Retrieving Sparse Patterns using a Compressed Sensing Framework: Applications to Speech Coding based on Sparse Linear Prediction

Daniele Giacobello, *Student Member, IEEE,* Mads Græsbøll Christensen, *Member, IEEE,* Manohar N. Murthi, *Member, IEEE,* Søren Holdt Jensen, *Senior Member, IEEE,* and Marc Moonen, *Fellow, IEEE*

***Abstract*—Encouraged by the promising application of compressed sensing in signal compression, we investigate its formulation and application in the context of speech coding based on sparse linear prediction. In particular, a compressed sensing method can be devised to compute a sparse approximation of speech in the residual domain when sparse linear prediction is involved. We compare the method of computing a sparse prediction residual with the optimal technique based on an exhaustive search of the possible nonzero locations and the well known Multi-Pulse Excitation, the first encoding technique to introduce the sparsity concept in speech coding. Experimental results demonstrate the potential of compressed sensing in speech coding techniques, offering high perceptual quality with a very sparse approximated prediction residual.**

I. Introduction

Finding a sparse approximation of the prediction residual in Linear Predictive Coding (LPC) has been an active field of research for the past thirty years. A significant result was found with the introduction of the Multi-Pulse Excitation (MPE) technique [1] providing a suboptimal solution to a problem of combinatorial nature. The purpose of this scheme is to find a prediction residual approximation with a minimum number of nonzero elements, still offering a high perceptual quality. MPE quickly evolved to Code Excited Linear Prediction (CELP), where the best residual approximation is selected from a codebook populated with pseudo-random white sequences. This choice was motivated by the statistic of the residual, ideally a sequence of i.i.d. Gaussian samples (due to the use of 2-norm minimization in the LP analysis).

In our recent work, we have utilized recent developments in convex optimization to define a new synergistic predictive framework that aims for a sparse prediction residual rather than the usual minimum variance residual [2], [3]. We have also shown that MPE techniques are better suited in this framework for finding a sparse approximation of the residual. Considering that MPE is itself a sub-optimal approach to modeling prediction residuals, a natural question is whether one can improve upon the performance of MPE by moving towards a more optimal approach of capturing prediction residuals without increasing complexity. Recent work on sparse solutions to linear inverse problems, commonly referred to as compressive sensing (CS), should be able to provide methods for tackling such issues [4]. While CS has been mainly applied to signals such as images with a natural underlying sparse structure, CS methods also seem to be appropriate for signals that are almost sparse, or for which sparsity is imposed [5]. Consequently, one expects that CS methods can be utilized to estimate a sparse residual within a suitably modeled predictive coding framework. In [6] the authors examined the use of CS within speech coding, resulting in a restricted approach in which a codebook of impulse response vectors is utilized in tandem with an orthonormal basis. In [7], one can find a CS formulation of sinusoidal coding of speech.

In this paper, we examine CS within predictive coding of speech. In contrast to the work in [6], we do not utilize a codebook of impulse response vectors, and instead examine the more familiar approach to predictive coding in which the impulse response matrix is specified. In particular, we demonstrate how a CS formulation utilizing the Least Absolute Shrinkage and Selection Operator (LASSO [8]) method allows for a trade-off between the sparsity of the residual and the waveform approximation error. Moreover, this CS approach leads to a reduction in complexity in obtaining sparse residuals, moving closer to the optimal 0-norm solution while keeping the problem tractable through convex optimization tools and projection onto a random basis. In addition, this paper also shows the successful extension of the CS formulation to the case where the basis is not orthogonal, a case which is rarely examined in the CS literature. In simulations, the CS-based predictive coding approach provides better speech quality than that of MPE-based methods at roughly the same complexity.

The paper is structured as follows. In Section II we briefly review the general CS theory. In Section III we introduce the CS formulation for the case of speech coding, providing some significant results in Section IV. Section V will then conclude our work.

II. Compressed Sensing Principles

Compressed sensing (CS) has arguably represented a shift in paradigm in the way we acquire, process and reconstruct

Copyright (c) 2008 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org

Daniele Giacobello and Søren Holdt Jensen are with the Department of Electonic Systems, Aalborg University, 9220 Aalborg, Denmark (e-mails: {dg,shj}@es.aau.dk).

Mads Græsbøll Christensen is with the Department of Media Technology, Aalborg University, 9220 Aalborg, Denmark (email: mgc@imi.aau.dk).

Manohar N. Murthi is with the Department of Electrical and Computer Engineering, University of Miami, Coral Gables, FL 33146, USA (e-mail: mmurthi@miami.edu).

Marc Moonen is with the Department of Electrical Engineering, Katholieke Universiteit Leuven, 3001 Leuven, Belgium (e-mail: marc.moonen@esat.kuleuven.be).

The work of Daniele Giacobello is supported by the Marie Curie EST-SIGNAL Fellowship (http://est-signal.i3s.unice.fr), contract no. MEST-CT-2005-021175.

The work of Manohar N. Murthi is supported by the National Science Foundation via awards CCF-0347229 and CNS-0519933.

signals. In essence, CS exploits prior knowledge about the sparsity of a signal $\mathbf{x}$ in a linear transform domain in order to develop efficient sampling and reconstruction. Let $\mathbf{x} \in \mathbb{R}^N$ be the signal for which we would like to find a sparse representation and $\mathbf{\Psi} = \{\psi_1, \ldots, \psi_N\}$ be the orthonormal basis (or *orthobasis*). Considering the expansion of $\mathbf{x}$ onto the basis $\mathbf{\Psi}$ as:

$$\mathbf{x} = \mathbf{\Psi}\mathbf{r} = \sum_{i=1}^{N} r_i \psi_i \tag{1}$$

where $\mathbf{r}$ is the vector of the scalar coefficients of $\mathbf{x}$ in the orthobasis. The assumption of sparsity means that only K coefficients, with $K << N$, of $\mathbf{r}$ are significant to represent $\mathbf{x}$. In particular, $\mathbf{x}$ is said to be K-sparse if only K nonzero samples in $\mathbf{r}$ are sufficient to represent $\mathbf{x}$ exactly.

In CS we do not observe the K-sparse signal $\mathbf{x}$ directly, instead we record $M < N$ nonadaptive linear measurements:

$$\mathbf{y} = \mathbf{\Phi}\mathbf{x} = \sum_{i=1}^{N} \phi_m(i)x(i), \quad 1 \leq m \leq M < N, \tag{2}$$

where $\mathbf{\Phi} \in \mathbb{R}^{M \times N}$ is a measurement matrix made up of random orthobasis vectors. CS theory states that we can reconstruct $\mathbf{x}$ (or, equivalently $\mathbf{r}$) accurately from $\mathbf{y}$ if $\mathbf{\Phi}$ and $\mathbf{\Psi}$ are incoherent ($\mu(\mathbf{\Psi}, \mathbf{\Phi}) \approx 1$, where $\mu(\mathbf{\Psi}, \mathbf{\Phi})$ is the coherence measure, the largest correlation between any two columns of the basis matrix and the random matrix). This property is easily achievable when the entries of the random matrix $\mathbf{\Phi}$ are i.i.d. Gaussian variables. In this case, the recovery works with high probability if M is in the order of $K \log(N)$ [9]. If the incoherence holds, the following linear program gives an accurate reconstruction with very high probability:

$$\min_{\mathbf{r} \in \mathbb{R}^N} \|\mathbf{r}\|_1 \quad \text{s.t.} \quad \mathbf{y} = \mathbf{\Phi}\mathbf{\Psi}\mathbf{r}, \tag{3}$$

where $\|\mathbf{r}\|_1 = (\sum_{n=1}^{N} |r(n)|)$ is the 1-norm and it is used as a convex relaxation of the so-called 0-norm, the cardinality of a vector.

A very interesting property of CS is that if $\mathbf{x}$ is not K-sparse (or, not exactly K-sparse), the quality of the recovered signal $\mathbf{r}$ (or, equivalently $\mathbf{x}$) is as good as if we were to select only the K largest values before the calculations, and measure them directly. To quote [9]:

> the reconstruction is nearly as good as that provided by an oracle which, with full and perfect knowledge about $\mathbf{r}$, would extract the K most significant pieces of information for us.

This important property, stated elegantly in [10], extends the use of CS to all kinds of signals for which we would like to find a sparse representation. In particular, it allows us to apply CS to signals where K is not defined by the signal $\mathbf{x}$ but by our "need" for sparsity, therefore allowing an approximation error:

$$\mathbf{e} = \mathbf{y} - \mathbf{\Phi}\mathbf{\Psi}\mathbf{r}. \tag{4}$$

The formulation in (4) will then become:

$$\min_{\mathbf{r} \in \mathbb{R}^N} \|\mathbf{r}\|_1 \quad \text{s.t.} \quad \|\mathbf{y} - \mathbf{\Phi}\mathbf{\Psi}\mathbf{r}\|_2^2 \leq \epsilon, \tag{5}$$

where ϵ is the bound for the approximation error. This inequality constrained convex problem can also be rewritten using Lagrange multipliers as:

$$\min_{\mathbf{r} \in \mathbb{R}^N} \|\mathbf{r}\|_1 + \gamma \|\mathbf{y} - \mathbf{\Phi}\mathbf{\Psi}\mathbf{r}\|_2^2. \tag{6}$$

This latest formulation, also called *Least Absolute Shrinkage and Selection Operator* (LASSO [8]), shows more clearly the robustness of CS to signals that are not necessarily sparse and in particular, the trade-off between the sparsity of $\mathbf{r}$ and the approximation error $\mathbf{e}(\gamma, \mathbf{r})$.

Summarizing, if we wish to perform CS, two main ingredients are needed: a domain where the analyzed signal is sparse and the sparsity of this signal. The domain is found through a linear transform while the level of sparsity can be either known or assumed. In the next section we will see how can we define the CS formulation in speech coding.

III. Compressed Sensing Formulation for Speech Coding

A. Definition of the transform domain

In speech coding, the transform domain where the representation is required to be sparse is the prediction residual. In our previous work, we have indeed found very few nonzero samples in the residual when sparse linear prediction is involved [2], [3]. Considering the simple case in which we would like to find a linear predictor $\mathbf{a}$ of order P that provides a sparse residual, the formulation becomes:

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a} \in \mathbb{R}^P} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_1; \tag{7}$$

where:

$$\mathbf{x} = \begin{bmatrix} x(N_1) \\ \vdots \\ x(N_2) \end{bmatrix}, \mathbf{X} = \begin{bmatrix} x(N_1 - 1) & \cdots & x(N_1 - P) \\ \vdots & & \vdots \\ x(N_2 - 1) & \cdots & x(N_2 - P) \end{bmatrix}$$

and $\|\cdot\|_1$ is the 1-norm. The start and end points N_1 and N_2 can be chosen in various ways assuming that $x(n) = 0$ for $n < 1$ and $n > N$. An appropriate choice is $N_1 = 1$ and $N_2 = N + P$ (in the case of 2-norm minimization, this leads to the autocorrelation and to the Yule-Walker equations). The more tractable 1-norm is used as a linear programming relaxation of the sparsity measure, just like in (4). Given a prediction filter $\mathbf{a}$ the residual vector can be expressed as:

$$\mathbf{r} = \mathbf{A}\mathbf{x}, \tag{8}$$

where $\mathbf{A}$ is the $N \times N$ matrix that performs the whitening of the signal, constructed from the coefficients of the predictor $\mathbf{a}$ of order P [11].

Equivalently, we can write:

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{r} = \mathbf{H}\mathbf{r}, \tag{9}$$

where $\mathbf{H}$ is the $N \times N$ inverse matrix of $\mathbf{A}$ and is commonly referred to as the synthesis matrix [11] that maps the residual representation to the original speech domain. In practice, this inversion is not computed explicitly and $\mathbf{H}$ is constructed directly from the impulse response $\mathbf{h}$ of the all pole filter that corresponds to $\mathbf{a}$. Furthermore, the usual approach is to have

$N+P$ columns in $\mathbf{H}$ bringing in the effects of P samples of the residual of the previous frame (the filter state/memory).

It is important to notice that the column vector $\mathbf{r}$ will be now composed of $N+P$ rows, but the first P elements belong to the excitation of the previous speech frame and therefore are fixed and do not affect the minimization process.

It is now clear that the basis vectors matrix is the synthesis matrix $\mathbf{\Psi} = \mathbf{H}$. We can now write:

$$\mathbf{x} = \sum_{i=1}^{K} r_{n_i} \mathbf{h}_{n_i}, \quad \{n_1, n_2, \ldots, n_K\} \subset \{1, \ldots, N+P\}. \tag{10}$$

where $\mathbf{h}_i$ represents the $i-$th column of the matrix $\mathbf{H}$. The formulation then becomes:

$$\hat{\mathbf{r}} = \arg\min_{\mathbf{r}\in\mathbb{R}^N} \|\mathbf{r}\|_1 + \gamma \|\mathbf{y} - \mathbf{\Phi}\mathbf{H}\mathbf{r}\|_2^2 \tag{11}$$

where $\mathbf{y} = \mathbf{\Phi}\mathbf{x}$ is the speech signal compressed through the projection onto the random basis $\mathbf{\Phi}$ of dimension $M \times N$. The second term is now the 2-norm of the difference between the original speech signal and the speech signal with the sparse representation, projected onto the random basis. Assuming that:

$$\|\mathbf{y} - \mathbf{\Phi}\mathbf{H}\mathbf{r}\|_2^2 = \|\mathbf{\Phi}(\mathbf{x} - \mathbf{H}\mathbf{r})\|_2^2 \approx \|\mathbf{x} - \mathbf{H}\mathbf{r}\|_2^2, \tag{12}$$

the problem in (11) can now be seen as a trade-off between the sparsity in the residual vector and the accuracy of the new speech representation $\hat{\mathbf{x}} = \mathbf{H}\hat{\mathbf{r}}$. To ensure simplicity in the preceding and following derivations, we have assumed that no perceptual weighting is performed. The results can then be generalized for an arbitrary weighting filter.

An important aspect that should be taken into consideration is that, if the transformation matrix $\mathbf{\Phi}$ is not exactly orthogonal, such as in the case of $\mathbf{\Phi} = \mathbf{H}$, the recovery is still possible, as long as the incoherence holds ($\mu(\mathbf{\Phi}, \mathbf{H}) \approx 1$) [4].

B. Defining the level of sparsity

CS theory states that for a vector $\mathbf{x}$ of length N with sparsity level K ($K << N$), $M = O(K \log(N))$ random linear projections of $\mathbf{x}$ are sufficient to robustly (i.e., with overwhelming probability) recover $\mathbf{x}$ in polynomial time. With a proper random basis, so that $\mathbf{\Phi}$ and $\mathbf{H}$ are incoherent ($\mu(\mathbf{\Phi}, \mathbf{H}) \approx 1$) [12], as a rule of thumb, four times as many random samples as the number of non-zero sparse samples should be used; therefore, we can simply choose $M = 4K$ [9]. It is now clear that the size of the random matrix $\mathbf{\Phi}$ depends uniquely on the sparsity level K that we expect in the residual vector. Now the question is how sparse do we expect the residual to be? An interesting case for the choice of K is obtained for voiced speech. In this case, the residual $\mathbf{r}$ is a train of impulses. Each impulse is separated by T_p samples, the pitch period of the voiced speech which is inversely proportional to the fundamental frequency f_0. It is now clear that K will depend on T_p; for a segment of voiced speech of length N, we can reasonably assume to find only N/T_p significant samples in the residual, belonging to the impulse train. A coarse estimation of the integer pitch period T_p can be easily obtained by an open-loop search on the autocorrelation function of the vector $\mathbf{x}$. Then the number of random projections sufficient for recovering $\mathbf{x}$ will be $M = 4\frac{N}{T_p}$. In the case of unvoiced speech the choice of K is not direct, however we can use a heuristic approach where $K = k$ is picked when the improvement in the accuracy of the representation between the choice of $K = k$ and $K = k+1$ is negligible.

C. Similarities with Multi-Pulse Excitation

In Multi-Pulse Excitation (MPE) coders the prediction residual consists of K freely located pulses in each segment of length N. This problem is made impractical by its combinatorial nature and a suboptimal algorithm was proposed in [1] where the sparse residual is constructed one pulse at a time. Starting with a zero residual, pulses are added iteratively adding one pulse in the position that minimizes the error between the original and reconstructed speech. The pulse amplitude is then found in an Analysis-by-Synthesis (AbS) scheme. The procedure can be stopped either when a maximum fixed number of amplitudes is found or when adding a new pulse does not improve the quality. MPE provides an approximation to the optimal approach, when all possible combinations of K positions in the approximated residual of length N are analyzed, i.e.:

$$\hat{\mathbf{r}} = \arg\min_{\mathbf{r}\in\mathbb{R}^N} \|\mathbf{x} - \mathbf{H}\mathbf{r}\|_2^2 \quad \text{s.t.} \quad \|\mathbf{r}\|_0 = K. \tag{13}$$

The compressive sensing formulation in (11) can then be seen to approximate (13), finding a trade-off between the information content of the prediction residual and the quality of the synthesized speech.

IV. Experimental Results

To evaluate our method, we have analyzed about one hour of clean speech coming from several different speakers with different characteristics (gender, age, pitch, regional accent) taken from the TIMIT database, re-sampled at 8 kHz. The order of the sparse linear predictor is $P = 10$, the length of the speech frame is $N = 160$ (20 ms). Three methods are compared: the MPE, the CS based approach in (11) and the optimal combinatorial approach (OPT) in (13). For simplicity, no long-term pitch prediction is performed. In the CS formulation, the random matrix $\mathbf{\Phi}$ is populated with Gaussian samples with distribution $N(0,1)$ and the size is chosen according to the level of sparsity we want to retrieve using the relation $M = 4K$. The regularization parameter γ is chosen as the point of maximum curvature of the $L-$curve, using the method presented in [13].

In Figure 1, we present the unquantized results of the three methods in term of the normalized error $\|\mathbf{x} - \hat{\mathbf{x}}\|_2 / \|\mathbf{x}\|_2$, with $\hat{\mathbf{x}} = \mathbf{H}\hat{\mathbf{r}}^{(K)}$ averaged over all frames, choosing different levels of sparsity K. It is clear that for $K > 10$, the CS solution performs similarly to the optimal solution. While for very few samples $K < 5$, the performance is comparable to that of MPE.

In the quantized case, we concentrate our experimental analysis for the two most significant cases ($K = 10$ and $K = 20$). The quantization process uses 20 bits to encode the predictor using 10 line spectral frequencies (providing

TABLE I
COMPARISON BETWEEN THE SPARSE RESIDUAL ESTIMATION METHODS. A 95% CONFIDENCE INTERVALS IS GIVEN FOR EACH VALUE.

METHOD	K	AS-SNR	MOS	t
OPT	10	21.2±3.1	3.25±0.13	343±5
	20	27.2±1.6	3.52±0.09	581±3
CS	10	20.6±2.6	3.13±0.16	0.3±0.1
	20	25.9±1.9	3.49±0.13	0.5±0.1
MPE	10	17.2±4.1	3.03±0.15	0.1±0.2
	20	20.3±3.2	3.22±0.12	0.9±0.3

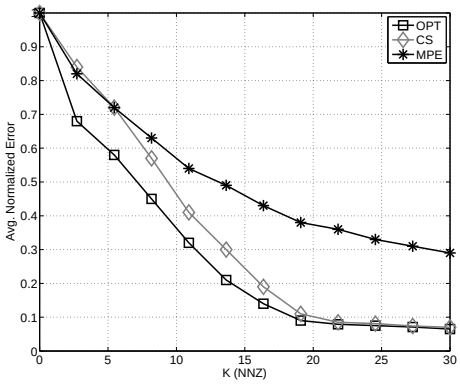

Fig. 1. Number of nonzero samples K versus the average normalized reconstruction error $\|\mathbf{x} - \hat{\mathbf{x}}\|_2/\|\mathbf{x}\|_2$ for a speech segment $\mathbf{x}$. The values corresponding to $K > 30$ are not shown for clarity as the error rates converge to zero.

transparent coding) using split vector quantization. A 3 bit uniform quantizer that goes from the lowest to the highest magnitude of the residual pulses is used to code the residual pulses; 5 bits are used to code the lowest magnitude and 2 bits are used to code the difference between the lowest and highest magnitudes. The signs are coded with 1 bit per each pulse. We postpone the efficient encoding of the positions to further investigation, for now we just use the information content of the pulse location $\log_2 \binom{N}{K}$ bits. The bit rate produced is respectively 5900 bits/s for $K = 10$ and 9500 bits/s for $K = 20$. In Table I, we present the results in terms of Average-Segmental SNR, MOS and empirical computational time t in elapsed CPU seconds of the three methods for the quantized case. It is now clear that the CS formulation achieves similar performances to the optimal case, in a computational time similar to that of MPE.

As mentioned in the previous section, the CS recovery seems also particularly attractive for the analysis and coding of stationary voiced signals. In Figure 2 we see an example of CS recovery of pitch excitation. The open-loop pitch search gives us a coarse approximation of the pitch period of $T_p = 35$ ($f_0 \approx 229$Hz). We then impose $K = \lceil N/T_p \rceil = 10$ and $M = 40$, using the relation $M = 4K$. From the solution we take the $K = 10$ pulses with largest magnitude. We can clearly see that this kind of approximation works very well in the case of voiced speech, retrieving the K pulses belonging to the train of impulses with very high accuracy. The distance between pulses is then approximately T_p.

V. CONCLUSIONS

In this paper we have introduced a new formulation in the context of speech coding based on compressed sensing. The

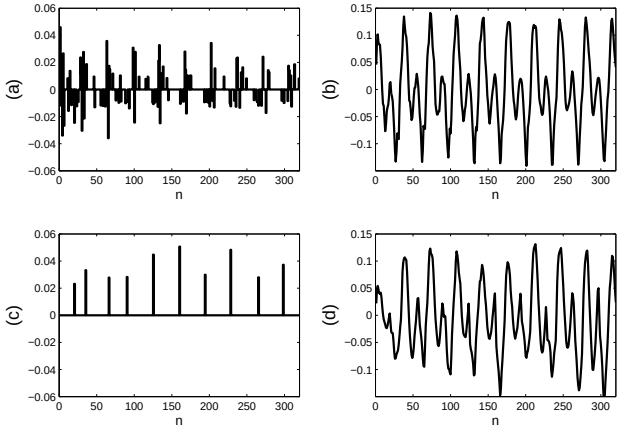

Fig. 2. Example of CS recovery of the pitch excitation for a segment of stationary voiced speech. In (a) we show the estimated excitation using (7) and in (b) the original speech segment. In (c) we show the CS recovered excitation with $K = N/T_p = 320/35 \approx 10$ and in (d) the reconstructed speech segment.

CS formulation based on LASSO has shown to provide an efficient approximation of the 0-norm for the selection of the residual allowing a trade-off between the sparsity imposed on the residual and the waveform approximation error. The convex nature of the problem, and its dimensionality reduction through the projection onto random basis, makes it also computationally efficient. The residual obtained engenders a very compact representation, offering interesting waveform matching properties with very few samples, making it an attractive alternative to common residual encoding procedures. The results obtained also show clearly that CS performs quite well when the basis are not orthogonal, as anticipated in some CS literature.

REFERENCES

[1] B. S. Atal and J. R. Remde, "A new model of LPC excitation for producing natural sounding speech at low bit rates," *Proc. IEEE ICASSP*, vol. 7, pp. 614–617, 1982.

[2] D. Giacobello, M. G. Christensen, J. Dahl, S. H. Jensen, and M. Moonen, "Sparse Linear Predictors for Speech Processing," *Proc. Interspeech*, pp. 1353–1356, 2008.

[3] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen, "Speech Coding Based on Sparse Linear Prediction," *Proc. EUSIPCO*, pp. 2524–2528, 2009.

[4] D. L. Donoho, "Compressed sensing," *IEEE Trans. on Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[5] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Com. on Pure and App. Math.*, vol. 57, pp. 1413–1457, 2004.

[6] T. V. Sreenivas and W. B. Kleijn, "Compressive sensing for sparsely excited speech signals," *Proc. ICASSP*, pp. 4125–4128, 2009.

[7] M. G. Christensen, J. Østergaard, and S. H. Jensen, "On compressed sensing and its applications to speech and audio signals," to appear in *Asilomar Conf.*, 2009.

[8] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Royal. Statist. Soc.*, vol. 58, no. 1, pp. 267-288, 1996.

[9] E. J. Candés, and M. B. Wakin, "An introduction to compressive sampling," *IEEE Sig. Proc. Mag.*, vol. 25, no. 2, pp. 21–30, 2008.

[10] E. J. Candés, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Com. on Pure and App. Math.*, vol. 59, no. 8, pp. 1207-1223, 2006.

[11] L. Scharf, *Statistical Signal Processing*, Addison-Wesley, 1991.

[12] E. J. Candés and J. Romberg, "Sparsity and incoherence in compressive sampling," *Inv. Problems*, no. 23(3), pp. 969–985.

[13] P. C. Hansen and D. P. O'Leary, "The use of the L-curve in the regularization of discrete ill-posed problems", *SIAM J. on Sci. Comp.*, vol. 14, no. 6, pp. 1487–1503, 1993.