COMPRESSIVE SENSING FOR SPARSELY EXCITED SPEECH SIGNALS

T.V. Sreenivas[1] *and W. Bastiaan Kleijn*[2]

[1]Department of ECE, Indian Institute of Science, Bangalore-12, India
[2]ACCESS Linnaeus Center, Electrical Engineering KTH - Royal Institute of Technology, Stockholm
tvsree@ece.iisc.ernet.in, bastiaan.kleijn@ee.kth.se

ABSTRACT

Compressive sensing (CS) has been proposed for signals with sparsity in a linear transform domain. We explore a signal dependent unknown linear transform, namely the impulse response matrix operating on a sparse excitation, as in the linear model of speech production, for recovering compressive sensed speech. Since the linear transform is signal dependent and unknown, unlike the standard CS formulation, a codebook of transfer functions is proposed in a matching pursuit (MP) framework for CS recovery. It is found that MP is efficient and effective to recover CS encoded speech as well as jointly estimate the linear model. Moderate number of CS measurements and low order sparsity estimate will result in MP converge to the same linear transform as direct VQ of the LP vector derived from the original signal. There is also high positive correlation between signal domain approximation and CS measurement domain approximation for a large variety of speech spectra.

Index Terms— sampling, compressed sensing, matching pursuit, sparse signal reconstruction

1. INTRODUCTION

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]. The traditional approach of sampling a signal at the Nyquist rate (twice the bandwidth) and then removing redundant information before efficient transmission or storage, requires usually a lot of signal processing at the transmitter (encoder), although the receiver (decoder) is relatively simple. Of course, in full-duplex (two-way) communication, there is need for an encoder and decoder at each user terminal and hence, the complexity difference may appear unimportant. In several new wireless applications such as wireless sensors and hands free communication, it is important that the signal acquisition be as efficient as possible in terms of power consumption and hence the computational complexity. This is similar to the requirement in sensor networks used in remote sensing.

Compressive sensing (or compressed sampling) provides for both sampling as well as compression, along with encryption of the source information, simultaneously. Also, signal reconstruction quality can be traded with the available complexity at the wireless receiver node. All these four advantages are very important for a communication application and we would like to explore the use of CS for communicating speech and audio signals.

The theory of compressive sensing provides for signal reconstruction from random projections of a signal vector, provided the signal is known to be sparse in a vector space. To quote [3], *"CS theory asserts that one can recover certain signals and images from far fewer samples or measurements than traditional methods use,"* such as Nyquist sampling. Many practical signals do satisfy the sparse property in some linear transform domain of the signal, such as the wavelet domain for images. The CS theory assertion that the number of measurements required is proportional to the sparse factor and has nothing to do with the Fourier bandwidth of the signal is very significant. The sparse property is a measure of signal redundancy (such as in DCT domain truncation) and CS permits to exploit this redundancy right at the signal acquisition stage, instead of a subsequent stage of compression.

Application of CS to speech and audio is not straight forward, since the signals constitute a very large class of production mechanisms, emphasizing different characteristics of the signal at different times. The domain in which their sparsity can be exploited is also not clear and their degree of sparsity. The perceptual properties of the reconstructed signal and the computational constraints also become important for a practical application, since the basic CS formulation is very computation intensive. In this paper, we show that recovery is possible from sub-Nyquist rate CS of speech and joint estimation of sparse excitation and the linear system, using the matching pursuit (MP) based iterative estimation.

2. COMPRESSIVE SENSING FORMULATION

Let $\mathbf{x} \in \mathcal{R}^{\mathbf{N}}$ be the signal and let $\mathbf{\Psi} = \{\psi_1, \psi_2, ...\psi_N\}$ be the basis vectors spanning $\mathcal{R}^N$. The signal is said to be "T-sparse" if

$$\mathbf{x} = \sum_{i=1}^{T} s_{n_i}\psi_{n_i}, \quad \{n_1, n_2, ...n_T\} \subset \{1, ..., N\} \tag{1}$$

where s_{n_i} are scalar coefficients and $T << N$. We can say that $\mathbf{\Psi}$ is our knowledge about $\mathbf{x}$ that provides the key to compressive sensing. Hence, $\mathbf{x} = \mathbf{\Psi} \cdot \mathbf{s}$ where $\mathbf{s}$ is the sparse vector with only T non-zero elements, indexed by n_i that are unknown. According to CS theory, we can do sampling of $\mathbf{x}$ through projections onto random bases and reconstruct the signal at a receiver with full knowledge of the random bases; i.e., the sampling (sensing) measurements are:

$$y_m = \sum_{i=1}^{N} \phi_m(i) \cdot x(i), \;\; 1 \le m \le M < N \tag{2}$$

or $\mathbf{y} = \mathbf{\Phi} \cdot \mathbf{x}$, where $\mathbf{\Phi}$ is a $M \times N$ measurement matrix. The $\mathbf{\Phi}$ is made up of orthonormal random basis vectors $\phi_{\mathbf{m}}$. Under the condition that $\mathbf{\Phi}$ and $\mathbf{\Psi}$ are "incoherent" it has been proved that $\mathbf{x}$ can be reconstructed from $\mathbf{y}$ with high probability if $M > T\log(N)$

Thanks to ACCESS Linnaeus Centre, Electrical Engineering KTH, for the financial support to the first author for the summer visit to KTH.

[3]. The reconstruction method proposed in [3] is through convex optimization:

$$\begin{aligned} \hat{\mathbf{s}} =& \underset{\mathbf{s}}{argmin}\ \|\mathbf{s}\|_1,\ subject\ to\ \mathbf{y} = \mathbf{\Phi}\cdot\mathbf{\Psi}\cdot\mathbf{s} \\ and\ \ \hat{\mathbf{x}} =& \mathbf{\Psi}\cdot\hat{\mathbf{s}}, \end{aligned} \tag{3}$$

where $\|.\|_1$ is the ℓ_1 norm. The measurements being projections onto random vectors, they are inherently encrypted and the number of measurements is related to the sparsity of the signal, hence compressed. The computation at the sensor is minimal whereas reconstruction at the receiver is iterative, which can also provide for graded reconstruction depending on the complexity of the receiver. The CS formulation is also referred to as *"basis pursuit"* (BP) since a subset of the column vectors of $\mathbf{\Phi}\cdot\mathbf{\Psi}$ is being determined. One of the efficient algorithms to solve CS by interpreting it as a sparse approximation is using *"orthogonal matching pursuit"* (OMP) [4], which can be formulated as:

$$\hat{\mathbf{s}} = \underset{\mathbf{s}}{argmin}\ \|\mathbf{y} - \mathbf{\Phi}\cdot\mathbf{\Psi}\cdot\mathbf{s}\|_2,\ and\ \|\mathbf{s}\|_0 = T. \tag{4}$$

The CS problem can also be interpreted as one of sparse signal approximation, since the measurement vector $\mathbf{y}$ is likely to have noise in most practical cases. Hence, CS reconstruction has been posed as one of several optimization problems: [6]

$$\hat{\mathbf{s}} = \underset{\mathbf{s}}{argmin}\ \|\mathbf{s}\|_1,\ subject\ to\ \|\mathbf{y} - \mathbf{\Phi}\cdot\mathbf{\Psi}\cdot\mathbf{s}\|_2 < \epsilon \tag{5}$$

$$or\ \ \hat{\mathbf{s}} = \underset{\mathbf{s}}{argmin}\ \|\mathbf{y} - \mathbf{\Phi}\cdot\mathbf{\Psi}\cdot\mathbf{s}\|_2,\ subject\ to\ \|\mathbf{s}\|_1 < \tau \tag{6}$$

$$or\ \ \hat{\mathbf{s}} = \underset{\mathbf{s}}{argmin}\ \left[\|\mathbf{y} - \mathbf{\Phi}\cdot\mathbf{\Psi}\cdot\mathbf{s}\|_2 + \tau.\|\mathbf{s}\|_1\right] \tag{7}$$

Various solutions to sparse approximation have been proposed, such as matching pursuit (MP), LASSO (least absolute shrinkage and selection operator), basis pursuit (BP), gradient pursuit (GP). The performance of sparse signal reconstruction does show some interdependence between the number of measurements, measurement noise, signal sparsity and the reconstruction algorithm itself [5, 6]. It has been found that low complexity greedy solutions, such as matching pursuit or even direct thresholding, can provide comparable performance to that of iterative convex optimization using BP, under certain conditions. MP is also attractive because it is extendable to the case of distributed compressive sensing and also the possibility of perceptually motivated extensions [7].

The sparse excitation formulation presented here is different from eq.(4). To exploit excitation sparsity of speech, we use a signal dependent impulse response matrix $\mathbf{h}$ in place of $\mathbf{\Psi}$. This requires estimation of both $\mathbf{h}$, $\mathbf{s}$ jointly, which is shown to be possible through the matching pursuit approach.

3. SPARSITY IN SPEECH

The successful models of generating speech and audio signals have been (i) linear system model for speech and (ii) sinusoidal model (AM-FM) for both speech and music. [8] Both are parametric models and parsimoniously represent the time-varying nature of these signals. (For high (transparent) quality reconstruction, such as for music, production models along with a residual signal model is used.) Because of the time-varying nature, we need to do sensing and compressing of a short duration of the signal. It is known that the perceptually significant features of spectral resonances (formants) and the harmonicity due to periodic excitation, are the most important and basic parameters in speech and audio. In speech, we perform either linear prediction analysis or cepstrum analysis of the signal $\mathbf{x}$ to separate the formant and periodicity information.

To explore sparsity, similar to $\mathbf{x} = \mathbf{\Psi}\cdot\mathbf{s}$ in CS, we can consider several alternative representations of a speech frame:

$$\mathbf{x} = \mathbf{C}^{-1}\cdot\boldsymbol{\theta_1} \tag{8}$$

$$\mathbf{x} = \mathbf{F}^{-1}\cdot\boldsymbol{\theta_2} \tag{9}$$

$$\mathbf{x} = \mathbf{F}^{-1}\cdot\exp\{\mathbf{F}\cdot\boldsymbol{\theta_3}\} \tag{10}$$

$$\mathbf{x} = \mathbf{h}\cdot\mathbf{r} \tag{11}$$

Eq.(8) describes a DCT where $\mathbf{C}$ is the real valued transform matrix and $\boldsymbol{\theta_1}$ are the DCT coefficients. Similarly, $\boldsymbol{\theta_2}$ corresponds to the DFT matrix $\mathbf{F}$, which is complex valued. $\boldsymbol{\theta_3}$ corresponds to the homomorphic mapping to the cepstrum domain. The formant and periodicity parameters become additive in the cepstrum domain, whereas they are multiplicative in the spectral domain. However, the cepstrum relation to $\mathbf{x}$ is non-linear, although homomorphic; this would lead to non-linear constraints in eq.(3) and hence, it will be even more difficult to solve the CS recovery problem. The DCT or DFT mapping is linear, but the degree of sparsity would be in question for speech or audio. The signal periodicity induces harmonicity in the spectrum as well as *spreading of the harmonics due to the implicit windowing of the signal.* (In sinusoidal modeling, a more compact representation of the amplitude, frequency and phase parameters is possible because of the generative model of synthesis, which cannot be expressed as a linear transform required in CS.)

The fourth alternative given by eq.(11) provides a linear sparse representation of speech in the time domain itself. The periodicity in the signal gets reflected as harmonics in the spectrum and the number of periods in a typical window is far fewer than the number of harmonics in the spectrum, resulting in a greater sparsity of the representation. Aperiodicities in speech causes spreading of the harmonics, further reducing sparsity; but, this does not affect signal domain sparsity. Using the linear model of speech production, the signal spectrum $\boldsymbol{\theta_2}$ can be composed as the product of excitation and vocal-tract spectra and the excitation then can be obtained through inverse filtering based on the spectral envelope. Let the convolution relation be $x[n] = h[n] * r[n]$, where $h[n]$ is the signal domain impulse response of the smooth spectral envelope and $r[n]$ is the residual excitation component. The convolution relation can be expressed in a matrix form given in eq.(11), where $\mathbf{h}$ is $N \times N$ impulse response matrix and $\mathbf{r}$ is a $N \times 1$ excitation vector. The matrix $\mathbf{h}$ would be Toeplitz lower triangular for linear convolution and circulant Toeplitz for circular convolution. The representation using θ_2 is related to eq.(11) through the respective Fourier transforms: $\theta_2[k] = H[k]\cdot r[k]$. Often, $H[k]$ is derived from $\mathbf{x}$ using linear prediction or cepstrum formulation [9].

The AR (autoregressive) parametric approximation results in an IIR response which is truncated to FIR to obtain $\mathbf{h}$; the truncated length K can be $> N$. This leads to $\mathbf{h}$ being $N \times K$, $K > N$, bringing in the effect of previous frame excitation signal also.

4. SPARSE SIGNAL RECONSTRUCTION

It has been shown that redundant dictionaries are useful for the success of MP; hence, $\mathbf{\Psi}$ in (4) need not be $N \times N$ but $N \times K$, $K > N$ [6], inducing a higher dimensionality for the sparse vector $\mathbf{s}$. Thus, $\mathbf{\Phi}\cdot\mathbf{\Psi}$ would be a $M \times K$ matrix with $M < N < K$.

For speech, a sparse model based on DCT or DFT is directly usable in eqs.(3) or (4) with the substitution of $\mathbf{\Psi} = \mathbf{C^{-1}}$ or $\mathbf{\Psi} =$

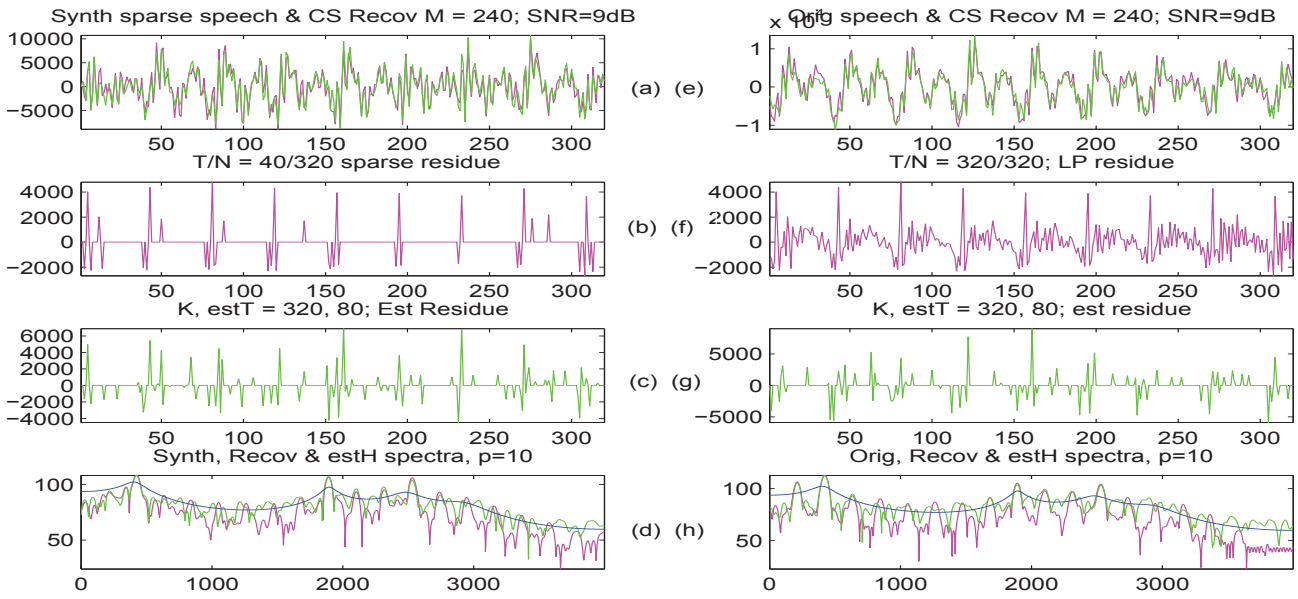

Fig. 1. *Example of CS recovery of speech: (a,e) 40ms speech frame, (b,f) residue signal (c,g) estimated residue (d,h) spectra of estimated LP, speech signal and recovered signal. Left column: signal with exact sparsity. Right column: original speech with approximate sparsity.*

$\mathbf{F^{-1}}$. For the inverse filter based sparse model, $\mathbf{x} = \mathbf{h} \cdot \mathbf{r}$, the impulse response matrix $\mathbf{h}$ is signal dependent, unlike the DCT or DFT matrices. $\mathbf{h}$ can be chosen to be an $N \times N$ circulant or an $N \times K$ Toeplitz for either circular convolution or linear convolution of the impulse response. Since $\mathbf{h}$ is signal dependent and unknown, we propose to construct a codebook of size L of such matrices from the training speech data: $\mathbf{h} \in \{\mathbf{h_1}, \mathbf{h_2}, ..., \mathbf{h_L}\}$. Then, the CS optimization problem in the time domain can be stated as:

$$[\hat{\mathbf{r}}, \hat{\mathbf{h}}] = \underset{\mathbf{r},\mathbf{h}}{argmin} \, \|\mathbf{y} - \mathbf{\Phi} \cdot \mathbf{h} \cdot \mathbf{r}\|_2, \; and \; \|\mathbf{r}\|_0 = T, \qquad (12)$$

$$and \; \hat{\mathbf{x}} = \hat{\mathbf{h}} \cdot \hat{\mathbf{r}}$$

where the search is over the cross product of matrix codebook and the basis vector set $\mathbf{\Phi}$. The solution complexity gets magnified linearly by a factor of L, the size of the codebook. Since we use MP, this is manageable. It may be noted that the measurement vector $\mathbf{y}$ is $M \times 1$ and we apply matching pursuit to the product matrix $\mathbf{\Phi} \cdot \mathbf{h}$, which is $M \times K$. We choose $K > N > M$ resulting in a redundant basis in the measurement space. The use of redundant bases has been reported to aid the MP approach to sparse reconstruction [5] and it also suits the IIR nature of $\mathbf{h}$; hence, we experiment with $K = N$ and $K > N$ for signal recovery using matching pursuit.

5. EXPERIMENTS

We construct a codebook of $\mathbf{h}$ using the LSF representation of speech. Using clean speech, sampled at $8KHz$, of a male speaker of duration $\approx$ $70s$, successive frames of $40ms$ duration signal, with $20ms$ overlap, are analyzed using the autocorrelation LP formulation, of order $p = 10$, to obtain the LP coefficients and the corresponding LSF coefficients. The residue signal power is also saved as a parameter along with LSF, to obtain a 11 dimension VQ codebook of size $L = 128$ using the LBG algorithm [9].

For the CS experiment, the same speech signal is analyzed in successive frames of $40ms$ duration, using a different Φ for each frame. (This provides the matched speaker CS performance which is then extended to mismatch speaker condition). The Φ, of size $M \times N$, is populated with ZMUV (zero mean unit variance) Gaussian samples, whose columns are nearly uncorrelated. Fixing $N = 320$, varying $M = 10 : 40 : 320$, $K = 320, 640$ and $\hat{T}$ are modified as experimental parameters. Exact sparsity of the signal is obtained by thresholding the LP residual to T number of largest magnitude samples and reconstructing the signal; T is varied to control sparsity, where T=320 gives the original signal. We use the performance measures of recovered signal SNR with respect to the synthesized signal, log-likelihood ratio (LLR) [9] of the estimated LP parameters with respect to those used for synthesis and SNR of CS recovery in the $\mathbf{y}$ measurement domain.

Fig.1 shows an example frame of CS recovery; the left column signal has a sparsity of T=40 and the right column is original signal (T=320); $\hat{T}$ is fixed as 80 for both and the number of measurements M=240. It can be seen that the quality of reconstruction is fairly good for both the cases, indicating that MP is effective not only for sparse signals, but nearly sparse also, such as the original speech. More importantly, it is found that the estimated LP spectrum is nearly the same as the direct VQ of exact LP parameters, in both cases. The estimated sparse residue also depicts some degree of periodicity in the signal, better with the original signal.

The scatter plot in Fig.2 shows the statistical behaviour over the whole $70s$ of speech. Each of the colored regions shows the strong positive correlation between the signal domain performance and that of CS measurement domain. Over the whole speech, reconstruction error is in the range of $5 - 30dB$ SNR, indicating that the technique works well for a large variety of speech spectra and atleast $5dB$ SNR. The best performance (blue) is for higher measurements, $M = 320$, and large redundancy $K = 640$. As expected, redundant bases provide some advantage, but the performance is only slightly

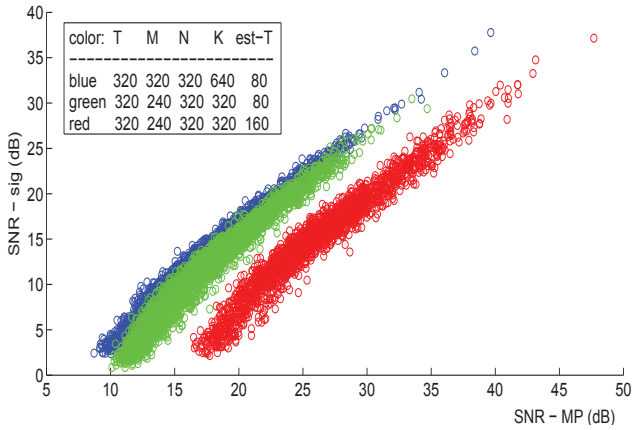

Fig. 2. *Scatter plot of signal SNR Vs measurement space (MP) SNR for different M, K and est-T; approx.* 3000 *frames of speech is shown.*

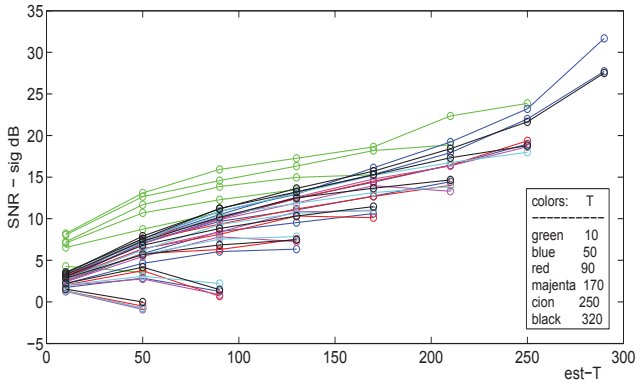

Fig. 3. *Average recovered signal SNR for different T (color), increasing M and est-T; M indicated by the final value of est-T.*

sacrificed for lower M and K (green), requiring much less computation. Increasing $\hat{T}$ keeping other parameters same as in green, increased the CS domain performance, but did not alter SNR much. To examine this more closely, a few frames of the signal are analyzed in a Monte-Carlo fashion, by averaging the SNR and LLR performance for ten repetetions of different $\mathbf{\Phi}$, for each combination of the parameter set: T, M, $\hat{T}$, fixing $K = 320$. These results are presented in Figs.3 and 4. It is found that, except for the case of very low sparsity ($T = 10$), full residue ($T = 320$) SNR-sig performance is close to that of $T = 50$ indicating robustness to approximate sparsity. Moderate number of measurements ($M > 130$) is required for a monotonic increase in performance with respect to $\hat{T}$. Interestingly, this comes out in Fig.4 where the LP estimation error (LLR) is lowest for $\hat{T}$ in the range of $50 - 90$. For small M, the LLR is significantly high, whereas otherwise, the LP estimate is close to the optimum.

We note that low LLR distortion implies good intelligibility of speech and high SNR implies good listening quality. Thus, with a combination of moderate M, K and $\hat{T}$ it is possible to achieve useful speech reconstruction through sub-Nyquist sampled CS. Informal listening to the recovered speech also corroborate this conclusion.

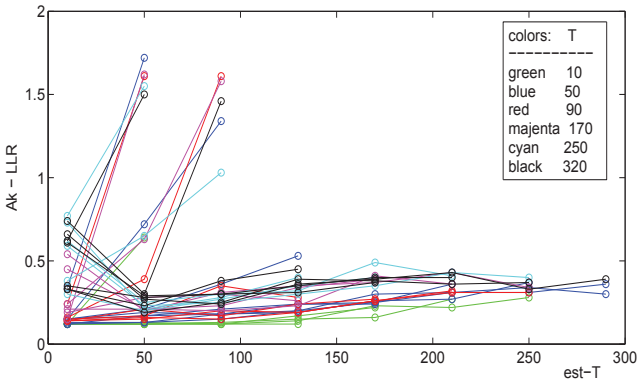

Fig. 4. *Average Log-likelihood ratio of estimated LP spectrum wrt signal, for different T (color), increasing M and est-T; M indicated by last point of est-T.*

6. CONCLUSIONS

We have shown that compressive sensing recovery is possible for sparsely excited signals, *even when the sparsity inducing impulse response matrix is unknown,* such as in speech. Matching pursuit is found to be effective and efficient to jointly estimate both the sparse excitation and the impulse response matrix. The signal reconstruction accuracy can be in the range of $10 - 30dB$ which can be useful for specific speech applications, such as recognition or coding.

7. REFERENCES

[1] Donoho, D., "Compressed sensing," IEEE Trans. Inf. Th., Vol.52, No.4, pp 1289-1306, Apr 2006.

[2] Baraniuk, R.G., "Compressive Sensing," IEEE Sig. Proc. Mag., pp 118-120, Jul 2007.

[3] Candes, E.J. and Wakin, M.B., "An introduction to compressive sampling," IEEE Sig Proc. Mag., pp 21-30, Mar 2008.

[4] Tropp, J.A. and Gilbert, A.C., "Signal recovery from random measurements via orthogonal matching pursuit," IEEE Trans. Inf. Th., Vol.53, No.12, pp 4655-4666, Dec 2007.

[5] Rauhut, H., Schnass, K. and Vandergheynst, P., "Compressed sensing and redundant dictionaries," IEEE Trans. Inf. Th., Vol.54, No.5, May 2008.

[6] Figueiredo, M.A.T., Nowak, R.D. and Wright, S.J., "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," IEEE J. Sel. Top. Signal Proc., 2007.

[7] Heusdens, R. Vafin, R., and Kleijn, W.B., "Sinusoidal Modeling Using Psychoacoustic-Adaptive Matching Pursuits," *IEEE Signal Proc. Letters,* Vol.9, No.8, pp 262-265, Aug 2002.

[8] Goodwin, M., "Adaptive Signal Models: Theory, Algorithms and Audio applications," Ph.D. Thesis, Univ California, Berkeley, USA, 1997.

[9] Rabiner, L.R. and Juang, B.H., *Fundamentals of Speech Recognition,* Prentice Hall, USA, 1993.