that

$$\varepsilon^2 = C_a E\left(\|e_{i+1}\|_{l_2}^2\right) \qquad (86)$$

As $\|e_{i+1}\|_{l_2}^2 \le \varepsilon^2$, we have

$$\|e_{i+1}\|_{l_2}^2 \le C_a E\left(\|e_{i+1}\|_{l_2}^2\right) \qquad (87)$$

In this paper, we are just concerned with the impact of quantization on reconstruction. Therefore, we assume that $\|\theta - \theta_K\|_{l_1}$ extends to zero. While the voiced speech signal is compressible with respect to an orthonormal basis, we have

$$\begin{aligned}\left\|x_{i+1} - x_{i+1}^*\right\|_{l_2}^2 &= \left\|\Psi\left(\theta_{i+1} - \theta_{i+1}^*\right)\right\|_{l_2}^2 \\ &= \left\|\theta_{i+1} - \theta_{i+1}^*\right\|_{l_2}^2 \le 3.317 \times 10^{-3} C_1^2 C_a M \sigma_{i+1}^2 \end{aligned} \qquad (88)$$

where $x_{i+1}{}^* = \Psi\theta_{i+1}{}^*$ and $\theta_{i+1}{}^*$ is the solution to

$$min\|\theta_{i+1}\|_{l_1} \text{s.t} \left\|\hat{\mathrm{y}}_{i+1} - \Phi\Psi\theta_{i+1}\right\|_{l_2} \le \varepsilon. \qquad (89)$$

Therefore,

$$\begin{aligned}\mathrm{SNR}_a &\ge 10log_{10}\left(\frac{M\sigma_{i+1}^2}{3.317 \times 10^{-3} C_1^2 C_a M \sigma_{i+1}^2}\right) \\ &= 24.792 - 10log_{10}\left(C_1^2 C_a\right). \end{aligned} \qquad (90)$$

However, for a fixed quantizer, when $\frac{\sigma_i}{\sigma_{i+1}} = 1.25$, according to Eq. (56), we establish

$$E\left((e_{i+1}(k))^2\right) = 4.309 \times 10^{-3} \sigma_{i+1}^2. \qquad (91)$$

Then, we have

$$E\left(\|e_{i+1}\|_{l_2}^2\right) = ME\left((e_{i+1}(k))^2\right) = 4.309 \times 10^{-3} M \sigma_{i+1}^2. \qquad (92)$$

Let $\varepsilon_1 > 0$ and we have

$$\varepsilon_1^2 = \sup_{j \in V} \|e_j\|_{l_2}^2. \qquad (93)$$

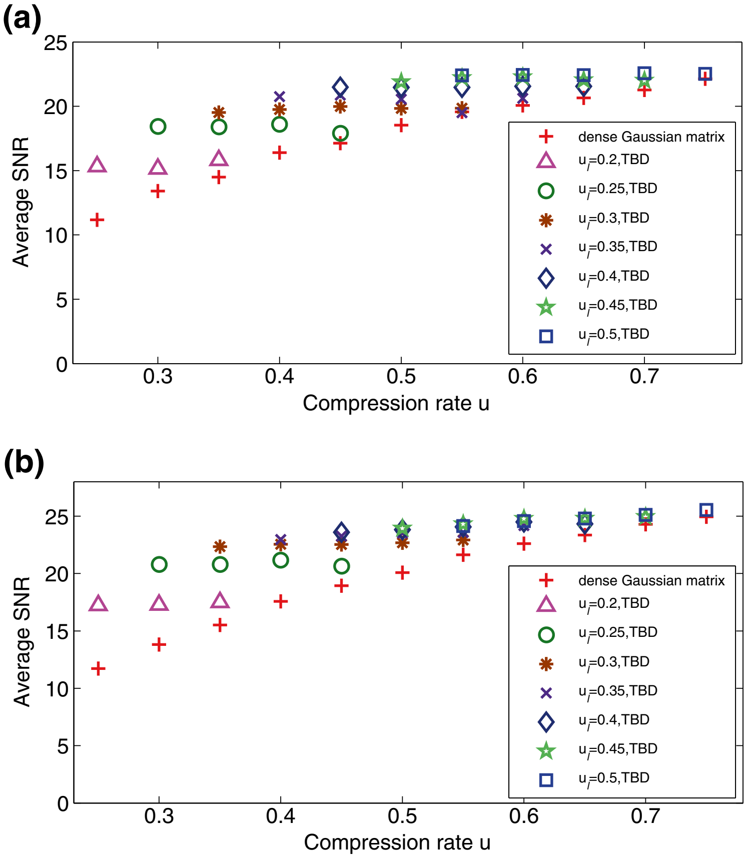

Figure 15 Average SNR of adaptive quantization of entire speech signals with different quantization levels: (a) *Q*=32. (b) *Q*=64.

There exist a constant C_{f_1} so that

$$\varepsilon_1^2 = C_{f_1} E\left(\|e_{i+1}\|_{l_2}^2\right) = 4.309 \times 10^{-3} C_{f_1} M \sigma_{i+1}^2 \tag{94}$$

Then we have

$$\begin{aligned} \mathrm{SNR}_f &\geq 10log_{10}\left(\frac{M\sigma_{i+1}^2}{4.309 \times 10^{-3} C_1^2 C_{f_1} M \sigma_{i+1}^2}\right) \\ &= 23.656 - 10log_{10}\left(C_1^2 C_{f_1}\right) \end{aligned} \tag{95}$$

Similarly, when $\frac{\sigma_i}{\sigma_{i+1}} = 0.75$, we have

$$E\left((e_{i+1}(k))^2\right) = 8.305 \times 10^{-3} \sigma_{i+1}^2 \tag{96}$$

Thus, we can establish that

$$\begin{aligned} \mathrm{SNR}_f &\geq 10log_{10}\left(\frac{M\sigma_{i+1}^2}{8.305 \times 10^{-3} C_1^2 C_{f_2} M \sigma_{i+1}^2}\right) \\ &= 20.8067 - 10log_{10}\left(C_1^2 C_{f_2}\right) \end{aligned} \tag{97}$$

Let $C_q = max\left(C_a, C_{f_1}, C_{f_2}\right)$ and then we obtain

$$\mathrm{SNR_a} \geq 24.792 - 10log_{10} C_1^2 C_q.$$

When $\frac{\sigma_i}{\sigma_{i+1}} = 1.25$, we have

$$\mathrm{SNR}_f \geq 23.656 - 10log_{10} C_1^2 C_q.$$

When $\frac{\sigma_i}{\sigma_{i+1}} = 0.75$, we have

$$\mathrm{SNR}_f \geq 20.8067 - 10log_{10} C_1^2 C_q.$$

Proof of Corollary 2 When the CS matrix is the TBD matrix, then we have

$$y_{i+1} = A\theta_{i+1} = \begin{bmatrix} \Phi_1 & 0 \\ 0 & \Phi_2 \end{bmatrix} \theta_{i+1} = \begin{bmatrix} \Phi_1 & 0 \\ 0 & \Phi_2 \end{bmatrix} \begin{bmatrix} \theta_{i+1,1} \\ \theta_{i+1,2} \end{bmatrix}$$

where θ_{i+1} is the coefficient vector of x_{i+1} with respect to DCT. In terms of Eqs. (50), (51), (52), denote then

$$\sigma_{i+1,1}^2 = \frac{1}{M_1} \left\|\theta_{i+1,1}\right\|_{l_2}^2 \tag{98}$$

and

$$\sigma^2_{i+1,2} = \frac{1}{M_2}\left\|\theta_{i+1,2}\right\|^2_{l_2}. \tag{99}$$

Moreover, according to the characteristic of the voiced segments, $\sigma_{i+1,1} \gg \sigma_{i+1,2}$. As an adaptive quantizer, $[-m\sigma_{i+1,1}, m\sigma_{i+1,1}]$ is used as the quantization range of the $(i+1)^{th}$ projection vector y_{i+1} and $\Delta = \frac{2m\sigma_{i+1,1}}{Q}$. In light with Eq. (54), for an adaptive quantizer, we have

$$E\left((e_{i+1}(k))^2\right) = 3.317 \times 10^{-3}\sigma^2_{i+1,1}\,(k = 1, 2\cdots M_1) \tag{100}$$

And in terms of Eq. (56), we have

$$E\left((e_{i+1}(k))^2\right) \approx \frac{\Delta^2}{12} = 2.738 \times 10^{-3}\sigma^2_{i+1,1}\,(k = M_1 + 1, M_1 + 2, \cdots M_1 + M_2) \tag{101}$$

and

$$E\left(\|e_{i+1}\|^2_{l_2}\right) = M_1 E\left((e_{i+1}(M_1))^2\right) + M_2 E\left((e_{i+1}(M_1 + M_2))^2\right) = 3.317 \times 10^{-3} M_1\sigma^2_{i+1,1} + 2.738 \times 10^{-3} M_2\sigma^2_{i+1,1} \tag{102}$$

We can find a subset in X denoted by V that can be represented as

$$V = \left\{k : \left\|\theta_{k,1}\right\|^2_{l_2} = \left\|\theta_{i+1,1}\right\|^2_{l_2}, \left\|\theta_{k,2}\right\|^2_{l_2} = \left\|\theta_{i+1,2}\right\|^2_{l_2}, \theta_k \text{ is the DCT coefficients vector of } x_k, x_k \in X\right\} \tag{103}$$

We define that $\varepsilon^2 = \sup_{j \in V}\left\|e_j\right\|^2_{l_2}$. There exist a constant C_b such that

$$\varepsilon^2 = C_b E\left(\|e_{i+1}\|^2_{l_2}\right) \tag{104}$$

Therefore, we establish

$$\|e_{i+1}\|^2_{l_2} \le C_b E\left(\|e_{i+1}\|^2_{l_2}\right) \tag{105}$$

As stated in corollary 1, we extend $\|\theta - \theta_K\|_{l_1}$ to zero. Thus, we establish $\left\|x_{i+1} - x^*_{i+1}\right\|^2_{l_2} = \left\|\Psi\left(\theta_{i+1} - \theta^*_{i+1}\right)\right\|^2_{l_2} = \left\|\theta_{i+1} - \theta^*_{i+1}\right\|^2_{l_2} \le C_1^2 C_b\left(3.317 \times 10^{-3} M_1\sigma^2_{i+1,1} + 2.738 \times 10^{-3} M_2\sigma^2_{i+1,1}\right)$ where θ^*_{i+1} is the solution to

$$min\|\theta_{i+1}\|_{l_1}\,\text{s.t.}\left\|\hat{y}_{i+1} - A\theta_{i+1}\right\|_{l_2} \le \varepsilon \tag{106}$$

and then we have

$$x^*_{i+1} = \Psi\theta^*_{i+1}. \tag{107}$$

Then, we have

$$\begin{aligned}\mathrm{SNR_a} &\ge 10log_{10}\frac{M_1\sigma^2_{i+1,1} + M_2\sigma^2_{i+1,2}}{C_1^2 C_b\left(3.317\times10^{-3}M_1\sigma^2_{i+1,1} + 2.738\times 10^{-3}M_2\sigma^2_{i+1,1}\right)}\\ &\ge 10log_{10}\frac{M_1\sigma^2_{i+1,1}}{C_1^2 C_b\left(3.317\times10^{-3}M_1\sigma^2_{i+1,1} + 2.738\times10^{-3}M_2\sigma^2_{i+1,1}\right)}\\ &= 10log_{10}\frac{u_l}{C_1^2 C_b(3.317 \times 10^{-3}u_l + 2.738 \times 10^{-3}u_h)}\end{aligned} \tag{108}$$

Moreover, for a fixed quantizer, when $\frac{\sigma_{i,1}}{\sigma_{i+1,1}} = 0.75$, we can prove in the same way that there exist a constant C_{f_3} so that

$$\mathrm{SNR}_f \ge 10log_{10}\frac{u_l}{C_1^2 C_{f_3}(8.305\times10^{-3}u_l + 1.534\times10^{-3}u_h)}. \tag{109}$$

And when $\frac{\sigma_{i,1}}{\sigma_{i+1,1}} = 1.25$, we can prove in the same way that there exist a constant C_{f_4} so that

$$\mathrm{SNR}_f \ge 10log_{10}\frac{u_l}{C_1^2 C_{f_4}(4.39\times10^{-3}u_l + 4.2775\times 10^{-3}u_h)} \tag{110}$$

Let $C_p = max\left(C_b, C_{f_3}, C_{f_4}\right)$, and then we can conclude that

$$\mathrm{SNR}_a \ge 10log_{10}\frac{u_l}{C_1^2 C_p(3.317\times10^{-3}u_l + 2.738 \times 10^{-3}u_h)}.$$

When $\frac{\sigma_{i,1}}{\sigma_{i+1,1}} = 0.75$, we have

$$\mathrm{SNR}_f \ge 10log_{10}\frac{u_l}{C_1^2 C_p(8.305\times10^{-3}u_l + 1.534 \times 10^{-3}u_h)}.$$

When $\frac{\sigma_{i,1}}{\sigma_{i+1,1}} = 1.25$, we have

$$\mathrm{SNR}_f \ge 10log_{10}\frac{u_l}{C_1^2 C_p(4.39\times10^{-3}u_l + 4.2775\times10^{-3}u_h)}.$$

Competing interests

The authors declare that they have no competing interests.

Acknowledgement

This work is supported by the National Science Foundation of China (Grant No. 60971129 & 61201326 & 61271335), the National Research Program of China (973 Program) (Grant No. 2011CB302303), the Natural Science Fund for Higher Education of Jiangsu Province (Grant No. 12KJB510021) and the

Scientific Innovation Research Program of College Graduate in Jiangsu Province (Grant No. CXLX11_0408).

Author details

[1]College of Communication and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing, Jiangsu 210003, China. [2]Key Lab of Broadband Wireless Communication and Sensor Network Technology, Nanjing University of Posts and Telecommunications, Nanjing, Jiangsu 210003, China.

Received: 2 December 2011 Accepted: 10 October 2012
Published: 31 October 2012

References

1. D Donoho, Compressed sensing. IEEE Trans Inf Theory **52**(4), 1289–1306 (2006)
2. EJ Candès, *Compressive sampling* (Proceedings of the International Congress of Mathematicians, Madrid, Spain, 2006), pp. 1433–1452
3. RG Baraniuk, Compressive sensing. IEEE Signal Process Mag **24**(4), 118–121 (2007)
4. G Reeves, M Gastpar, *"Compressed" compressed sensing* (IEEE International Symposium on Information Theory, Austin, 2010), pp. 1548–1552
5. TV Sreenivas, WB Kleijn, *Compressive sensing for sparsely excited speech signals* (IEEE International Conference on Acoustics, Speech and Signal Processing, Taipei, 2009), pp. 4125–4128
6. D Giacobello, MG Christensen, MN Murthi, SH Jensen, M Moonen, Retrieving sparse patterns using a compressed sensing framework: applications to speech coding based on sparse linear prediction. IEEE Sigl Proc Letters **17** (1), 103–106 (2010)
7. EJ Candès, T Tao, Decoding by linear programming. IEEE Trans Inf Theory **51**(2), 4203–4215 (2005)
8. RG Baraniuk, MA Davenport, R Devore, MB Wakin, A simple proof of the restricted of isometry property for the random matrices. Constr Approx **28**(3), 253–263 (2008)
9. VK Goyal, AK Fletcher, S Rangan, Compressive sampling and lossy compression. IEEE Signal Process Mag **25**(2), 48–56 (2008)
10. JN Laska, PT Boufounos, MA Davenport, RG Baraniuk, Democracy in action: quantization, saturation, and compressive Sensing. Appl Comput Harmon Anal **31**(3), 429–443 (2011)
11. W Dai, HV Pham, O Milenkovic, *Quantized compressive sensing*, 2009. Arxiv preprint: http://arxiv.org./abs/0901.0749
12. JN Laska, P Boufounous, RG Baraniuk, *Finite range scalar quantization for compressive sensing* (Proc. International Conference On Sampling Theory and Applications, Marseille, 2009), pp. 1433–1452
13. PT Boufounos, RG Baraniuk, *1-Bit compressive sensing. in Proc* (42nd annual Conference on Information Science and Systems, Princeton, NJ, 2008), pp. 16–21
14. Y Baig, EM-K Lai, JP Lewis, *Quantization effects on compressed sensing video* (17th International Conference on Telecommunications, Doha, 2010), pp. 935–940
15. P Boufounos, R Baraniuk, *Quantization of sparse representation*. Data Compression Conference (Snowbird, UT, 2007), p. 378
16. EJ Candès, *The restricted isometry property and its implications for compressed sensing* (Compte Rendus de l'Academie des Science, Paris, 2008), pp. 589–592
17. S Chen, DL Donoho, MA Saunders, Atomic decomposition by basis pursuit. SIAM Rev **43**(1), 33–61 (2001)
18. J Tropp, A Gilbert, Signal recovery from random measurements via orthogonal matching pursuit. IEEE Trans Inf Theory **53**(12), 4655–4666 (2007)
19. W Dai, O Milenkovic, Subspace pursuit for compressive sensing signal reconstruction. IEEE Trans Inf Theory **55**(5), 2230–2249 (2009)
20. DL Donoho, Y Tsaig, JL Strack, Sparse solution of underdetermined linear equation by stagewise orthogonal matching pursuit. IEEE Trans Inf Theory **58**(2), 1094–1121 (2012)
21. D Needell, R Vershynin, Signal recovery from incomplete and inaccurate measurements via regularized orthogonal matching pursuitr. IEEE J Selected Topics in Siganl Processing **4**(2), 310–316 (2010)
22. TT Do, L Gan, N Nguyen, TD Tran, *Sparsity adaptive matching pursuit algorithm for practical compressed sensing* (42nd Asilomar Conference on Signals, Systems and Computer, Pacific Grove, USA, 2008), pp. 581–587
23. BS Atal, MR Schroeder, Adaptive predictive coding of speech signals. Bell Syst Techn J **49**, 1973–1986 (1970)
24. JX Dong, JJ Zhou, YZ Chao, The structure of symmetric r-cyclic matrices and their eigenvalues. J Centl Chin Normal Univ **31**(2), 129–132 (1997)
25. XB Li, RZ Zhao, SH Hu, *Blocked polynomial deterministic matrix for compressed sensing* (6th International Conference on Wireless Communication, Networking and Mobiles, Chengdu, 2010), pp. 1–4
26. L Gan, T DO, T Tran, *Fast compressive imaging using scrambled block hadamard ensemble, in Proc* (European Signal Processing Conference, Switzerland, 2008), pp. 1281–1284
27. HS Chang, Y Weiss, WT Freeman, *Informative sensing of natural images* (IEEE International Conference on Image Processing, Cario, Egypt, 2009), pp. 3025–3028
28. HL Yap, A Eftekhair, MB Wakin, CJ Rozell, *The restricted isometry property for block diagonal matrices* (45th Annual of the Conference on Information Science and Systems, Baltimore, 2011), pp. 1–6
29. JY Park, HL Yap, CJ Rozell, MB Wakin, Concentration of measure for block diagonal matrices with application to compressive sensing. IEEE Trans Signal Process **59**(12), 5859–5875 (2011)
30. GA Gray, GW ZEOLI, Quantization and saturation noise due to analog-to-digital conversion. IEEE Trans Aerosp Elect Syst **7**(1), 222–223 (1971)
31. S Dasgupta, A Gupta, An elementary proof of the Johnson-Lindenstrauss lemma. Random Struc Algorithms **22**(1), 60–65 (2003)
32. JN Laska, MA Davenport, RG Baraniuk, *Exact signal recovery from sparsely corrupted measurements through the pursuit of justice* (43rd Asilomar Confernce on Signals, Systems and Computers, Pacific Grove, 2009), pp. 1556–1560

doi:10.1186/1687-6180-2012-232
Cite this article as: Yunyun and Zhen: **Study on adaptive compressed sensing & reconstruction of quantized speech signals.** *EURASIP Journal on Advances in Signal Processing* 2012 **2012**:232.