Sensors & Transducers, Vol. 178, Issue 9, September 2014, pp. 188-193

www.sensorsportal.com

Sensors & Transducers

http://www.sensorsportal.com

Compressed Sensing Adaptive Speech Characteristics Research

Long Tao
Science & Research Division, Hunan International Economics University, Changsha, 410205, China
Tel.: 86-0731-88127113
E-mail: matlab_wjf@126.com

Received: 24 May 2014 /Accepted: 29 August 2014 /Published: 30 September 2014

Abstract: The sparsity of the speech signals is utilized in the DCT domain. According to the characteristics of the voice which may be separated into voiceless and voiced one, an adaptive measurement speech recovery method is proposed in this paper based on compressed sensing. First, the observed points are distributed based on the voicing energy ratio which the entire speech segment occupies. Then the speech segment is enflamed, if the frame is an unvoiced speech, the numbers of measurement can be allocated according to its zeros and energy rate. If the frame is voiced speech, the numbers of measurement can be allocated according to its energy. The experiment results shows that the performance of speech signal based on the method above is superior to utilize compress sensing directly. *Copyright © 2014 IFSA Publishing, S. L.*

Keywords: Compressed sensing, Speech recovery, Sparsity, OMP, Adaptive.

1. Introduction

Traditional signal acquisition and processing mainly includes sampling, compression, transmission, decompression four steps, the sampling process must satisfy the Nyquist sampling theorem. Compression process is first the transformed signal, and relatively large absolute value coefficients are coded for transmission, and coefficient close to zero is discarded. Most of the sampled data are abandoned in the compression process. Candes has proven theoretically the original signal is reconstructed accurately from the part of the Fourier transform coefficients，this is the compressed sensing theory foundation.

Compressed sensing (CS) [1-3] theory suggests that the required information may not be lost in the approximation of the original signal, and the signal is sampled with the least number of observations, the dimension of the signal is reduced. That signal is directly sampled less to get compressed signal expression. The cost is saved in the sampling and transmission, the purpose is achieved simultaneously in the sampling and compressed signal, it has broken the limitations of Nyquist sampling theorem. When the signal has a sparse or compressible, the reconstruction signal can be achieved accurately or approximately by collecting a small amount of the signal projection value.

Since CS is advent, it has been widely studied in many fields of the information theory and coding, signal restoring, lossy compression, machine learning, the sensor network. The image signal processing is also a wide range of applications. For example, in the image processing, since the high frequency component is very smaller than the low frequency component in the majority case, the image signal time-frequency coefficients can be regarded as sparse, the digital rate can be greatly reduced by using CS technology to capture images.

Speech signal for CS is not much research currently, the characteristics of the speech signal is researched in this paper, the observation points are assigned adaptively for each frame to reconstruct the speech signal, and the reconstructed signal CS directly is compared.

2. Compressed Sensing Theory Foundation

2.1. Signal Sparse Representation

Signal $X=\{x_1, x_2,\ldots, x_N\}^T$ belongs to a column vector of R^N space, Ψ belongs to a group of orthogonal basis of R^N space, $\Psi=\{\varphi_1, \varphi_2,\ldots, \varphi_N\}$, the signal X can be expressed as:

$$X = \sum_{i=1}^{N} \theta_i \varphi_i = \Psi\ \Theta \tag{1}$$

$\|\Theta\|_0 = K$, which means that non-zero element number, K is a small, an it represents that X signals are sparse in the orthogonal basis Ψ, we call that X is K-sparse. At this point, you can use a matrix Φ (M * N, M << N), which is irrelevant to with Ψ, and which is a linear transformation on X:

$$Y=\Phi\ X \tag{2}$$

If the observation matrix Φ satisfies RIP nature [3], the original signal X can be reconstructed approximately and lossless by the observation vector Y, which is equal to solving the matrix equation:

$$\begin{aligned} Y &= \Phi X = \Phi\Psi\Theta = T\Theta \\ s.t. &\|\Theta\|_0 = K \end{aligned} \tag{3}$$

Because M << N, there is non-unique solution in equation (3), the coefficients of signal sparse domain are reconstructed by the l_0 optimization algorithm:

$$\Theta = \min \|\Theta\|_0, s.t. Y = \Phi\Theta \tag{4}$$

Since (4) is difficult to solve, so the l_0 problem is transformed into l_1 problem:

$$\Theta = \min \|\Theta\|_1, s.t. Y = \Phi\Theta \tag{5}$$

2.2. Select the Observation Matrix

How to select the observation matrix is also an important consideration of the issue, it is needed to ensure that the observation matrix and sparse groups are irrelevant, RIP meet the criteria, which was expressed as:

$$\begin{aligned} (1-e)\|X\|_2^2 \le \|\Theta X\|_2^2 \le (1+e)\|X\|_2^2 \\ e \in (0,1) \end{aligned} \tag{6}$$

It can be understood, after signal is observed by the observed matrix, the observed value energy is not much change with of the original signal energy, and energy is the main characteristic parameters of a signal. If the signal sparsity is known, and most of the signal energy is not lost, a good recovery of the original signal is possible.

Compressed sensing is also understand availability from the physics point of view [4], for example, Young's double-slit experiment is token, lights pass two small parallel gaps, the light and dark stripes are showed on the screen. In fact, these light and dark fringes can be interpreted as the superposition of the two light signals and weakened, the superposition and weakening of the two light sources is because both light sources are coherent (they come from the same source). In the field of signal processing, if the signal is sparse, it is necessary to make that the sparse base and the signal is coherent, if the observed signal is obtained by the observation matrix without loss, you must make that the observation matrix is non-coherent with the signal.

Gaussian random matrix Φ has a useful property: For a Gaussian random $M \times N$ matrix Φ, when $M \ge c*k*\log(N/K)$ (c is a small constant), Φ Ψ has RIP nature in great probability, the K-dimensional sparse signals with length N can be restored with a high probability from the M observations.

2.3. Reconstruction Algorithm

After the compressed sensing theory was proposed, there has been a variety of sparse signal reconstruction algorithm [5]. They are mainly attributed to two categories, the greedy algorithm is proposed based on the l_0 norm, which includes matching pursuit algorithm, orthogonal matching pursuit algorithm, and so on. The convex optimization algorithm is made on another norm [6], which includes the gradient projection method, based tracking algorithms and so on.

Convex optimization algorithm is a kind of algorithm with reconstruction better, greater computational complexity, the solved model is:

$$\min \|\Theta\|_1 \quad \text{s.t} \|Y - T\Theta\| \le \varepsilon, \tag{7}$$

where ε is the control error factor. Or it is converted into unconstrained optimization problem [7]:

$$\min \|\Theta\|_1 + \lambda \|Y - T\Theta\|, \tag{8}$$

where λ is the relaxation factor, the balance is done between the control sparsity and reconstruction error.

In comparison, the greedy algorithm has a small amount of calculation, remodeling effect better and easy to implement, its solving model is:

$$\min \|\Theta\|_0 \qquad \|Y - T\Theta\| \le \varepsilon \qquad (9)$$

In engineering, we usually use OMP (Orthogonal Matching Pursuit) [8-9] algorithm. OMP algorithm follows the MP algorithm, the atoms are continuously selected with the observed signal or iterative margin last match from atomic matrix T. The difference is that atomic is selected by OMP algorithm, and it is made in orthogonal treatment, and then the signal components and iterative margin are solved. OMP algorithm iterative process:

Input: perception matrix T, the measured value Y, the number of iterations m

Output: X's K sparse solution x_n, as well as the reconstruction error r_n

Initialization: $r_0 = Y$, Atomic collections $A_{\Lambda_0} = \varnothing$, atom index number $\Lambda_0 = \varnothing$

I: to calculated the inner product for margin and T's each column, to find the column vector A_k and the column number k which the maximum inner product value is correspond to;

II: to update the index collection $\Lambda_n = \Lambda_{n-1} \cup \{k\}$, updating atomic matrix $A_{\Lambda_n} = A_{\Lambda_{n-1}} \cup \{A_k\}$;

III: to use the least squares method for the approximate solution: $x_n = (A_{\Lambda_n}^T A_{\Lambda_n})^{-1} A_{\Lambda_n} Y$;

IV: to update margin $r_n = Y - Tx_n$;

V: If the number of iterations is less than m, go to Step I, otherwise stop the iteration, while the output x_n.

3. Adaptive Compressed Sensing of Speech Signal

3.1. Characteristics of the Speech Signal [10]

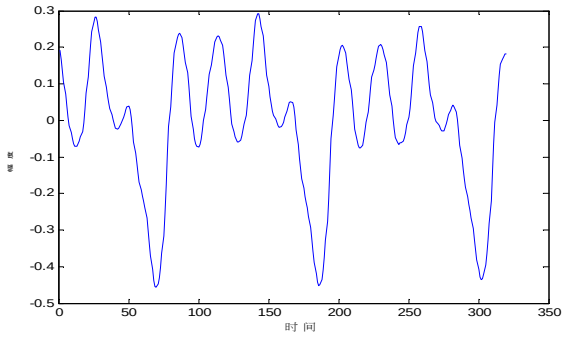
Speech signal is a time-varying, non-stationary random process. However, because human muscle movement speed is slow in speech organs, speech in 10 ~ 30 ms can be approximated as smooth. In short, the feature of the voice signal varies with time, in a short time frame, the speech signal maintains relatively stable and consistent characteristics, there is short-term stationarity in the speech signal. In the speech signal analysis and processing, the speech signal must be divided into frame to be processed.

The main component of the speech signal is voiced, unvoiced portions occupy only a small proportion. A speech is used in the experiment, the voiced signal energy accounts for about 93 % of the speech segment, and there are small amount of unvoiced segments between voiced segments, we focus on the voiced signal treatment of speech.

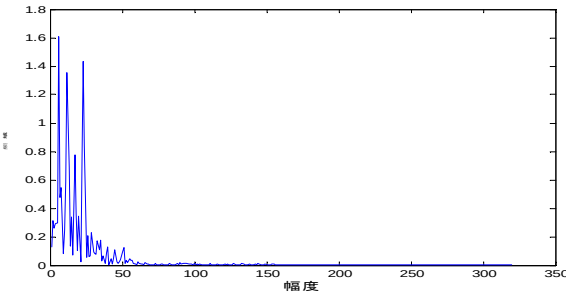
3.2. Speech Signal Sparsity

If there is compressibility in speech signals, a sparse group must be chosen legitimately, speech signal have a good sparsity in the group, and common sparse transformation are: cosine transform, Fourier transform, wavelet transform. We use the DCT transform to process a random frame voiced signal of a certain speech.

Fig. 1 shows that only a small portion of the value is relatively large in the DCT coefficients of the voiced, most of the coefficient values are small and they can be ignored. This indicates that the speech signals are approximately sparse in DCT domain, so the CS reconstruction of speech is reasonable and feasible in DCT domain.



(a)



(b)

Fig. 1. Voiced and its DCT transform coefficients, (a) frame voiced, (b) DCT transform coefficients.

3.3. Observation Points Allocation Principles

Speech can be divided into voiceless and voiced according to whether the vocal cords are vibrating, the information is contained in voiceless or voiced, and their the importance is not the same, voiced energy accounts for more than 90 % of the total

energy, and it is the more important. Each frame speech contains information, their importance also is not the same. Therefore, because of the different types and the importance of each frame speech, the observation points M are distributed adaptively, the more observation points are assigned to the more important frame, whereas fewer observation points are allocated to achieve optimal the purpose of distribution. And there is an important parameter to measure the speech, which are the energy, the zero rate, zero energy ratio. We recorded some Girls 'molecular structure' voice, the sampling frequency is 16 kHz, the frame length N = 320 and the speech parameters are solved.

It can be seen from Fig. 2 (a) and (b) that the voiced energy is far greater than the voiceless energy, and voiced zero rate is very small. So based on energy and short-term zero-crossing rate, we can judge whether one frame speech is voiced or voiceless. As can be seen from the short-term energy, energy changes between each voiced frame are more obvious, so you can distinguish its importance between the voiced frames by the energy [11-12].

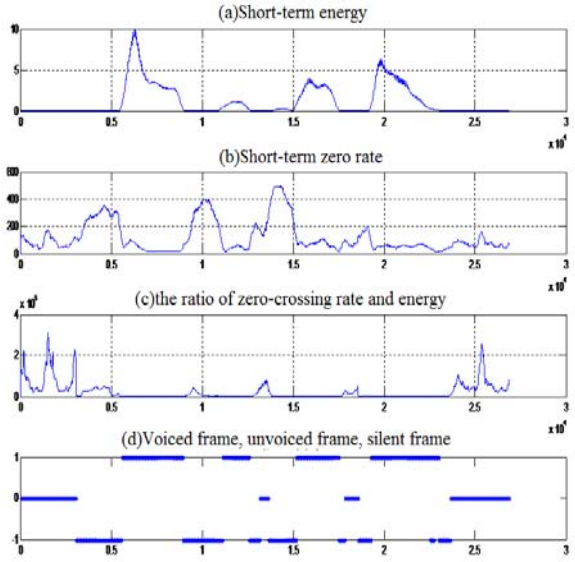


Fig. 2. Speech segment each parameter value (In Figure (d) , value 1 represents a voiced frame, value -1 indicates unvoiced frames, value 0 indicates silence frames).

It can be seen from (c) that the ratio change of zero-crossing rate and energy between the unvoiced frames is too obvious, so you can be able to distinguish their importance between each voiceless frame by the ratio of zero-crossing rate and energy (Because the ratio of zero-crossing rate and energy is the greater, that indicates that the classified unvoiced likelihood is the greater, this frame is not the more important, so we use the reciprocal of zero-crossing rate and energy as basis, which the observation points can be allocated).

We must first distinguish their nature according to the amount of a frame speech energy, and their different observation points are distributed by depending on the ratio of the total energy of each Voicing, silent segment (energy is almost zero, and it does not contain any semantic) is directly set to 0. Then according to one frame voiced energy occupies the proportion in the total voiced energy, the observation points are distributed, according to an unvoiced frame zero-crossing rate can possess the proportion in the overall frame voiceless one. The observation points are assigned for the frame. Due to the large gap between the voiced frame energy, there is the larger difference between the assigned points, in order to achieve a balanced distribution of points in each frame, we take the logarithm treatment for each frame energy, so that the energy difference is reduced between each frame. Similarly, the same approach can be also used for the ratio of voiceless energy and zero-crossing rate. If the distribution of observation points are over 320 points, we continue to increase M, an there will not be have any help to improve the signal to noise ratio, it will lead to decreased compression ratio, so the observation points are capped at 320 points. If the distribution of points is less than 10, we assign 10 observation points, non-critical frame reconstruction quality is ensured. Detailed expressed is as follows:

1) If frame energy is less than T1, frame will be classified as a silent, do not assign observation points, zeros are set directly in the corresponding position of the recovery end.

2) Then according to the ratio of the remaining frame energy and short-term zero zero-crossing rate, the remaining frames will be divided into voiceless frames and voiced frames.

3) Let the total observation points for M, each frame energy as $E_n = \sum_{m=1}^{N} x_n^2(m)$, the total voiceless energy is Eq in speech segments, the total voiced energy is Ez, the assigned dullness observation point number is M1 = M * Ez / (Eq + Ez), the voiceless observation point number is M2 = M-M1.

4) If it is the voiced, to calculate the energy e, then the observing distribution point number is m = e / Ez * M1.

5) If it is voiceless, to calculate the total ratio of the voiceless energy and zero-crossing rate for EZR (This paragraph can explain the greater voiceless energy and zero-crossing rate, the larger the voiceless probability, the distribution points should be less), and then to calculate ear which is the ratio of a frame energy and zero-crossing rate. The frame allocated observing point umber is m = ezr / EZR * M2.

6) Observation point M = min (320, m). For each frame speech is reconstructed by using OMP algorithm, and the full speech is recovered after the reconstruction.

The algorithm flow chart is shown in Fig. 3.

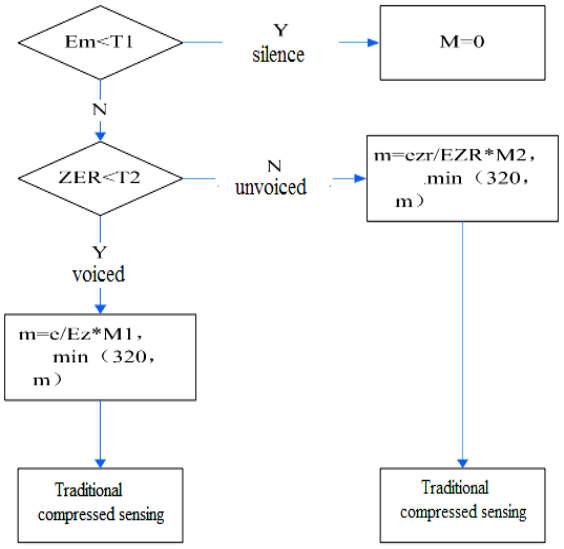


Fig. 3. Algorithm flowchart.

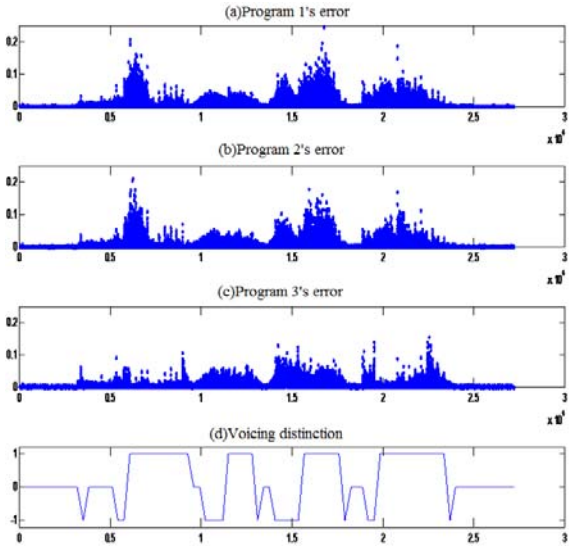


Fig. 4. The error comparison of three programs (In figure (d), value 1 represents a voiced frame, value -1 indicates unvoiced frames, value 0 indicates silence frames).

4. Experimental Evaluation

4.1. Error Analysis

In this study, there are 27,200 sampling points, the sampling frequency is 16 kHz, it is female voice, semantics is "molecular structure." Rectangular window is used in experiment framing, non-overlapping frames, each frame size is 320 points / frame, threshold parameter T1 = 0.001, T2 = 150. These parameters are obtained after several tests, which are the experience values.

In order to demonstrate the advantages of the algorithm, we will compare it with the following two options:

Scenario 1: we do not distinguish between voiceless and voiced， the speech signal is directly reconstructed with OMP.

Scenario 2: we distinguish voicings and muted tones, a fixed length M1of observation points are assigned for each voiced, the fixed-length observation points are allocated for each voiceless, it is M2, M1> M2. Scenario 3 is the method which is described in section 3.3.

As can be seen from Fig. 4, the maximum amplitude of Scheme 3 is smaller than the other two schemes, or even only one half of the maximum amplitude in Scheme 1.

Especially in the more important voiced frames, performance is better, reconstruction can be achieved more accurately. In terms of the unvoiced frame, it is obvious improvement, after the allocation of the adaptive observation points, unvoiced share points are less than the first two schemes, since the unvoiced is similar to white noise, if signal is reconstructed with a larger number of observed points, it is not much effect to improve the signal to noise ratio, we can assign these observations point to a more important voice frame, this is the advantage which adaptive allocation lies.

4.2. SNR Test

In this study, there are the same speech with 4.1 and pretreatment, while the compression ratio is defined as r=M/N.

$$SNR = 10\log_{10}(\|X\|^2 / \|X - \hat{X}\|^2) \cdot$$

The three schemes are compared, the three scenario relationship between SNR and r is shown in Fig. 5.

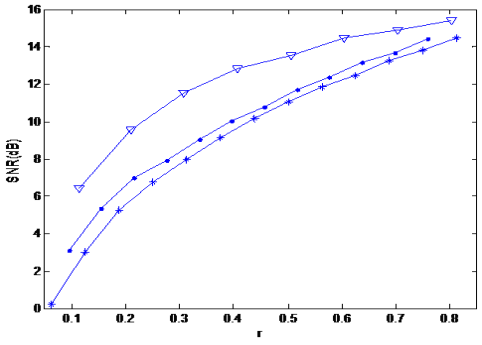


Fig. 5. SNR comparison of three algorithms.

As can be seen from Fig. 5, after we distinguish voiced and unvoiced, their SNR is slightly better than one in the original OMP, the SNR of the adaptive observation points will have significantly better than the first two algorithms, in particular, when the compression ratio r is small, effect is significantly better, it can achieve a greater ratio of signal to noise.

In experiments above, adaptive algorithm is only valid for the characteristics speech, in order to verify

its universality, in male and female voices, each 100 speeches, and three speeches were randomly selected to test the three methods, the compression ratio r = 0.4, the experimental results are in Table 1.

Table 1. Different speech SNR under three methods.

(a) Girls test group

r=0.4	Method 1	Method 2	Method 3
Female 1	7.4	8.2	8.35
Female 2	6.04	7.42	7.78
Female 3	7.13	8.33	8.89

(b) Male test group

r=0.4	Method 1	Method 2	Method 3
Male 1	17.63	18.50	20.15
Male 2	11.60	12.95	18.67
Male 3	18.33	18.04	22.26

As can be seen from Table 1, there is a corresponding increase in the adaptive method which is are compared with method 1 and method 2, and more excellent performance is showed in terms of the male, the signal to noise ratio has been greatly improved for each test speech. This illustrates that there is the general applicability to any speech in the method. We noticed that with the same treatment method, male SNR is larger than female SNR generally, which is because the female higher frequency component is more.

5. Conclusions and Outlook

This paper describes the theoretical framework of compressed sensing and sparse nature of the speech signal, the compressed sensing theory is used into speech signal processing. In the DCT-based, the OMP algorithm is used for speech signal processing, and on this basis, by the characteristics of the speech signal, they is divided into the voice signal Voicing. According to "voiced energy is large, the more important; voiceless energy is small, less important" in nature, "divide and rule" purposes are reached. Classifying the speech signal through the different features: the voiced is determined by energy, the unvoiced is determined by the ratio of energy and zero-crossing rate, the different observation points are assigned adaptively. Through experiments, it is further evidence that the performance of adaptive allocation method for observation points is superior to the OMP speech signal processing directly, but also it is superior to distinguish voicing, which is no adaptive allocation scheme for observation points.

Acknowledgements

This paper is sponsored by the Scientific Research Project (NO.12C0798) of Department of Education of Hunan Province, and Hunan Science and Technology Program "object-oriented IOT rapid discovery, location and access information technology research (NO: 2014FJ3040)".

References

[1]. Donoho D., Compressed sensing, *IEEE Transactions on Information Theory*, Vol. 52, Issue 4, 2006, pp. 1289-1306.
[2]. E. Candès, Compressive sampling, in *Proceedings of the International Congress* of *Mathematicians*, 2006.
[3]. Candess E., Romberg J., and Tao T., Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information, *IEEE Transactions on Information Theory*, Vol. 52, Issue 2, 2006, pp. 489-509.
[4]. Qi Dai, Wei Sha, The Physics of compressive sensing and the Gradient-Based Recovery Algorithms, *Research Report*, http://arxiv.org/abs/0906.1487.
[5]. Jianjing, Yuantao Gu, Shunliang Mei, An introduce to Compressive sampling and its application, *Journal of Electronics & Information Technology*, 2010
[6]. Tsaig Y. and Donoho D., Extensions of compressed sensing, *Signal Processing*, Vol. 86, Issue 3, 2006, pp. 533-548.
[7]. Can Zhou, Research on signal reconstruction algorithms based on compressed sensing, *Beijing Jiaotong University*, 2010.
[8]. J. Tropp and A. Gilbert, Signal recovery from random measurements via orthogonal matching pursuit, *IEEE Transactions on Information Theory*, Vol. 53, 2007, pp. 4655–4666.
[9]. D. Needell and R. Vershynin, Signal recovery from incomplete and inaccurate measurements via regularized orthogonal matching pursuit, *IEEE Journal of Selected Topics in Signal Processing,* Vol. 4, Issue 2, 2010, pp. 310-316.
[10]. Jiqing Han, Lei Zhang, Tieran Zhen, Speech signal processing, *Beijing Tsinghua University Publishing House*, Vol. 9, 2004, pp. 43-55.
[11]. Haiyan Guo, Tianjing Wang, Zhen Yang, Adaptive speech compressed sensing in the DCT domain, *Chinese Journal of Scientific Instrument*, Vol. 31 Issue 6, 2010.
[12]. Haiyan Guo, Zhen Yang, Compressed speech signal sensing based on approximate KLT, *Journal of Electronics & Information Technology*, Vol. 23, Issue 31, 2009.