International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Speech Compression and Reconstruction Algorithm Based on Multi-Scale Wavelet Packet and Compressed Sensing

Jihua Cao *，Xing Xiong，Jinpeng Yuan

School of electronic engineering, Tianjin University of technology and education, Tianjin 300222, China
* Corresponding author's Email: caojihua@sina.com

Abstract: Wavelet transform is widely applied in the field of the speech signal processing, but the wavelet decomposition is only decomposing the low frequency coefficients, so higher resolution can not be obtained for the high frequency components. In order to further decompose the high frequency coefficient, in this paper, the speech signal was decomposed by multi-scale wavelet packet. By analyzing the sparsity of the decomposed coefficients, the reconstruction algorithm based on compressed sensing was proposed by using the sparse representation of speech signal, at the same time, the effect of compression ratio on the reconstruction quality was analyzed. In determining the compression ratio, speech signal reconstruction quality and coding speed were taken into account, so the proposed compressed and reconstruction method had good real-time property. A large number of simulation experiments showed that the proposed method in this paper had excellent reconstruction quality and efficient coding efficiency.

Keywords: Speech signal; Compressed sensing; Sparse representation; Wavelet transform; Wavelet packet transform

1. Introduction

Over the past 20 years, with the rapid development and the wide application of communication and electronic technology, the speech processing technology has been a significant development at the alarming rate and made a lot of achievements, which promotes the development of communication technology, and the communication between people becomes more and more convenient.

The main purposes of speech coding are to reduce the coding rate, and have a higher reconstruction quality. It should also reduce the decoding delay, and make the complexity of the algorithm as low as possible, so the coding rate, the reconstructed quality, the decoding delay and complexity of the algorithm are the basic index to evaluate a speech coding and decoding performance.

The coding rate reflects the degree of compression on the information, and is proportional to the sampling frequency. For the signal acquisition, the usual practice is to obtain the data according to Nyquist method, but the sampling frequency may not be less than two times the bandwidth of the signal. This raises a higher requirements for the hardware, so a new methods need to be explored in order to reduce the sampling frequency.

Compressed Sensing (CS) theory [1] makes full use of signal sparsity or compressibility, and is not a direct acquisition signal itself, but a small amount of projection of acquisition signal. Through the acquisition of projection value, the exact or approximate reconstructed signal can be achieved, so it is possible to reduce coding rate by reducing the sampling frequency.

The speech signal is close to sparse in the wavelet domain or the Discrete Cosine Transform (DCT) domain, so the compressed sensing theory is used in the processing of the speech signals. At present, the application of the compressed sensing theory has got involved with many fields, such as: compressed sensing radar[2-6], Distributed Compressed Sensing (DCS) theory[7,8,9], wireless sensor network[10,11], image acquisition equipment development [12,13], medical image processing [14,15], Analog-to-Information[16,17], spectral analysis [18-19], hyperspec-

tral image processing and remote sensing image processing[20,21,22]. However research on the compressed sensing of the speech signal is still in the early stage. Gemmeke et al.[23] applied the compressed sensing theory to speech recognition and makes the speech recognition system to obtain better anti-noise performance.

It can be predicted that based on the study of the compressed sensing performance on speech signal a new mode will be set up for observation, then there might be a major breakthrough in the existing theory and technology on speech compression, speech recognition, speech synthesis and speech enhancement, so the research on the compressed sensing performance on speech signal has important theoretical significance and practical value.

In this paper, based on the sparse representation of the speech signal, the speech compression and reconstruction algorithm by using multi-scale wavelet packet was proposed, and the validity of simulation experiment enriched the application achievement of the compressed sensing.

The remainder of the paper is organized as follows. Section 2 presents the principle of the wavelet packet decomposition and the sparse analysis of the wavelet packet decomposition coefficient and DCT coefficient. In Section 3 , the compressed sensing algorithm for speech signal is proposed. The simulation experiments and results analysis are provided in Section 4. We make a summary to the full text in Section 5.

2. Sparse Analysis on the Wavelet Packet Decomposition Coefficients of the Speech Signal

2.1 Wavelet packet decomposition

Fast binary wavelet packet decomposition algorithm for signal $f(t)$ is as follows:

$$\begin{cases} d_0^0(n) = f(n) \\ d_{j+1}^{2i}(n) = \sum_k h_0(k-2n) d_j^i \quad i = 0,1,.....,2^j - 1 \\ d_{j+1}^{2i+1}(n) = \sum_k h_1(k-2n) d_j^i \end{cases} \quad (1)$$

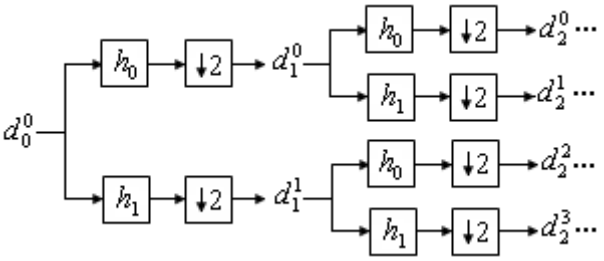
where $h_0(n)$ and $h_1(n)$ are a pair of quadrature mirror filter for decomposition, and $h_0(n)$ is low pass filter, and $h_1(n)$ is high pass filter. In the time domain, they satisfy equation (2):

$$h_1(k) = (-1)^k h_0(1-k). \quad (2)$$

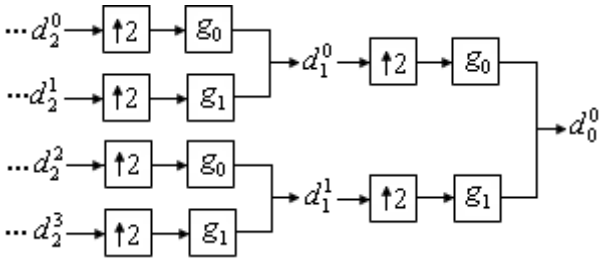
The reconstruction equation of wavelet packet transform is expressed by equation (3).

$$d_j^i(t) = \sum_k g_0(t-2k) d_{j+1}^{2i} + \sum_k g_1(t-2k) d_{j+1}^{2i+1} \quad (3)$$

where $g_0(n)$ and $g_1(n)$ are a pair of quadrature mirror filter for reconstruction, and $g_0(n)$ is low pass filter, and $g_1(n)$ is high pass filter. The algorithms of wavelet packet decomposition and reconstruction are shown in Figure 1 (a) and (b).



(a) Wavelet packet decomposition algorithm



(b) Wavelet packet reconstruction algorithm

Figure 1 Typical wavelet packet transform algorithm

In order to achieve accurate reconstruction, the wavelet filter must meet the following conditions

$$H_0(\Omega)G_0(\Omega) + H_1(\Omega)G_1(\Omega) = 1, \quad (4)$$

$$H_0(\Omega+\pi)G_0(\Omega) + H_1(\Omega+\pi)G_1(\Omega) = 0, \quad (5)$$

where $H_0(\Omega)$, $G_0(\Omega)$, $H_1(\Omega)$ and $G_1(\Omega)$ are the Fourier transform of $h_0(n)$, $g_0(n)$, $h_1(n)$ and $g_1(n)$ respectively.

In the family of functions generated by wavelet packets, wavelet packet can further decomposed the high frequency components, which can make the sound more clearly. According to the characteristics of signal decomposition, the suitable wavelet packet

decomposition tree was choose, which improved the time-frequency resolution and provided more precise decomposition method. Therefore, for the sake of observing high-frequency components more carefully, this paper adopted the wavelet packet decomposition. In order to reduce the complexity of the algorithm, this paper used *db3* wavelet packet to realize the 2 level decomposition of the speech signal, and look for the sparse representation of the speech signal. The decomposition process is shown in Figure2, in the diagram, the nodes (2,0) represents the lowest frequency coefficient by two layer wavelet packet decomposition, and (2,1), (2,2) and (2,3) represents the sub low frequency coefficients, the sub high frequency coefficients and the highest frequency coefficients respectively.

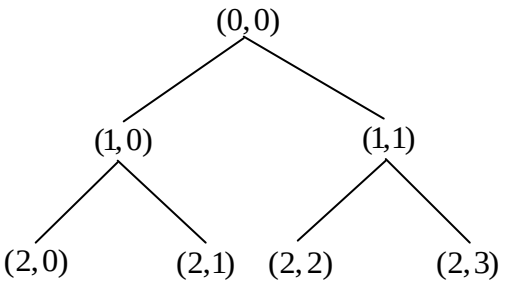


Figure 2 The tree structure of the two layer wavelet packet decomposition

2.2 Sparse analysis

Take the male and female student speech signal as an example, after being decomposed by wavelet packet, the sparsity of male and female speech signals was relatively similar. A frame (the length was selected as 320 points) male speech signal was decomposed by the two layer wavelet packet, where Figure 3 is a male original speech frame signal, and Figure 4 is the decomposed results.

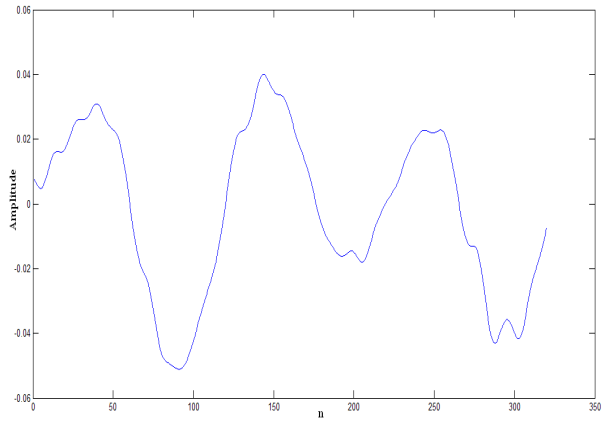


Figure 3 The original speech frame signal (male)

From Figure 4, it can be seen that the node (2,0) retained a lot of information of the original speech frame signal, and this node did not have sparse characteristics and the dynamic range of the coefficients was relatively large. However, the coefficients of the node (2,1), (2,2) and (2,3) were close to zero, so the sparseness of the three nodes was better. But when DCT was done on the node (2,0), the *n*th DCT transform coefficient was equation (6).

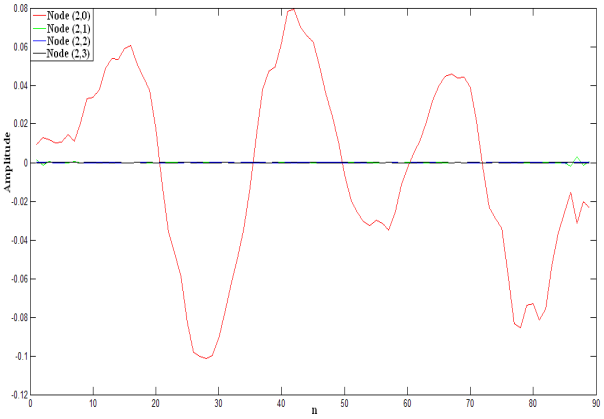


Figure 4 The two layer wavelet packet coefficients (male)

$$\alpha_n = w[n]\sum_{i=1}^{N} x[i]\cos\left(\frac{\pi(2i-1)(n-1)}{2N}\right), \quad (6)$$

where $x[i]$ is the ith coefficient of the nodes (2,0), and $w[n] = \begin{cases} \frac{1}{\sqrt{N}}, & n=1 \\ \sqrt{\frac{2}{N}}, & 2 \le n \le N \end{cases}$. Equation (6) can be written in matrix form:

$$\boldsymbol{\alpha} = \boldsymbol{\psi}_{DCT}^{T}\mathbf{x} \quad . \quad (7)$$

The DCT coefficients of the node (2,0) are shown in Figure 5. From Figure 5 it can be seen that most of the DCT coefficient was close to zero, and only a small part of the absolute values of the coefficients were relatively large. This showed that the node (2,0) in the DCT domain was close to sparse.

Next, the sparsity of the wavelet packet coefficients of the female speech signal was analyzed. Figure 6 is a female original speech frame signal, and Figure 7 is the decomposed results. From Figure 7, it can be seen that the node (2,0) retained a lot of information of the original speech

frame signal, and it was known that the sparsity of the male and female speech signal was relatively similar by the absolute amplitude of the node (2,1), (2,2) and (2,3).

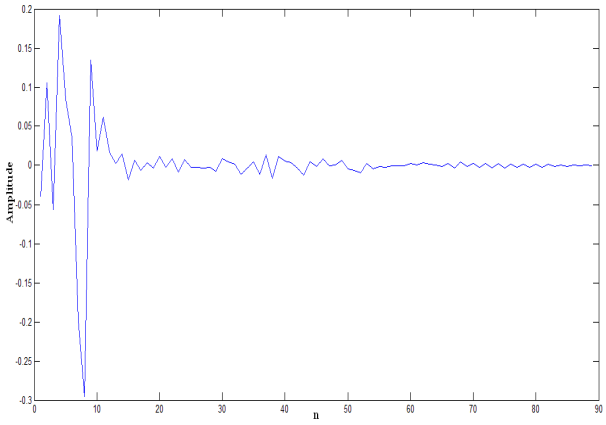


Figure 5 The coefficients of the node (2,0) in DCT domain

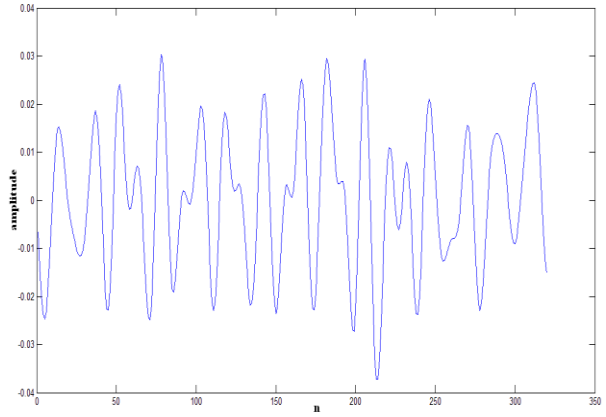


Figure 6 The original speech frame signal (female)

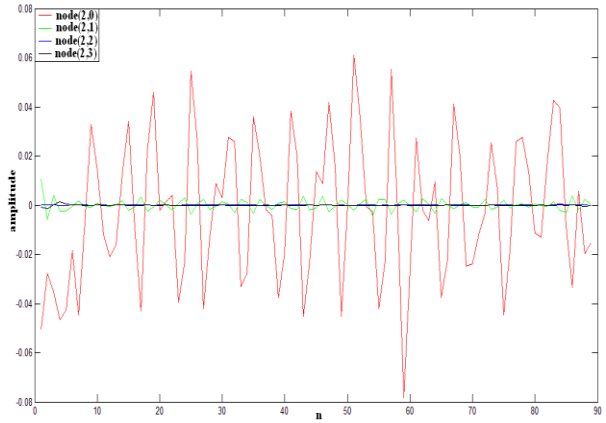


Figure 7 The original speech frame signal (female)

3. Compressed Sensing of the Speech Signal

Through the analysis of the speech signal, it can be seen that the speech signal was approximately sparse, so the speech signal met the prerequisite of applying the compressed sensing theory. *N* samples of the linear projection in a speech frame was obtained by the observation matrix associated not with the sparse basis of the speech signal. In order to save sources, a small amount of sample observations were stored and transferred. According to the theory of compressed sensing, the original speech frame can be recovered through a sequence of observations received at the receiver. The following random Gauss matrices could be taken as the observation matrix, explaining the specific steps of compression sampling and reconstruction for the *i*th frame speech signal

(1)Projection

To structure ***M*×*1*** observation matrix y_i for x_i

$$\mathbf{y}_i = \mathbf{\Phi}_i \mathbf{x}_i, \tag{8}$$

where $\mathbf{\Phi}_i$ is the $M \times N$ ($M < N$) dimension of the random Gauss observation matrix.

(2)Reconstruction

To solve the optimization problem to reconstruct the DCT coefficient of x_i or the second order wavelet packet decomposition coefficient $\hat{x}_{ci}$

$$\hat{x}_{ci} = \arg\min \| x_{ci} \|_1 \ s.t. \ y_i = \Phi_i x_i = \Phi_i C_n x_i, \tag{9}$$

where C_n is a $N \times N$ dimensional DCT matrices or two level wavelet packet coefficient matrix, and equation (9) is the coefficient of BP reconstruction algorithm, and we can also use the coefficient of OMP reconstruction algorithm.

By the coefficient $\hat{x}_i$ to restore the original speech signal x_i

$$x_i = C_n^T \hat{x}_{ci}, \tag{10}$$

where C_n^T is the transpose matrix of C_n, That is, $C_n^T = C_n^{-1}$.

For a period of speech signal x, all it frames were processed by using the above method, combining all the reconstructed finally and then, the reconstruction of the original speech signal will be obtained.

This paper used the average frame reconstruction SNR to estimate the reconstruction quality of compressed sensing for speech signal, and defined as follows.

$$AFSNR = \frac{1}{K}\sum_{k=1}^{K} 10\log 10\left(\frac{\| x_k \|_2^2}{\| x_k - \hat{x}_k \|_2^2}\right), \quad (11)$$

Where K is the total number of frames， and x_k and $\hat{x}_k$ are the *K*th frame speech signal and its reconstructed signal respectively. The compression ratio is defined as the ratio of the number of measurements per frame and frame length. When r is fixed, the bigger $AFSNR$ value shows better reconstruction performance.

4. Simulation Experiments and Results Analysis

Experiment 1： This experiment explained that the choice of wavelet and wavelet packet impacted on the quality of the reconstructed speech signal. In code side, a window was added to the recorded speech signal, dividing it into frames. Both the reconstruction quality and the bit rate of the speech signal, the length of the frame was selected as the 320 point, then the three layers of wavelet transform of the speech frames was calculated, and the wavelet base selection was *sym*7. The high frequency wavelet coefficients were processed by the compressed sensing theory, the low frequency coefficients were not processed, then measurement values were coded with *A* law for non-uniform PCM quantization. The decoding process and the coding process is just the opposite, and we will get to the average frame reconstruction SNR under different observation numerical. Then the three layer wavelet decomposition were transformed into wavelet packet decomposition of the two layer, the decoder change accordingly. The simulation results of the two schemes are shown in Figure 8.

From Figure 8, it can be seen that under the same observation points, the average frame SNR of the wavelet packet transform was higher than the ones of the wavelet transform. From the diagram, it can be seen that:

(1) The average frame SNR of the wavelet packet transform in the left side of the curve was much higher than the ones of the wavelet transform, and the reason was that the low frequency coefficients of the two layers of wavelet packet transform and the three layers of wavelet transform was different, and the low frequency coefficients of the two layers of wavelet packet transform was 89 while the low frequency coefficients of the three layers of wavelet transform was 57.

(2) From left to right, The growth rate of the curve of the wavelet packet transform was far less than the ones of wavelet transform, and the reason was that the wavelet packet transform not only decomposed low frequency coefficients, but also decomposed high frequency coefficients, while the wavelet transform only decomposed low frequency

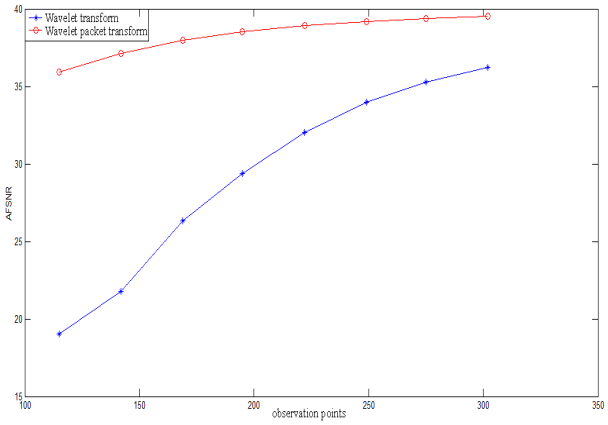


Figure 8 The average frame SNR of the wavelet and wavelet packet transform of the speech signal

coefficients. This made that the sparse characteristics of the high frequency coefficients of the wavelet decomposition was much weaker than the ones of the wavelet packet decomposition, that is, the sparse characteristic of the wavelet packet decomposition

Table 1 under different compression ratios the average frame SNR of the reconstructed speech signal

	q3=0.1	q3=0.2	q3=0.3
q1=0.1,q2=0.1	35.5153	35.4706	35.4217
q1=0.1,q2=0.2	35.5699	35.5005	35.5308
q1=0.1,q2=0.3	35.4628	35.5160	35.5606
q1=0.2,q2=0.1	35.9614	36.0579	36.1192
q1=0.2,q2=0.2	36.0097	35.9772	36.1075
q1=0.2,q2=0.3	36.0351	36.0127	36.1317
q1=0.3,q2=0.1	37.0635	37.1451	37.1467
q1=0.3,q2=0.2	37.1726	37.1296	37.1964
q1=0.3,q2=0.3	37.0725	37.0544	37.1462

was better, so the wavelet packet required less observation points and there was a better reconstructed quality. However, if it continued to increase observation points, the ascension of the average frame SNR of the reconstructed speech signal would be relatively small, consequently the growth slower.

Experiment 2：In the second experiment, the node (2,1), node (2,2) and node (2,3) with different compression ratio of q1, q2, q3 were studied, and their influence on the quality of the reconstructed speech signals. When q1, q2, q3 were taken as 0.1, 0.2, 0.3 respectively, the average frame SNR of the reconstructed speech signal are shown in Table1.

From Table 1 it can be seen that when q1 and q2 were fixed and q3 was taken different values, the difference between the average frame SNR was within 0.15dB. When q1 and q3 were fixed and q2 was taken different values, the difference between the average frame SNR was within 0.15dB too. However, when q2 and q3 were fixed and q1 was taken different values, the difference between the average frame SNR was between 0.4dB and 1.6dB. The reason was that the sparsity of the wavelet packet coefficients of the node (2,2) and node (2,3) were more similar, but the sparsity of the wavelet packet coefficients of the node (2,1) was relatively poor.

Experiment 3：In the above experiments, we used the non-uniform PCM method and completed the quantization coding of all nodes. Since a better reconstructed signal requires higher coding rate, in this experiment, a different encoding method was used to get a better reconstructed speech quality an very low bit rates. The experimental studied speech signal compression and reconstruction method based on wavelet packet transform and compressed sensing(Figure 2 signal as an example).

In the experiment, frame length was 320 points, and frame shift was 160 points, and the node (2,2) and node (2,3) compression ratio q1 and q2 were determined as 0.1. Taking into account the observation sequence variation range of the node (2,1), node (2,2) and node (2,3) was very small, the observation sequence of three nodes were coded by using uniform PCM quantization method, quantization order by K1, K2, K3 said respectively, and their value were set to 4. For the node (2,0), because its coefficient variation range was bigger, the sequence of the node (2,0) was coded by using vector quantization method. If the uniform PCM quantization coding is used, more orders will be needed to result in high code rate. Therefore, vector quantization coding was used, and vector quantization codebook N was 128, 256 and 512 respectively, and q1 was 0.1, 0.2 and 0.3 respectively. The average frame SNR of the reconstructed speech signal are shown in Table 2.

Table 2 In different codebook sizes and different q1, the average frame SNR of the reconstructed speech signal

	N=512	*N*=256	*N*=128
r_1=0. 1	18. 4169	8. 2805	4. 3134
r_1=0. 2	18. 7830	8. 4254	4. 3683
r_1=0. 3	19. 3983	8. 6346	4. 4430

From Table 2, it can be seen that when *N* was the same, the average frame SNR of the reconstructed speech signal had smaller enhance with the increase of q1. However, when q1 was the same, the average frame SNR of the reconstructed speech signal had larger enhance with the increase of N. When N was 512, the average frame SNR of the reconstructed speech signal was close to 20dB. As for the quality of the reconstructed speech signals and code rate, taking q1=0.3, N=512, when K1, K2 and K3 were increased, the average frame SNR of the reconstructed speech signal improve somewhat but very little. So in speech processing, all the parameters were identified as: q1=0.3, q2=q3=0.1, N=512 and K1=K2=K3=4 respectively. the OMP reconstruction algorithm was used, and the average frame SNR of the reconstructed speech signal was 19.3983dB, and coding rate was 9.40kb/s

5. Conclusion

In this paper, through the sparse analysis on the wavelet packet and DCT coefficients for the speech signal, and a speech compression and reconstruction method was proposed based on wavelet packet transform and compressed sensing. In view of the reconstruction quality and coding rate, the selection of the parameters of the proposed algorithm was studied. After the algorithm parameter was determined, the average frame SNB of the reconstructed speech signal and coding rate were calculated, and the effectiveness of the algorithm was verified.

Acknowledgment

This work was supported by the National Nature Science Foundation of China under grant number

61271412.

Reference

[1] Donoho D,"Compressed sensing",*IEEE Trans. Information Theory*, pp.1289-1306, 2006.

[2] Jindong Zhang, Daiyin Zhu,"Adaptive Compressed Sensing Radar Oriented Toward Cognitive Detecti on in Dynamic Sparse Target Scene", *IEEE Trans. Signal Processing*, pp.7137-7140,2012.

[3] Li Fanghua, Zeng Fanzi, "Radar Target Location Based on Compressive Sensing Technique", *Proceedings of the International Conference on Comp uter Science & Service System*. Nanjing, China, pp.1018-1021, 2012.

[4] Potter L.C.,Ertin E.,Parker J.T.,Cetin M, "Sparsity and Compressed Sensing in Radar Imaging", *Proceedings of the IEEE*, pp.1006-1020,2010.

[5] Qilian Liang, "Compressive Sensing for Radar Sensor Networks", *Proceedings of the International Conference on Global Telecommunications Conference*, Miami, FL, pp.1-5, 2010.

[6] Yao Yu,Athina P,Petropulu H,Vincent Poor,"CSSF MIMO RADAR:Compressive-Sensing and Step-Frequency Based MIMO Radar", *IEEE Trans. Aerospace and Electronic Systems*, pp.1490-1504, 2012

[7] Vinuelas-Peris P,Artes-Rodriguez A, "Sensing matrix optimization in Distributed Compressed Sensing", *IEEE/SP 15th Workshop on Statistical Signal Processing*, Cardiff, England, pp.638-641, 2009.

[8] Wenger S, Ament M, Guthe S etc, "Visualization of Astronomical Nebulae via Distributed Multi-GP U Compressed Sensing Tomography",*IEEE Trans. Visualization and Computer Graphics*,pp.2188-219 7, 2012.

[9] Zhuang Zhe Min, Wu Li Ke, Li Fen Lan, "Distributed Compressed Fire Signal Sensing Based on Unbalance Expander", *Fifth International Conference on Intelligent Computation Technology and Automation*. Zhangjiajie, China, pp.486-489, 2012.

[10] Nguyen N, Jones D.L, Krishnamurthy S, "Netcompress: Coupling network coding and compressed sensing for efficient data communication in wireless sensor networks", *IEEE Workshop on Signal Processing Systems*, San Francisco, CA, pp.356-361, 2010.

[11] Pudlewski S,Prasanna A,Melodia T, "Compressed-Sensing-Enabled Video Streaming for Wireless Multimedia Sensor Networks", *IEEE Trans.Mo -bile Computing*, pp.1060-1070, 2012.

[12] Yan Zhou,Yong Zhong,Dong Wang, "The Reconstruction of High Resolution Image Based on Compressed Sensing", *International Conference on Machine Learning and Cybernetic*, Qingda o, China, pp.828-832, 2010.

[13] Zhaorui Liu,Elezzabi A.Y,Zhao H.V, "Maximum Frame Rate Video Acquisition Using Adaptive Compressed Sensing", *IEEE Trans. Circuits and Systems for Video Technology*,pp.1704-1718,201 1.

[14] Majumdar A,Ward R.K,Aboulnasr T,"Compresse -d Sensing Based Real-Time Dynamic MRI Rec -onstruction",*IEEE Trans.Medical Imaging*,pp.225 3-2266, 2012.

[15] Murphy Mark, Alley M, Demmel J, "Fast l_1 -SP IRIT Compressed Sensing Parallel Imaging MRI: Scalable Parallel Implementation and Clinically Feasible Runtime", *IEEE Trans. Medical Imag ing*,pp.1250-1262, 2012.

[16] Trakimas M,Hancock T,Sonkusale S, "A Compr -essed Sensing Analog-to-Information Converter with Edge-Triggered SAR ADC Core", *IEEE International Symposium on Circuits and System s*. Seoul, Korea, pp.3162-3165, 2012.

[17] Mangia M, Rovatti R, Setti G, "Rakeness in th eDesign of Analog-to-Information Conversion of Sparse and Localized Signals", *IEEE Trans. Ci r-cuits and SystemsI:Regular Papers*,pp.1001-101 4,2012.

[18] Zhenghao Zhang, Zhu Han, Husheng Li, "Belief Propagation Based Cooperative Compressed Spectrum Sensing in Wideband Cognitive Radio Ne tworks", *IEEE Trans. Wireless Communications*, pp.3020-3031, 2011.

[19] Ji-xin Liu, Quan-Sen Sun., "Mass spectrum data processing based on compressed sensing recognition and sparse difference recovery", *9th Inte -rnational Conference on Fuzzy Systems and K -nowledge Discovery*, Sichuan, China,pp. 1413-1417,2012.

[20] Aguilera E,Nannini M,Reigber A., "Multisignal Compressed Sensing for Polarimetric SAR To mography", *IEEE Geoscience and Remote S-en sing Letters*, pp.871-875, 2012.

[21] Aguilera E,Nannini M,Reigber A, "A Data-Adap tive Compressed Sensing Approach to Polarime -tric SAR Tomography of Forested Areas",*IEE EGeoscience and Remote Sensing Letters*, pp.5 43-547, 2013.

[22] Potter L.C,Ertin E,Parker J.T etc, “Sparsity and Compressed Sensing in Radar Imaging”, *Proc eedings of the IEEE*, pp.1006-1020, 2010.

[23] Gemmeke J.F, Virtanen T, Hurmalainen A, “Ex -emplar-Based Sparse Representations for Noise Robust Automatic Speech Recognition”, *IEEE Trans. Audio Speech and Language Processi ng*,pp.2067-2080, 2011.