Compressive Sensing in Speech Processing: A Survey Based on Sparsity and Sensing Matrix

Siddhi Desai¹, Prof. Naitik Nakrani²

¹Student of M.tech Electronics and Communication Engineering in Chhotubhai Gopalbhai Patel Institute of Technology, Bardoli
²Assistant Professor, Department of Electronics and Communication Engineering, Chhotubhai Gopalbhai Patel Institute of Technology, Bardoli

Abstract—Compressive sampling is an emerging technique that promises to effectively recover a sparse signal from far fewer measurements than its dimension. The compressive sampling theory assures almost an exact recovery of a sparse signal if the signal is sensed randomly where the number of the measurements taken is proportional to the sparsity level and a log factor of the signal dimension. Encouraged by this emerging technique, this paper briefly reviews the application of Compressive sampling in speech processing. It comprises the basic study of two necessary condition of compressive sensing theory: sparsity and incoherence. In this paper, various sparsity domain and sensing matrix for speech signal and different pairs that satisfy incoherence condition has been compiled.

Keywords—Compressive sensing, Incoherence, Sensing Matrix, Sparsity Domain

I. INTRODUCTION

Compressive sensing or C.S. is a very simple, efficient, non-adaptive and parallelizable compressed data acquisition protocol that provides both sampling and compression along with encryption of source information simultaneously [1],[3]. The theory of compressive sensing was developed by Candès et al and Donoho in 2004 [8]. This method is different from traditional method as it sampled the signal below the Nyquist rate and it permits to exploit the sparse property at the signal acquisition stage of compression [3]. In this method the signal is first transformed into a sparse domain and then the signal is reconstructed using numerical optimization technique using small number of linear measurements [8]. Implementation of Compressive sensing Theory in specific application reduced sampling rates, or reduced use of Analog to Digital converter resources [1]. Compressive sensing is a new paradigm of acquiring signals, fundamentally different from uniform rate digitization followed by compression, often used for transmission or storage [1],[2],[3]. Compressive sensing can be used in image compression, radar system, A to D converter, Medical Imaging, speech compression [8].

This paper is organized as follows. This section gives an introduction about compressive sensing. In section II, a review about compressive sensing theory is presented. In section III, literature survey about Sparsity and sensing matrix and incoherence is given. The conclusion is given in section VI.

II. COMPRESSIVE SENSING

The basic principle of Compressive Sensing is shown in Fig.1. It consist two main parts: transmitter and receiver. Transmitter side input signal x is given with N samples. First x has to be converted into some domain in which x has sparse representation. For example, DCT, DFT etc. after this conversion signal x is transformed into K – sparse signal. Where K is largest coefficients obtained using thresholding. These K largest coefficients contain most of the information about signal. Then it is multiplied with sensing matrix φ and result will give M – length measurement matrix.

At the receiver side, different optimization techniques are used for reconstruction of original signal. First multiplication of signal with sensing matrix is computed which gives N samples from M measurements. Then convex optimization techniques are used to recover K-sparse signal. Once again inverse sparsity is applied to obtain original signal.

![Block Diagram](image-url)
In short, working of compressive sensing theory is mathematically expressed by following manner:

Let \( x \in \mathbb{R}^N \) be the speech signal and let \( \psi = [\psi_1, \psi_2, \ldots, \psi_N] \) be the basis vectors spanning \( \mathbb{R}^N \).

The speech signal is said to be sparse if,

\[
x = \psi \cdot s = \sum_{i=1}^{k} s_{ni} \cdot \Psi_{ni}, [n_1, n_2, \ldots, n_k] \subseteq [1, \ldots, N]
\]

Where, \( S_{ni} \) are scalar coefficients and \( K \ll N \), i.e. \( s_n \) or simply \( s \) is the sparse vector with only \( K \) non-zero elements.

Based on CS theory, perform sampling of \( x \) through projections onto random bases and reconstruct the speech signal at the receiver with full knowledge of the random bases.

In other words, the sampling (sensing) measurements can be defined as:

\[
y_m = \sum_{i=1}^{N} \phi_m(i) \cdot x(i), \quad 1 \leq m \leq M < N
\]

Or \( y = \phi \cdot x \), where \( \phi = M \times N \) is measurement matrix. The \( \phi \) is made up of orthonormal random basis vector \( \phi_m \). If the incoherence condition between \( \phi \) and \( \psi \) are satisfied, then there is a high probability that \( y \) can be reconstructed perfectly if \( M > K \log N \) measurements.

At Receiver side, for reconstruction of signal, convex optimization techniques are used.

Convex optimization then can be utilized as follows:

\[
\min \| s \|_p \quad \text{such that} \quad y = \psi \cdot \phi \cdot s = A \cdot s
\]

Where, A=reconstruction matrix and \( \| s \|_p \) is the p-norm given by

\[
\left( \sum_{n=1}^{N} (S_n)^p \right)^{\frac{1}{p}}
\]

A very interesting property of CS is that if signal is not K–sparse (or, not exactly K-sparse), the quality of the recovered signal \( s \) (or, equivalently \( x \)) is as good as to select only the K largest values before the calculations, and measure them directly.

In mathematics the above problem is called underdetermined problem due to more number of coefficients than variables to solve & is a linear programming problem for \( p=1 \). Various algorithms have been developed to solve this problem.

The best solution is achieved by \( L_0 \)-norm i.e. \( p=0 \) in equation (4), but it is NP hard problem. \( L_2 \)-norm i.e. \( p=2 \) in equation (4) gives not sparse solution. \( L_1 \)-norm i.e. \( p=1 \) in equation (4) gives sparse solution & good reconstruction probabilities.

III. LITERATURE SURVEY

Compressive Sensing mainly relies on two things: Sparsity –which relates to the signals of interest and Incoherence–which relates to the sensing modality.

A. Sparsity:

Signals that are mostly populated with zeros and have a small number of non-zero components are called sparse signals. A sparse representation is one in which small number of coefficients contain large proportion of energy.

Fig. 2 shows that the Original sampled signal is mostly populated with non-zero samples. And in Fig. 3 its Fourier transform is given that contain large proportion of zero samples. This is known as sparse representation.
Different Sparsity basis for speech signal: [3] [7] [10]

1. Discrete Fourier transform:
   It is the most basic sparsifying basis. But due to following reasons it is not suitable for speech: DFT is more suitable for audio than speech because audio tones have very less frequency components compare to speech. It also gives the complex numbers as result. Due to the large amplitudes values of most of the components we cannot discard spectrum component to increase sparsity as shown in Fig.4 [10].

![Fig. 4: Sparsity using DFT](image)

2. Discrete Wavelet Transform:
   DWT is better than the DFT. A wavelet exists only within a finite domain, and is zero elsewhere, in contrast to Fourier transforms. A waveform that is bounded by both frequency and duration (time) is called wavelet. But then also it cannot be used for speech because DWT can’t achieve good efficiency at near-transparent compression ratios as compare to DCT and its Computational complexity limits the algorithmic implementation of a codec [10].

![Fig. 5: Sparsity using Bessel function](image)

3. Cepstrum domain:
   Cepstrum domain analysis used to get formant and periodicity information of speech. Equation (5) is Cepstrum domain representation of signal \( x \).

\[
x = F^{-1} \cdot \exp\{F \cdot \theta_3\} \quad \text{Where} \quad x = \psi \cdot s \quad (5)
\]

Here, \( \theta_3 \) = Homomorphic mapping of Cepstrum domain.

It cannot be used for speech because its Cepstrum relation to original speech signal non-linear so it is not suitable for linear reconstruction techniques used in Compressive Sensing [3].

4. Bessel Basis:
   It is constructed using the Bessel function of first kind which resembles damped sinusoids in case of speech signals which can be used as basis function for expanding speech signal. Bessel Basis gives real valued coefficient. But designing of such basis becomes computational difficult & time consuming as dimension of signal \( N \) increases [10].

![Fig. 6: Sparsity using DCT](image)

5. Discrete Cosine Transform:
   DCT remove redundancy between neighboring values which leads to uncorrelated transform coefficients which can be encoded independently. It is better for concentrating the energy into lower order coefficients. All the spectral coefficients are purely real in it [10].

Fig.6 shows the DCT of speech signal. It has very few numbers of non-zero spectrum components. Thus we can even increase sparsity of speech by thresholding low value components without considerably affecting speech quality [10].

6. Spectral domain:
   Spectral domain representation is linear sparse representation of signal in time domain and it gives periodicity and formant parameters of speech in form of residual excitation component. Equation (6) is spectral domain representation of signal \( x \).

\[
x = h \cdot r 
\]

Here, \( h = N \times N \) Impulse response matrix
\( r = N \times 1 \) Residual excitation vector

Signal dependent impulse response matrix \( h \) is used \( \psi \) and it is Toeplitz lower triangular and circular Toeplitz.

The Auto Regressive (AR) parametric approximation results the IIR response so it is truncated to FIR to get \( h \); the truncated length is \( K > N \).This makes size of \( h = N \times K \). So, \( h \) can be chosen to be an \( N \times N \) circulant or \( N \times K \) Toeplitz of impulse response matrix [3].
7. Hybrid dictionary:

Sparsity of speech signal is also obtained by ‘Hybrid dictionary’. It is made up by combining LPC Model (For Unvoiced) and DFT model (For voiced) as basis of speech signal. A speech signal is quasi-periodic in its voiced parts; hence a Discrete Fourier Transform (DFT) basis can provide a better approximation and the linear prediction coding (LPC) is an efficient tool for speech compression, as the speech is considered to be an AR process. So, hybrid dictionary combined with LPC and DFT model is proposed as the basis for speech signal [7].

8. Linear Prediction Coefficients:

LPC is efficient technique for speech processing. LP residues are sparse and give a large amount of information about speech like pitch; formants [10]. Equation (7) gives representation of \( x \) in terms of LP residual vector.

\[
x = A \cdot r
\]  
(7)

Where, \( A \) is the matrix that performs the whitening of the signal, constructed from the coefficients of the predictor a of order P.

Equivalently, it can be written as,

\[
x = A^{-1} \cdot r = H \cdot r
\]  
(8)

Where, \( r \) is residue vector and \( H \) is the inverse matrix of \( A \) and it is commonly referred to as the synthesis matrix that maps the residual representation to the original speech domain. And \( y = H \cdot r \).

\( H \) is constructed directly from the impulse response \( h \) of the all pole filter that corresponds to \( a \) and \( r \) is composed of \( N+P \) rows with the first \( P \) elements belong to the excitation previous speech frame [4].

Analysis done after study of various sparsity domains:

- After evaluating DCT, DWT, and DFT for speech signals using gini index, DCT achieved higher gini index compared to DWT and DFT, so, DCT representation of speech is sparser than DWT, DFT [9].

After evaluating DWT (DB1, DB2 Haar), DFT, LP coefficients and DCT by L1 norm. Among them, the DCT and LP coefficients give more sparse representation so they are best choices as sparsity domain [10].

B. Sensing Matrix:

Two types of measurements matrices that can be used in compressive sensing are given below:

1. The Random measurement matrix:

In this rows are randomized from the same random seed vector. The random matrix is transposed and then orthogonalized. So, this will create the matrix that represents an orthonormal basis.

Examples of Random Sensing Matrix: [2]

1) form \( \phi \) by sampling \( n \) column vectors uniformly at random on the unit sphere of \( R^m \);

2) By sampling independent Identical Distribution entries from the normal distribution with mean \( 0 \) and variance \( 1/m \);

3) By sampling a random projection \( P \) as in “Incoherent Sampling” and normalize: \( A = \sqrt{n/m}P \);

4) By sampling independent identical distribution entries from a symmetric Bernoulli distribution (\( P(Ai, j= \pm 1/\sqrt{m}) = 1/2 \)) or other sub-Gaussian distribution.

2. The predefined measurement matrix:

The matrix is created by using function like the Dirac functions and Sine functions [8].

Several examples of deterministic matrix:

1) Fourier Matrix:

It has randomly selected \( M \) rows out of the \( N \times N \) Fourier matrix. Computation complexity can be reduced by using the fast Fourier transform algorithms. It is used for time domain only [10].

2) Wavelets:

They can be used as sparse basis and sensing matrices because of its inherent property of orthogonality among rows & columns. A combination of wavelets with DCT can give very good reconstruction results in less time [10].

3) Orthogonal Symmetric Toeplitz Matrices (OSTM):

They are easy to generate as only \( N \) numbers need to be stored. Both sampling and reconstruction are more efficient to implement. They are well suited to some applications such as system identification, channel estimation etc [10].
Results derived after study of various sensing matrices:

- Reconstructed speech is very good in case of random Gaussian sensing matrix as compared to Haar, Hadamard matrices. But the problem with Random matrix is that it takes large reconstruction & sensing time and doesn’t guarantee similar results. So, deterministic matrices like wavelets can be used which give comparable & guaranteed results i.e. the quality of reconstructed speech is same each time processing is done \[^{10}\].
- Bernoulli matrices have reduced storage & complexity as compared to random matrices \[^{10}\].
- Toeplitz matrix is proposed using Golay sequences. These are easy to generate & especially designed for speech signals. But they are based on statistical RIP & thus reconstruction results are only comparable to random matrices and its works for lower values of N to achieve optimal coherence values \[^{10}\].

C. Incoherence:

The coherence between $\phi$ and $\psi$ is,

$$\mu(\phi, \psi) = \sqrt{n} \cdot \max_{k,j} \left| \left( \phi_k, \psi_j \right) \right|$$

(9)

The coherence measures the correlation between $\phi$ and $\psi$. If $\phi$ and $\psi$ contains correlated elements, the coherence is large. Otherwise, it is small. It follows from linear algebra that $\mu(\phi, \psi) \in [1, \sqrt{n}]$.

Compressive sampling theory is mainly concerned with low coherence pairs \[^{2}\].

Examples of pairs that satisfy Incoherence condition:

1. If $\psi$ is the Fourier basis then $\phi$ can be Canonical or Spike basis.
2. Time–frequency pair gives maximum incoherence.
3. If $\psi$ is wavelet basis then $\phi$ can be noiselet.
   - The coherence between noiselets and Haar wavelets is $\sqrt{2}$.
   - Noiselets and Daubechies D4 and D8 wavelets are, respectively, about 2.2 and 2.9.
   - Noiselets are also maximally incoherent with spikes and incoherent with the Fourier basis.
4. Independent Identical Distribution Gaussian or Gaussian ±1 binary entries have a very low coherence with any fixed representation.
5. Spikes and sinusoids are maximally incoherent \[^{2}\].

D. Restricted Isometry Property:

The Restricted Isometry Property (RIP) is a sufficient condition on CS matrices which can ensure the performance of signal recovery \[^{10}\]. The RIP is given as

$$1 - \varepsilon \leq \frac{\|\psi \cdot s\|_2}{\|s\|_2} \leq 1 + \varepsilon$$

(10)

Here, $\varepsilon$ is RIP constant.

For perfect reconstruction $\phi$ and $\psi$ satisfy the RIP and Incoherence condition \[^{8}\].

E. Four -to-one practical rule:

The four-to-one rule says that for exact recovery, one needs about four incoherent samples per unknown non-zero term \[^{2}\].

IV. CONCLUSION

Compressive sensing theory can be efficiently used in speech processing applications. Due to the Sensing matrix and Sparsity domain conversion of signal in compressive sensing, sampling, compression and encryption is obtained. Sensing matrix and sparsity domain should be selected in such a way that satisfy incoherence and RIP condition. There are various domains for speech as discussed above among them DCT, Spectral domain, Hybrid dictionary, LPC are better. As sensing matrix various choices like Random Gaussian matrix, Bernoulli and different deterministic matrices can be used.

REFERENCES


