Abstract

Compressive Sensing (CS) signal recovery has been formulated for signals sparse in a known linear transform domain. We consider the scenario in which the transformation is unknown and the goal is to estimate the transform as well as the sparse signal from just the CS measurements. Specifically, we consider the speech signal as the output of a time-varying AR process, as in the linear system model of speech production, with the excitation being sparse. We propose an iterative algorithm to estimate both the system impulse response and the excitation signal from the CS measurements. We show that the proposed algorithm, in conjunction with a modified iterative hard thresholding, is able to estimate the signal adaptive transform accurately, leading to much higher quality signal reconstruction than the codebook based matching pursuit approach. The estimated time-varying transform is better than a 256 size codebook estimated from original speech. Thus, we are able to get near “toll quality” speech reconstruction from sub-Nyquist rate CS measurements.

Index Terms: Compressive Sensing, Iterative hard thresholding, Signal adaptive transform.

1. Introduction

Compressive Sensing theory developed in [1, 2] permits sampling of a signal at its intrinsic information rate which could be much lower than Nyquist rate, while guaranteeing good quality reconstruction for signals sparse in a linear transform domain. CS framework provides sensing in a compressed and encoded fashion through incoherent projections onto a lower dimensional subspace. The signal could then be recovered almost perfectly if the signal is sparse, which is the case for many practical signals; for example, images are sparse in DWT domain. Though the sensing mechanism is simple, the decoder complexity is high and involves complex iterative optimization.

Sparse representations of signals have been studied using signal independent orthogonal transforms such as DCT, DWT, or DFT, and signal dependent transforms such as learnt dictionaries. Compressive sensing framework has been explored with respect to both these approaches when the dictionary is known priori. Speech signals are highly time-varying and non-stationary in nature. The information content in speech is spread over a large dynamic range of 30-40 dB and hence difficult to characterize sparsity. Also, for speech, there is no global orthogonal transform which can represent all the speech frames sparsely. Hence, we have to consider signal adaptive transforms such as vocal tract impulse response in the case of speech. Application of CS framework to speech signals has been studied in the context of sinusoidal modeling in [3] and linear predictive model in [4, 5]. The approach in [4] uses a code book of vocal tract impulse response which is the sparsifying transform that results in reconstruction from CS. In [5], the authors assume a scenario in which the impulse response is specified.

In this paper, we focus on iterative estimation of signal adaptive transform i.e., the vocal tract impulse response matrix and the sparse excitation signal jointly from just the CS measurements. We examine the CS for speech signals using the linear prediction (LP) framework; unlike the approaches in [4, 5], we estimate the impulse response matrix from the CS measurements. In particular, we show that using an iterative procedure in combination with a modified iterative hard thresholding (IHT) algorithm, it is possible to recover both the vocal tract impulse response matrix and the sparse excitation signal. We also show that the estimated LP parameters are close to the actual LP parameters obtained using the original speech, which makes them suitable for further processing such as verification and recognition tasks.

2. Compressive Sensing fundamentals

Let $x$ be a signal in the $N$-dimensional Hilbert space. Let $s$ be the representation of $x$ in the linear transform domain, i.e., $x = \Psi s$. When $\Psi$ is orthonormal, $x$ and $s$ are said to be equivalent representations. $x$ is said to be $K$-sparse in the $\Psi$ domain, if $s$ contains only $K \ll N$ number of non-zero elements. In CS formulation, the transformation need not be orthogonal, and the signal can be represented sparsely in any coherent dictionary [6].

The measurement process in compressive sensing consists of projecting the signal vector, onto a set of sensing waveforms. Let $\phi_{j=1..M}$ be $M$ sensing waveforms, the measurement $y_j$ is computed as $y_j = <\phi_j, x>$. By stacking the measurements into a vector $y$, we get

$$y = \Phi x = \Phi \Psi s$$

(1)

The issues associated with CS are broadly categorized as (i) design of the sensing matrix ($\Phi$), and (ii) the reconstruction algorithm. The measurement matrix ($\Phi$) should satisfy the “Restricted Isometry Property (RIP)” and “Incoherence” with any fixed basis as discussed in [1, 2]. It has been shown that random matrices with the above two properties permit the CS reconstruction with high probability when the number of CS measurements is, $M \geq cK \log (N/K)$. The reconstruction is formulated as the sparse inverse problem using $l_0$-norm minimization which is $NP$-hard computationally; but, $l_1$-minimization is shown to be equivalent to $l_0$-minimization with high probability [1]. The reconstruction using $l_1$ is formulated as,

$$\min ||s||_1 \text{ subject to } y = \Phi \Psi s.$$

(2)
This can be solved using “linear programming” techniques, and is termed as basis pursuit [7]. However, greedy algorithms such as “matching pursuit (MP)” [8], which are more effective computationally, result in a sub-optimal solution when the sparsity is assumed to be known. MP solves the following quadratic optimization problem with a sparsity constraint,

\[
\min \|y - \Phi \Psi s\|_2 \text{ s.t. } \|s\|_0 = K
\]

(3)

Recently, hard thresholding algorithms [9] have also gained importance because of the simplicity and the promising reconstruction guarantees.

3. Problem formulation

Let us consider the case in which the sparsifying transform \(\Psi\) is not known and only the CS measurements are available. The reconstruction goal in such a case is to estimate both the transform and the sparse signal, which can be formulated as:

\[
[\hat{\Psi} \hat{s}] = \arg \min_{\Psi, s} \|y - \Phi \Psi s\|_2 \text{ s.t. } \|s\|_0 = K
\]

(4)

\[\text{Sparse signal Reconstruction: Eqn. (3)}\]

\[\text{Init: } i = 0\]

\[\Psi^0\]

\[\text{Estimation of signal-adaptive transformation } \Psi\]

\[\hat{x}\]

\[\text{}\]

\[\text{}\]

\[\text{}\]

\[\text{\Figure 1: Block diagram representation of proposed method.}\]

Figure 1 gives an overview of the proposed algorithm to solve the cost function in (4). The transformation matrix is initialized to a random matrix, and the sparse signal recovery block in the diagram solves for the optimization problem in (3) assuming the transformation matrix is known. The signal \(\hat{x}\) thus estimated is used to estimate the sparsifying transform. The estimated transform is again used for sparse reconstruction in an iterative manner. The algorithm is repeated until convergence, and the estimate of the signal and transform at convergence gives the actual transform and the signal. The formulation presented here is applied to time-varying speech signals, as shown in next section.

3.1. The Speech Problem:

Let \(x[n]\) be the speech signal, and using the source-excitation model of speech production, we can write \(x[n] = h[n] \circ r[n]\); where \(h[n]\) is the vocal tract impulse response and \(r[n]\) is the excitation signal. For a signal frame of length \(N\), We can express the linear model as:

\[
x = Hr
\]

(5)

Where \(H\) is the convolution matrix formed using the impulse response function \(h[n]\). Since, the CS measurements are obtained by projecting \(x[n]\) onto a random basis \(\Phi\), we can write:

\[
y = \Phi x = \Phi H r, \text{ } ||r||_0 = K
\]

(6)

The reconstruction goal is to estimate both \(H\) and \(r\). The algorithm to solve for both \(H\) and \(r\) is summarized in Table 1.

<table>
<thead>
<tr>
<th>Table 1: Algorithm to estimate both LP parameters and the speech signal from CS measurements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initialization: (i = 0)</td>
</tr>
<tr>
<td>(\hat{h}^0) = random (N)-dimensional vector</td>
</tr>
<tr>
<td>(H^0 = \text{Conv}(\hat{h}^0))</td>
</tr>
<tr>
<td>Iterations:</td>
</tr>
<tr>
<td>Step1: Solve the sparse inverse problem, (r^i = \arg \min |y - \Phi H^i r|_2 \text{ s.t. }</td>
</tr>
<tr>
<td>(\hat{x}^i = H^i r^i)</td>
</tr>
<tr>
<td>Step2: Estimate the signal adaptive transform (H):</td>
</tr>
<tr>
<td>(a): Estimate the vocal tract impulse response (h^{i+1}) given the signal estimate (\hat{x}^i)</td>
</tr>
<tr>
<td>(b): (H^{i+1} = \text{Conv}(h^{i+1}))</td>
</tr>
<tr>
<td>Step3: Increment (i): (i \leftarrow i + 1)</td>
</tr>
<tr>
<td>goto step 1 and repeat until convergence</td>
</tr>
<tr>
<td>Note: (\text{Conv}(\cdot)) denotes the circular toeplitz matrix formed using the vector (\hat{h})</td>
</tr>
</tbody>
</table>

The sparse inverse problem in step 1 of the algorithm can be solved using any of the sparse reconstruction algorithms such as BP, MP, OMP, or CoSaMP etc. Here, we propose a modification of Iterative Hard Thresholding (IHT) [9] technique to solve the sparse inverse problem. Later in section 4, we show that the proposed algorithm is more effective compared to Matching Pursuit for the problem at hand. The modified algorithm which we term as “Iterative Pseudo inverse based Hard Thresholding” (IPHT) is summarized below,

- Goal: Solve \(\min \|y - \Phi \Psi s\|_2 \text{ s.t. } ||s||_0 = K\)
- Initialize \(\hat{s}^0 = 0, \hat{z}^0 = y, n = 0\)
- Iterate with \(n \leftarrow n + 1\), until stopping criterion is met,

\[
* \ g^n = (\Phi \Psi)^T z^{n-1}
* \ \hat{s}^n = H_K[\hat{s}^{n-1} + g^n]
* \ \hat{z}^n = y - \Phi \hat{s}^n
\]

- Output: Sparse coefficient vector \(\hat{s}\) and the signal estimate \(\hat{x} = \Phi \hat{s}\)

Here \(H_K(\cdot)\) is a hard thresholding operator that retains the \(K\) largest magnitude coefficients and the remaining elements are zeroed. The algorithm at iteration-\(n\) minimizes the cost function \(||z^{n-1} - \Phi \Psi g^n||_2^2\) whose solution is obtained as \(g^n = (\Phi \Psi)^T z^{n-1}\). This is added to the solution at the previous iteration followed by the hard thresholding operation to retain the \(K\) largest elements. The IHT/IPHT algorithm is guaranteed to converge [9] with measurement space SNR monotonically decreasing.

Step 2 in the algorithm corresponds to estimation of the signal adaptive transform. The transformation matrix here is the vocal tract impulse response matrix, which can be estimated using (i) linear prediction analysis, or (ii) cepstrum analysis of the signal \(\hat{x}^i\). In linear prediction analysis, speech is modeled as the output of an all-pole system,

\[
x[n] = \sum_{k=1}^{p} a_k x[n - k] + u[n]
\]

(7)

The system parameters \(\{a_k\}_{k=1}^p\) are estimated from speech signal \(\hat{x}^i\) using the least squares criterion. The vocal tract impulse
response can then be obtained as the impulse response of the synthesis filter,
\[ H(z) = \frac{1}{1 + \sum_{k=1}^{p} a_k z^{-k}} \] (8)
Since all-pole modeling results in an IIR response, truncated impulse response is used as the impulse response vector \( h \). Through the iterations, the signal estimate \( x \) is constrained to be the output of an all-pole system described in (7) resulting in an accurate estimation of both the signal and the transform.

4. Experiments and Results
For these CS experiments, we consider a BBC broadcast news speech clip of duration 30sec spoken by a male talker and sampled at 8kHz. Speech signal is analyzed in non-overlapping frames of duration 40ms. The measurement matrix for CS measurements \( \Phi \) is populated with i.i.d normal random samples. The signal adaptive transform is obtained from reconstructed speech signal \( x \) using linear prediction of order \( p = 10 \). The sparse reconstruction step in the proposed iterative algorithm is performed using (i) IPHT algorithm discussed in section 3, and (ii) Matching Pursuit (MP). The algorithm is repeated for 40 iterations. The Performance measures to compare the reconstruction quality are the SNR in signal space \( (SNR_s) \), which compares the closeness of the estimate to the actual signal, SNR in measurement space \( (SNR_y) \), which accounts for the consistency of the solution with CS measurements, and the log-likelihood ratio (LLR), which measures the closeness of the estimated LP parameters with the actual LP parameters of the signal frame. LLR is a sensitive measure of speech intelligibility comparing the CS estimated impulse response and the target signal impulse response. All the performance measures are examined for different sub-sampling factors, less than the Nyquist rate of 8kHz.

Figure 2 shows a frame of the original and reconstructed signal along with their log-spectra and the variation of SNR as a function of iterations. We compare IPHT and MP performance effectiveness. We can see that the reconstructed waveforms are well matched with the original; also the signal log-spectra which shows a huge dynamic range of 60-80 dB is very well matched for the IPHT case and not so well for the MP case. From the SNR plots, we can see that the SNRs are generally increasing, although not exactly monotonic. It can be seen that IPHT reaches saturation level with very few (\( \sim 5 \)) iterations, whereas MP requires about 25 iterations. More importantly, IPHT provides an SNR nearly same as \( SNR_y \), whereas in MP these two have a gap of approximately 20dB. Thus IPHT performance excels in all respects compared to MP. The SNR level reached is also close to 25dB which is almost toll quality, much better than that of MP (\( \sim 18dB \)).

In Figure 3, we show the success of CS reconstruction, with respect to non-stationarity in speech. We can see that \( SNR_x \) is always > 20dB, which is very good for unknown signal adaptive transform. The -ve LLR plot also shows that estimation of the log-spectrum is uniformly good over the entire speech. Figure 3 shows the narrow band spectrogram of a 2.5 sec. segment of the original and the reconstructed speech. It can be seen that both formant contours (\( \cdot \) of \( H \)) and pitch contours (\( \cdot \) of \( r \)) are well reconstructed in the CS reconstruction. Best performance is obtained for frames which correspond to voiced regions and exhibit periodicity, and the worst performance is for the frames which have a transition from unvoiced to voiced or vice-versa.

CS addresses the issue of sub-Nyquist sampling, i.e., \( M < N \). Figure 4 gives the comparison of \( SNR_x, SNR_y \) as well as log-likelihood ratio for varying number of measurements and sparsity. As \( M \) is varied from 25% to 75% of the Nyquist rate, and varying sparsity \( K=20:20:M-20 \), we can see that the reconstruction SNR is increasing with increasing number of measurements which is as expected. We can also see a monotonic increase in average segmental SNR with increasing \( K \). Log-likelihood ratio is found to decrease with increasing measurements and sparsity as expected. For \( M=240 \), the LLR is less than 0.5 for all values of \( K \), which shows that the impulse response estimated using the iterative algorithm is close to the actual impulse response derived using the original signal frame. A closer look at \( SNR_x \), and LLR plots w.r.t sparsity \( K \) reveals that, for \( K < 100 \), though \( SNR_x \), obtained is almost the same for \( M=160 \) and 240, the LLR is low for higher number of measurements. This shows that number of measurements is important for accurate estimation of the transform. Figure 4 also compares the performance of the proposed method with the codebook constrained approach in [4] with a codebook of size 256. We can see that for \( K > 140 \), the proposed method gives a better reconstruction quality and parameter accuracy compared to the approach in [4].

5. Conclusion
We have presented an iterative CS reconstruction algorithm based on hard thresholding, which can estimate a signal adaptive transform as well as the sparse vector from the CS measurements. We have shown that using the source-excitation model of speech production, it is possible to estimate the all-pole system parameters and the sparse excitation vector of time-varying speech. The signal reconstruction accuracy is \( \sim 25dB \) and the all-pole model parameter accuracy is \( \sim 0.25 \), better than a 256 size optimum codebook distortion. This approach can be useful for several speech applications, such as speech recognition, speaker verification etc.

6. References
Figure 2: CS recovery: (a,b) Original and reconstructed signals, (c,d) log spectrum, and (e,f) SNR in signal space and measurement space w.r.t iterations. The left column shows performance of IPHT based CS and right column shows MP based CS. Frame size N=320 (40 msec), number of CS measurements M=240, and sparsity K=220.

Figure 3: Spectrogram of a 2.5 second segment of (a) the original speech signal, (b) the reconstructed speech signal, and (c) shows the segment wise SNR in signal space and the log-likelihood ratio. Frame size N=320 (40 msec), number of CS measurements M=240, and sparsity K=220.

Figure 4: Comparison of average segmental reconstruction performance for varying number of CS measurements (M) and the sparsity (K) for a fixed frame size of 320 samples (a) Signal space SNR, (b) Measurement space SNR, and (c) Log-Likelihood Ratio. The plot in “black” shows the results for the codebook constrained approach in [4].