An Adaptive Compressed Sensing Method in Speech

Xu Feng, Wang Xia, Zheng Xiao-Dong, Wang Hao

School of Electrical and Information Engineering, HuaiNan Normal University, China, Email: xufeng771125@sina.com

Abstract

The application of an adaptive compressive sensing method in the speech signal processing is proposed in this paper. First, the threshold of wavelet transform is used to preprocess the speech signal. Then, according to the parameters of the speech frame, each frame is adaptively assigned a measurement number. Finally, the measurement matrix is used to reconstruct the speech signal. Experimental results show that the proposed method can improve the SNR in the speech signal reconstruction and is of high robustness in different noise intensity.

Keywords: Speech Signal, Compressed Sensing, Adaptive, Measurement Matrix

1. Introduction

Compressed sensing (CS) [1,2], which have arisen in recent years, combines the sampling and compression technologies. The Nyquist theorem states that a signal must be sampled at least twice as fast as the bandwidth of the signal to accurately reconstruct the waveform. The CS technique (theory) states that if the (original) signal is sparse in some known transform domain, then, by utilizing the projections of the original signal into the known transform domain, the original signal can be approximately reconstructed. When using the CS technique, the number of measurements required to reconstruct the original signal is closely related to the sparsity of the signal and is irrelevant to the original signal’s bandwidth in the traditional frequency domain. In view of this property of CS technique, the sampling cost in the signal processing can be reduced. And undoubtedly, the CS technique is of great research importance and of high practical value.

This paper deals with the CS technique that is applied in the speech signal processing. Due to the physiological structure of human pronunciation and the characteristics of speech signal, the speech signal can be viewed as a sparse signal. Two perspectives the speech signal can be taken as a sparse signal are given as follows: 1) Because of the discontinuity of human speech, the speech signal is obviously a sparse signal in time domain; 2) The speech signal can be approximately viewed as sparse if it is represented in the discrete cosine transform domain [12]. Therefore, the CS technique can be employed in the speech signal processing. However, there is not much research on the application of CS technique to speech signal processing. In 2009, based on the sparsity of speech signal in the complete dictionary, the CS technique is used to process the speech signal by Tingting Xu [9]. In 2011, Another category of speech signal processing using CS technique is proposed from the perspective of wavelet transform [10]. Guo Haiyan proposed the representation of speech signal in approximate KLT domain in 2009 and studies the application of CS technique in the speech signal processing [11].

In the previous literature on CS technique, the construction of measurement matrix is generally not adaptive and only depends on the sparsity in some (transform) domain. The above CS technique is called non-adaptive CS technique. This paper proposes an adaptive CS technique that is applied in the speech signal processing. The characteristics of the speech signal are exploited to adaptively construct the measurement matrix when applying the CS technique to the speech signal processing. The details of our proposed method are given in...
Section 3. The rest of the paper is organized as follows: Section 2 introduces the framework of speech signal processing using CS technique and the reconstruction algorithm based on GPSR-BB. Section 3 details our proposed method, namely speech signal processing using adaptive CS technique. Experimental results are given in Section 4 in order to illustrate the effectiveness of our proposed method.

2. Speech signal CS framework and GPSR-BB algorithm

2.1 Speech signal CS framework

Suppose speech signal is \( x \), signal length is \( L_{sp} \). If \( L_{sp} \) is very large, constructing \( L_{sp} \) length orthogonal matrix is unrealistic, and which do not correspond speech short-time characteristics. Therefore, the signal needs to be divided into frames [13].

Frame length is \( l = L_{sp} / N_{mq} \), \( N_{mq} \) is the frames number, and usually frame length \( l \) is 2 integer power. Suppose orthogonal wavelet base \( \Psi_i = [\psi_{i,1}, \psi_{i,2}, \cdots, \psi_{i,N}] \), and

\[
x_i = \sum_{j=1}^{N} \theta_{i,j} \psi_{i,j},
\]

where \( x_i \) represents the i-th frame signal, \( \theta_{i,j} \) is the corresponding coefficient of \( x_i \) in the transform domain \( \Psi_i \), while \( \| \Theta \| = K \), the signal \( x_i \) is called K-sparse.

As the transformation matrix \( \Psi_i \) is an orthogonal matrix, then \( x_i = \Psi_i^T \Theta_i \). Design a stable \( m \times n \) dimension measurement matrix \( \Gamma_i \) that irrelevance to the transformation base \( \Psi_i \),

Observe \( \Theta_i \), then obtain measurement set \( y_i = \Gamma_i \Theta_i = \Gamma_i \Psi_i^T x_i \). CS theory supposed that more sparse the signal is, the better the performance of reconstruction is. Therefore, using 0-norm optimization problem to solve \( x_i \) precise or approximate \( \hat{x}_i \).

\[
\min \| \Psi_i^T x_i \|_0 \quad \text{s.t.} \quad \Gamma_i \Psi_i^T x_i = y_i
\]  
(1)

Because the 0-norm optimization problem is actually the NP problem, Thus equation (1) is often converted to equation (2).

\[
\min \| \Psi_i^T x_i \|_1 \quad \text{s.t.} \quad \Gamma_i \Psi_i^T x_i = y_i
\]  
(2)

Let \( s_i = \Psi_i^T x_i \), simultaneously, equation (2) will change the form

\[
F = \min \frac{1}{2} \| y_i - \Gamma_i s_i \|_2^2 + \lambda \| s_i \|_1
\]  
(3)

To solve the optimization problem and get the transform domain \( s_i \), then through inverse transform, obtain the original signal \( \hat{x}_i \), consequently, combination of all the obtain \( \hat{x}_i \) sub-frame signal and then get the original speech signal \( \hat{x} \).

2.2 GPSR-BB algorithm

Algorithm GPSR-Basic ensures that the objective function \( F \) decreases at every iteration. Recently, considerable attention has been paid to an approach that does not have this property. This approach was originally developed in the context of unconstrained minimization of a smooth nonlinear function \( F \). It calculates each step by the formula \( \delta^{(k)} = -H_k \nabla F(z^{(k)}) \),

where \( H_k \) is an approximation to the Hessian of \( F \) at \( z^{(k)} \). Barzilai and Borwein propose a particularly simple choice for the approximation \( H_k : H_k = \eta^{(k)} I \), where \( \eta^{(k)} \) is chosen so that this approximation has similar behavior to the true Hessian over the most recent step, that is

\[
\nabla F(z^{(k)}) - \nabla F(z^{(k-1)}) \approx \eta^{(k)} [z^{(k)} - z^{(k-1)}]
\]  
(4)
with \( \eta(k) \) chosen to satisfy this relationship in the least-squares sense. In the unconstrained setting, the update formula is

\[
z^{(k+1)} = z^{(k)} - (\eta^{(k)})^{-1}\nabla F(z^{(k)})
\]

(5)

Figueiredo extends the BB approach to GPSR algorithm in [14]. They choose \( \lambda_k \) as the exact minimizer over the interval \([0,1]\) and choose \( \eta^{(k)} \) at each iteration in the manner described above, except that \( \alpha^{(k)} = (\eta^{(k)})^{-1} \) is restricted to the interval \([\alpha_{\min}, \alpha_{\max}]\). The GPSR-BB algorithm is simply described as following:

**Step 0 (initialization):** Given \( z^{(0)} \), choose parameters \( \alpha_{\min} \) and \( \alpha_{\max} \), \( \alpha^{(0)} \in [\alpha_{\min}, \alpha_{\max}] \), and set \( k = 0 \).

**Step 1:** Compute step:

\[
\delta^{(k)} = (z^{(k)} - \alpha^{(k)} \nabla F(z^{(k)}))_+ - z^{(k)}
\]

(6)

where \( + \) denotes only when the evaluation of the expression in the bracket is positive; the expression with + is equal to the expression inside the bracket, otherwise, it is equal to zero.

**Step 2 (line search):** Find the scalar \( \lambda^{(k)} \) that minimizes \( F(z^{(k)} + \lambda^{(k)} \delta^{(k)}) \) on the interval \( \lambda^{(k)} \in [0,1] \), and set

\[
z^{(k+1)} = z^{(k)} + \lambda^{(k)} \delta^{(k)}
\]

(7)

**Step 3 (update \( \alpha \)):** compute

\[
y^{(k)} = (\delta^{(k)})^T B \delta^{(k)}
\]

(8)

if \( y^{(k)} = 0 \), let \( \alpha^{(k+1)} = \alpha_{\max} \), otherwise \( \alpha^{(k+1)} = \text{mid}\{\alpha_{\min}, \frac{\|y^{(k)}\|_1}{y^{(k)}}, \alpha_{\max}\} \).

**Step 4:** Perform convergence test and terminate with approximate solution \( z^{(k+1)} \) if it is satisfied; otherwise set \( k + 1 \rightarrow k \) and return to **Step 1**.

### 3. Speech signal adaptive compressed sensing

The speech signal is processed frame by frame in CS technique. According to the short-time stationary, the signal is sampled at 16000Hz and frame length is 20ms, namely, sampling point \( l \) is 320. In traditional CS method, measurement number \( m \) and measurement matrix \( \Gamma_i \) of each frame signal are fixed. Intuitively, information of each speech frame is not the same, therefore, in order to better reconstruct the original signal, the frame containing amount of information assign more measurement numbers, and frames contained less information allocated fewer measurement numbers. However, adaptive CS method first according to the type and the number of each speech frame of the entire speech, assign each frame measurement number \( m_i \), then according to the completely measurement components size of each speech frame, select \( m_i \times l \) dimensional measurement vector and the corresponding \( m_i \times l \) dimensional measurement matrix. In the reconstruction side, the process method is same with non-adaptive CS. Adaptive obtain measurement vector and measurement matrix of each frame, using reconstruction algorithms reconstruct the sparse coefficients of each speech frame, and finally obtains each frame reconstruction speech.

Adaptive algorithm using the speech frame type to realize high efficient measurement. Namely, voiced frame that contain more information assign more measurement number, and unvoiced frames which contain less information allocate less measurement number. First, according to the energy and zero-crossing rate of each frame to determine the signal is voiced or unvoiced, then according to each frame type and each type number of the entire speech allocate its measurement number.

The realizations of the methods are as follows:
The speech signal is divided into frames, frame length 20ms, obtain all speech frames $x_i$ ($1 \leq i \leq l$, $l$ is the total number of frames).

Compute each speech frame energy and zero-crossing rate, the formula is as follows;

$$E_i = x_i^T x_i$$  \hspace{2cm} (9)

$$Z_i = \frac{1}{2} \sum_{j=2}^{n} \text{sgn}[x_i(j)] - \text{sgn}[x_i(j-1)]$$  \hspace{2cm} (10)

$n$ is the number of sampling of per frame, in this paper $n = 320$.

According to each frame energy $E_i$ and zero-crossing rate $Z_i$, calculate each speech frame zero to energy ratio $EZR_i = \frac{E_i}{Z_i}$. If $EZR_i \leq T$, The frame can be judged to be unvoiced frames. Else $EZR_i > T$, the frame can be judged to be voiced frames. The threshold $T$ is 25. Through calculation of a large number of different male and female stable voiced frames and stable unvoiced frames, compute zero to energy ratio. Each frame according to the frame type which was determined in the forward step and the unvoiced frame number $L_1$ and voiced frame number $L_2$ of the entire speech, assign measurement number.

If $x_i$ is an unvoiced frame, then the frame measurement number

$$m_i = \frac{M}{L_1 + 1.5L_2}$$  \hspace{2cm} (11)

If $x_i$ is an voiced frame, then the frame measurement number

$$m_i = \frac{1.5M}{L_1 + 1.5L_2}$$  \hspace{2cm} (12)

$M$ is the total measurement number of the entire speech. Determine $m_i$ and then obtain random measurement matrix $\Gamma_i$.

4. Experimental results and analysis

Experimental compared method proposed in this paper the GPSR-BB algorithm reconstruction performance. In order to facilitate display, all the experimental data of reconstruction algorithm are taken from AV16.3 database 'seq01-1p-0000_array1_mic2.wav' speech segment, and intercept a speech which the sampling points from 136072 to 201607, speech content "one two three four".
Figure 1. Waveform of original speech and reconstructed speech

Figure 1 is the effect chart of GPSR-BB algorithm, sparse and reconstruction signal. Figure 1(a) presents the original speech signal, Figure 1(b) shows the sparse signal, wavelet threshold \( q = 0.07 \), sparseness 0.011. Figure 1(c) shows reconstruction signal that without using adaptive processing method, Figure 1(d) reconstruction signal that the use of adaptive processing method. The result showed that Figure 1(c) signal is heavily contaminated.

To define \( \text{SNR} = 20 \log \left( \frac{\| \hat{x} \|_2}{\| x - \hat{x} \|_2} \right) \). Compared with Figure 1(d), the SNR reconstructed signal without using adaptive processing method is low 5dB, while the CPU run time is 3 times higher.
Figure 2 shows the threshold and signal reconstruction performance relationship. The figure shows that within a certain range, the greater the threshold $q$ and the greater the sparseness of signals, then the reconstructed signal SNR will be higher. However, when the threshold is too large, it will filter out some useful information in signal, sequentially, making it impossible to restore the original signal. Therefore, a suitable threshold directly affects the signal reconstruction performance, as seen from Fig.4, when the SNR separately equal to 5dB, 0dB, 5dB and 10dB, system best threshold $q$ respectively be 0.0880, 0.0730, 0.0710 and 0.0680. Obviously, the optimal threshold $q$ will decreases with the increase of the SNR. Because the lower SNR, the relevant energy of noise signal will be greater, and wavelet transform coefficients will also correspondingly increased, consequently, leading to threshold $q$ increased. When $q$ reaches a certain value, continue increase $q$, SNR decreased, because excessive threshold filtering out some useful information, consequence, unable to reconstruct the original signal.

Experiment 3 compared in different SNR, adaptive strategy and non-adaptive strategy reconstruction performance. Speech signal SNR changes from 10dB to 20dB. As shown in chart, the output SNR of reconstruction signal of adaptive strategy is better than non-adaptive strategy. Especially in low SNR, it approximate increased 3dB. However, as the SNR increases, the improvement degree will be descend. Comprehensive comparison, the speech signal reconstruction performance is better when applying adaptive strategy, especially in low SNR.
5. Conclusions

This paper investigates the application of the CS technique to the speech signal processing. An adaptive CS technique is proposed to be employed in the speech signal processing. First, the threshold of wavelet transform is used to preprocess the speech signal. Then, according to the parameters of the speech frame, each frame is adaptively assigned a measurement number. Finally, the measurement matrix is used to reconstruct the speech signal. Experimental results show that the proposed method can improve the SNR in the speech signal reconstruction and is of high robustness in different noise intensity. How to exploit the characteristics of the speech signal in order to adaptively construct the measurement matrix will be further studied in our future research work.

6. Acknowledgment

This work is partially supported by Open Research Foundation of Key Laboratory of Underwater Acoustic Signal Processing of Ministry of Education of Southeast University under Grant (No.UASP1003), and Special Fund for Young Teachers of Chinese Central Colleges (No.2009B32044). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

7. References