# TIME VARYING LINEAR PREDICTION USING SPARSITY CONSTRAINTS

*Srikanth Raj Chetupalli, T. V. Sreenivas*

Department of Electrical Communication Engineering,
Indian Institute of Science,Bangalore, India-560012
{sraj,tvsree}@ece.iisc.ernet.in

## ABSTRACT

Time-varying linear prediction has been studied in the context of speech signals, in which the auto-regressive (AR) coefficients of the system function are modeled as a linear combination of a set of known bases. Traditionally, least squares minimization is used for the estimation of model parameters of the system. Motivated by the sparse nature of the excitation signal for voiced sounds, we explore the time-varying linear prediction modeling of speech signals using sparsity constraints. Parameter estimation is posed as a 0-norm minimization problem. The re-weighted 1-norm minimization technique is used to estimate the model parameters. We show that for sparsely excited time-varying systems, the formulation models the underlying system function better than the least squares error minimization approach. Evaluation with synthetic and real speech examples show that the estimated model parameters track the formant trajectories closer than the least squares approach.

***Index Terms***— Linear prediction, sparse representation, 1-norm minimization, speech analysis, non-stationary signals, time-varying systems.

## 1. INTRODUCTION

Linear predictive coding (LPC) [1] is by far the most successful signal processing technique used in the study of speech signals. In LPC formulation, speech signal is modeled as the output of a slowly time-varying linear system (vocal-tract) excited by a periodic impulse train input or a random noise input for voiced and un-voiced sounds respectively. Traditional LPC assumes the signal to be quasi-stationary over a short analysis interval, i.e., the system function is constant.

Time-varying linear prediction has been proposed in [2] to analyze the non-stationary speech signal over longer analysis intervals. A $p^{th}$ order all-pole filter with time-varying coefficients is used to model the vocal-tract system function. Time-varying filter coefficients are modeled as a linear combination of a set of known basis functions. The parameters of the system are estimated using a least square error minimization approach similar to the quasi-stationary approach. An inherent assumption in the least squares approach is that the distribution of the excitation signal is Gaussian [1]. This does not model the spiky impulse train excitation of voiced speech signals. To account for the non-Gaussian nature of the excitation for voiced signals, a robust parametric modeling approach is considered in [3], where a Huber's loss function based error metric is used in the minimization problem.

Motivated by the sparsity of the excitation signal for voiced speech segments, a sparse linear prediction frame-work for the study of speech signals is proposed in [4]. Quasi-stationary analysis of speech is performed under the sparse excitation constraints. The work in [4] has shown that the estimated system parameters model the spectral envelope of the signal more accurately, while providing shift-invariant and pitch independent estimates for the system function. Parameter estimation is posed as a $\ell_0$-norm minimization problem. Due to the non-convex nature of the cost function, the $\ell_0$-norm minimization is typically solved using $\ell_1$ relaxation techniques [5]. It has been shown that iterative re-weighted $\ell_1$-norm minimization [6] leads to a solution with enhanced sparsity compared to the $\ell_1$-norm minimization [7].

In this paper, we examine the sparse linear prediction formulation for the time-varying linear prediction problem. As in [2], the parameters of the time-varying linear predictor are assumed to be a linear combination of a set of known basis functions. The parameters of the system are estimated by minimizing the $\ell_0$ norm of the excitation signal. We show that for sparsely excited systems, the new approach is able to track the system changes better than the least-squares (MMSE) approach in [2].

## 2. PROBLEM FORMULATION

Let $x[n]$ be a signal modeled as the output of an auto-regressive (AR) linear system of order $p$

$$x[n] = \sum_{i=1}^{p} a_i[n]x[n-i] + e[n] \tag{1}$$

where $e[n]$ is the excitation signal for the AR linear system. For voiced sounds, the excitation signal $e[n]$ consists of a set of impulses with a period corresponding to the rate of vibration of the vocal-folds. Traditional LPC methods assume the signal to be quasi-stationary over a short interval of 10-30 msec, meaning that the time-varying system parameters $a_i[n]$ are considered to be constant with respect to $n$. Time-varying linear prediction relaxes the quasi-stationary assumption by allowing the system parameters to vary in a parametric manner. Parameters are assumed to be a linear combination of a set of known basis functions $\{u_k[n]\}_{k=1}^{q}$.

$$a_i[n] = \sum_{k=1}^{q} a_{ik}u_k[n] \tag{2}$$

Different choices for $u_k[n]$ are power series, trigonometric series, and piece-wise constant basis functions [2]. In this paper, we employ the power series basis function $u_k[n] = n^k$. Using (1,2) the prediction equation can be written as,

$$\hat{x}[n] = \sum_{i=1}^{p}\sum_{k=0}^{q} a_{ik}u_k[n]x[n-i] \tag{3}$$

The authors thank Google India Private Ltd., for the student travel grant.

and the prediction error is given by:

$$e[n] = x[n] - \hat{x}[n] \tag{4}$$

Using vector notation,

$$e[n] = x[n] - \mathbf{X}_n^T \mathbf{a} \tag{5}$$

where,

$$\begin{aligned} \mathbf{a} &= [a_{10}, a_{11}, \ldots, a_{1q}, \ldots, a_{p0}, a_{p1}, \ldots, a_{pq}]^T \\ \mathbf{X}_n &= [x[n-1], u_1[n] * x[n-1], \ldots, u_q[n] * x[n-1], \\ &\quad \ldots, x[n-p], u_1[n] * x[n-p], \ldots, u_q[n] * x[n-p]]^T \end{aligned} \tag{6}$$

For an analysis interval of $N$ samples, we can express the error vector as,

$$\mathbf{e}_{N\times 1} = \mathbf{x}_{N\times 1} - \mathbf{X}_{N\times p(q+1)} \mathbf{a}_{p(q+1)\times 1} \tag{7}$$

where the rows of $\mathbf{X}$ are formed using $\mathbf{X}_n$ for $n \in [0, N-1]$ and $x[n]$ is defined over $-p \leq n \leq N-1$. The parametric variation of the system function allows for longer analysis intervals, and hence compact modeling of the time-varying properties of the system. The analysis interval for speech can be typically 150-400 msec, which is much longer compared to the quasi-stationary LPC (10-30 msec) interval. Also, the time-varying system is characterized by a total of $p(q+1)$ number of parameters, in comparison to $p$-poles for every analysis interval of 10-30 msec in quasi-stationary analysis. For smaller values of $q$, this might result in a compact model with fewer parameters compared to the quasi-stationary case.

In the time-varying LPC method proposed in [2], the parameters are estimated by minimizing the total squared prediction error (We refer to this method as Ls-TVLPC method). The minimization problem can be equivalently written as,

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_2^2 \tag{8}$$

For a system excited by an i.i.d Gaussian random excitation signal, least squares minimization for the prediction error is also the maximum-likelihood method of parameter estimation [1] and results in an accurate modeling of the system. However, the approach gives only an approximate representation in the case of systems with excitation signal which is non-Gaussian, which is the case for voiced segments of speech. For voiced sounds, the distribution of excitation consists of a large concentration of samples around the mean, and a few samples (excitation impulses) with larger values compared to the zero mean value.

In this paper, we pose the parameter estimation problem using sparsity promoting $\ell_0$-norm (referred to as Sp-TVLPC) based optimization .

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_0 \tag{9}$$

Motivation for the sparsity constraint comes from the impulsive nature of the excitation signal for voiced speech segments. Solving the $\ell_0$-norm minimization problem as such involves combinatorial search, which is NP-hard [8]. The advances in the field of compressive sensing (CS) have led to the development of effective solutions to solve the problem in (9). A standard approach in CS is to use a convex relaxation of the $\ell_0$-norm, i.e., an $\ell_1$-norm; $\ell_1$-norm minimization also results in a sparse solution to the error vector $\mathbf{e}$. In solving for the $\ell_1$ minimization, sparsity can be enhanced by using iterative re-weighted schemes studied in [7]. The $\ell_0$-minimization is thus approximated by solving a set of weighted $\ell_1$-minimization problems. In this paper, we use the iterative re-weighted $\ell_1$-minimization algorithm [6] to solve for 9. The weights $w_k$ are assumed to be unity for the first iteration, and the cost function $\|\mathbf{W}(\mathbf{x} - \mathbf{X}\mathbf{a})\|_1$ is minimized for $\mathbf{a}$, here $\mathbf{W}$ is a diagonal matrix formed using weights $w_k$. The estimated minimum prediction error vector $\hat{\mathbf{e}} = \mathbf{x} - \mathbf{X}\hat{\mathbf{a}}$ is used to update the weights for the next iteration; we use the update equation $w_k = (|\hat{e}[k]| + \epsilon)^{-1}$ in the present work (Alternate choices for updating weights are given in [7]). The parameter $\epsilon$ is typically chosen to be smaller than the expected maximum amplitude of the vector $\mathbf{e}$. The algorithm is summarized in Table. 1.

**Table 1**. *Iterative re-weighted $\ell_1$ minimization algorithm*

| | |
|---|---|
| Initialization: | $i = 0$ |
| | $\mathbf{W}^0 = I_{N\times N}$ |
| Iterations: | |
| step 1: | Solve for the $\ell_1$-minimization problem, |
| | $\hat{\mathbf{a}}^i = \arg\min_{\mathbf{a}} \|\mathbf{W}^i(\mathbf{x} - \mathbf{X}\mathbf{a})\|_1$ |
| | $\hat{\mathbf{e}}^i = (\mathbf{x} - \mathbf{X}\hat{\mathbf{a}}^i)$ |
| step 2: | Update the weight matrix $\mathbf{W}$: |
| | $\mathbf{W}^{i+1} \leftarrow \text{diag}\left(1/\left(|\hat{e}^i[k]| + \epsilon\right)\right)$ |
| step 3: | Increment $i$: $i \leftarrow i + 1$ |
| | goto step 1 and repeat until convergence |

## 3. EXPERIMENTS AND RESULTS

We evaluate the proposed method using synthetic and real speech examples. Power series basis $u_k[n] = n^k$ is used for the expansion of the AR model coefficients. A $(p, q)$ power series model denotes a system function with $p$-poles and a $q^{th}$ order power series basis used for the AR model. In each case, the proposed method is compared with the Ls-TVLPC method. Iterative re-weighted $\ell_1$-norm minimization is implemented using "cvx" tool box [9]. The parameter $\epsilon$ in the algorithm is chosen to be 0.01. The iterations are repeated till the difference in the norm of the excitation signal between successive iterations reduces below $10^{-4}$ (chosen empirically).

### 3.1. Synthetic examples

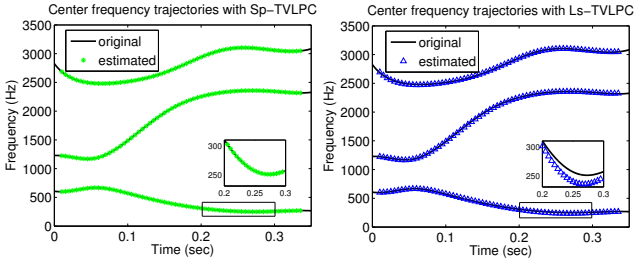
A synthetic speech signal is generated using a cascade of three all-pole time-varying filters excited by a periodic impulse train. The resonant frequencies of the filters are changed such that, the initial resonant frequencies correspond to the vowel '/a/', and the final resonant frequencies correspond to vowel '/i/'. Periodicity of the input excitation is also varied linearly from 100 Hz to 300 Hz and a sampling rate of $fs = 8000$ samples/sec is used. A (6,5) power-series model is used for the time-varying AR system. To evaluate the accuracy of the system representation, we compare the center frequency and the radius trajectories of poles using Sp-TVLPC and Ls-TVLPC methods with the actual values. The error in the estimation of a pole $z[n] = r[n]e^{j\theta[n]}$ is measured using the mean absolute frequency error (MAFE) and mean absolute radius error (MARE) defined as,

$$\text{MAFE} = \frac{1}{N}\sum_{n=1}^{N} \frac{fs}{2\pi} |\theta[n] - \hat{\theta}[n]| \text{ Hz} \tag{10}$$

$$\text{MARE} = 10\log_{10}\left(\frac{1}{N}\sum_{n=1}^{N} |r[n] - \hat{r}[n]|\right) \text{ dB} \tag{11}$$

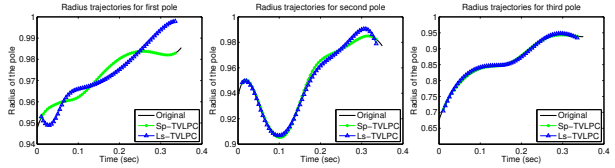
where $N$ is the number of samples in the analysis interval, $r[n]e^{j\theta[n]}$ and $\hat{r}[n]e^{j\hat{\theta}[n]}$ are the simulated and the estimated pole locations.

Fig. 1 shows the estimated center frequency trajectories. From the figure, we can see that the Sp-TVLPC method follows the original center frequency trajectory better than the Ls-TVLPC method.




(a) Sp-TVLPC (b) Ls-TVLPC

**Fig. 1**. *Estimated center frequency trajectories using (a) Sp-TVLPC and (b) Ls-TVLPC.*




(a) First (b) Second (c) Third

**Fig. 2**. *Estimated radius trajectories for the three formants.*

The trajectory estimated using Ls-TVLPC shows a ripple behavior around the original trajectory. The radius trajectories for the three formants are shown in Fig. 2. The radius trajectory estimated using Ls-TVLPC deviates from the actual trajectory very much compared to the Sp-TVLPC method. Figs. 1,2 show that the Sp-TVLPC method results in accurate modeling of the sparse time-varying linear system compared to the Ls-TVLPC method. The MAFE measure obtained is shown in Table 2. The average of MAFE for three formants is close to zero for Sp-TVLPC and it is 4.7 Hz for Ls-TVLPC. However average of MARE for three formants is -24.8 dB for Ls-TVLPC method and below -80 dB for Sp-TVLPC.

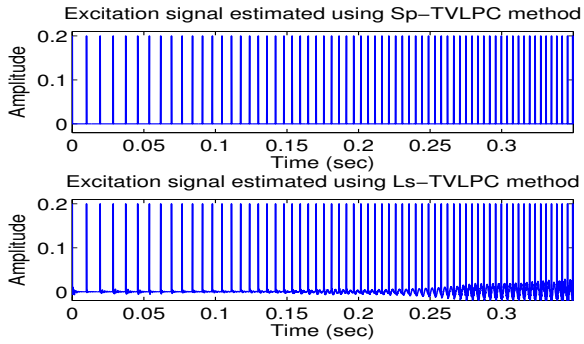



**Fig. 3**. *Comparison of the estimated excitation signal using Sp-TVLPC and Ls-TVLPC methods.*

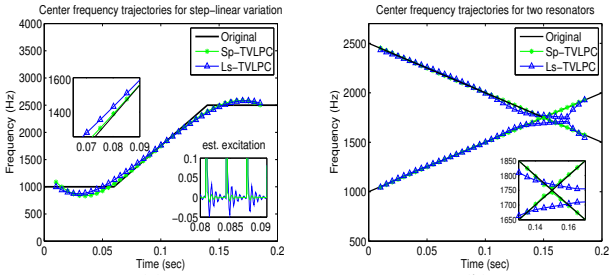
To further demonstrate the power of the Sp-TVLPC method, we compare the excitation signal estimated using the two methods. Fig. 3 shows the excitation signal estimated using Sp-TVLPC and Ls-TVLPC methods. The Sp-TVLPC method estimates the sparse impulse train, while the excitation signal estimated using Ls-TVLPC is not sparse. It can be seen that the estimate obtained using Sp-TVLPC is not affected by the changes in the formant frequencies of the system or the changes in properties of the excitation signal (pitch of the excitation). This demonstrates the robustness of the Sp-TVLPC method to changes in the excitation periodicity. On the other-hand, the error in the excitation signal estimated using Ls-TVLPC is less

| Method | f1 (Hz) | f2 (Hz) | f3 (Hz) | AVG (Hz) |
|---|---|---|---|---|
| Sp-TVLPC | 3.9e-5 | 1e-5 | 2.2e-5 | 2.3e-5 |
| Ls-TVLPC | 4.121 | 3.2897 | 6.8734 | 4.7 |

**Table 2**. *Comparison of MAFE measure for the three formants.*

when the pitch is close to 100 Hz, and more when the pitch is high, indicating the effect of pitch on the error in modeling the system.

The transition between the two vowels in a diphthong is typically smooth with formant frequencies varying gradually from the first vowel to the second vowel in a smooth manner. To evaluate such transitions, an experiment is conducted using a synthetic signal of duration 200 msec, generated using a 2-pole filter with a step-linear variation in the center frequency excited by a periodic impulse train of frequency 300 Hz. The center frequency is kept constant for the initial and final 60 msec of the signal at 1000 Hz and 2500 Hz respectively, and linear variation in between. Fig. 4(a) shows the estimated trajectories using a (2,3) power series model. The power-series model constrains the parameter variation to be smooth. With third order model chosen for AR coefficients, we can see that the system can be modeled only approximately using both the methods. MAFE is found to be 55.6 Hz and 63.04 Hz for Sp-TVLPC and Ls-TVLPC methods respectively, and MARE is -35.3 dB for Sp-TVLPC and -20.5 dB for Ls-TVLPC. Fig. 4(a) also shows the estimated excitation signal which is much sparser in the case of Sp-TVLPC compared to Ls-TVLPC method.




(a) (b)

**Fig. 4**. *Estimated Frequency trajectories for (a) step-linear variation of resonant frequency (estimated excitation during linear portion of the system is also shown), and (b) two cascaded resonators with linearly increasing and decreasing center frequencies.*

Another experiment is carried out to evaluate the usefulness of the proposed approach to track closely spaced center frequency trajectories. The synthetic signal is generated by passing a 300 Hz periodic impulse train through a 4-pole system, with center frequencies varying linearly. Fig. 4(b) shows the estimated center frequency trajectories. Sp-TVLPC method resolves the frequencies better compared to the Ls-TVLPC method. The trajectory estimated using Ls-TVLPC deviates from the actual trajectory as the center frequencies get closer, while Sp-TVLPC follows the trajectory very closely. The deviation from the target trajectory is less in case of Sp-TVLPC even at the cross-over point of the two trajectories. Average of the MAFE for two trajectories is 1.58 Hz for Sp-TVLPC and 15.44 Hz for Ls-TVLPC, and average MARE is -18 dB and -14 dB for Sp-TVLPC and Ls-TVLPC respectively.

Computationally, Sp-TVLPC is much complex compared to the Ls-TVLPC method. Solving one $\ell_1$ minimization takes 5 times more computation time compared to the $\ell_2$ minimization problem; on a Intel(R) Core(TM) i5 CPU M520 @2.40GHz with 4GB RAM, $\ell_1$ minimization takes an average 1.2584 sec, while $\ell_2$ problem is solved in 0.2457 sec. Additionally, computation time scales with the number of iterations in the re-weighted scheme for solving Sp-TVLPC.

| Comparison | AVG SPDIFF (dB) |
|---|---|
| Sp-TVLPC vs Ls-TVLPC | 1.17 |
| Sp-TVLPC vs quasiLPC | 1.25 |
| Ls-TVLPC vs quasiLPC | 1.85 |

**Table 3**. *AVG-SPDIFF comparison for diphthong segment '/$a^I$/'.*

### 3.2. Speech example

Experiments are conducted on real speech signals sampled at 8000 Hz. An analysis interval of 400 msec is chosen for time-varying linear prediction analysis. A (8,5) power series model is used for the time-varying AR coefficients. Since we do not know the actual formant tracks for real speech, we compare the TVLPC methods with the quasi-stationary LPC analysis (quasiLPC) using sparse linear prediction [4]. Quasi-stationary analysis is carried out on intervals of duration 20 msec, with a shift of 5 msec in the successive analysis intervals; the number of poles is chosen to be 8. Note that the total number of AR coefficients required in quasiLPC is 640 while TVLPC methods require only 48 parameters for a total signal duration of 0.4 sec. The signal is windowed using a Hamming window prior to parameter estimation. Estimated time-invariant filter is compared with the time-varying filter evaluated at the center of the analysis interval. The estimated system functions are compared using the average spectral difference measure (AVG-SPDIFF) as in [10]; i.e., the spectral difference is computed as the mean of the instantaneous spectral difference measure evaluated at a few analysis points of interest (for example, the position of pitch pulses in the excitation).

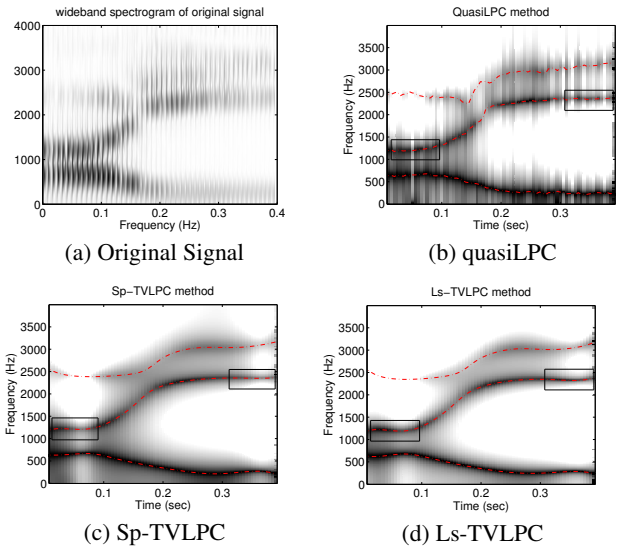
$$\text{AVG-SPDIFF} = \frac{10}{\log_e 10}\left[\frac{2}{L}\sum_{n=1}^{L}\sum_{k=1}^{p}(c_k[n]-c'_k[n])^2\right]^{1/2} \quad (12)$$

where $c_k[n]$ and $c'_k[n]$ are the cepstra of the two systems compared.

Fig. 5 shows the estimated center frequencies of the time-varying all-pole filter for diphthong '/$a^I$/'. Figs. 5(b,c,d) show the frequency response of the instantaneous system functions computed at the center of the analysis interval corresponding to quasi-stationary analysis. From the figures, we can see that the trajectories of both Ls-TVLPC and Sp-TVLPC match closely. The trajectories estimated using time-varying methods correspond to smoothed trajectories of the quasi-stationary analysis case. This can be attributed to the parametric variation of the system function constrained to the power-series basis. From Figs. 5(c,d), we can notice that, in the stationary portions of the spectrum where the formant frequencies are nearly constant, Ls-TVLPC method shows deviation from the expected trajectory showing a ripple behavior, where as Sp-TVLPC estimates are smooth without ripples. Table. 3 shows the AVG-SPDIFF measure between the three methods compared. AVG-SPDIFF is less than 2 dB between the three methods. AVG-SPDIFF between Sp-TVLPC and quasiLPC is less than the difference between Ls-TVLPC and quasiLPC. This is because both Sp-TVLPC and quasiLPC methods are based on $\ell_0$-norm minimization, while Ls-TVLPC uses least-squares minimization.

## 4. CONCLUSIONS

We have examined the sparse linear prediction formulation for the time-varying linear prediction case. Allowing the system function to vary in a parametric manner results in modeling the time-varying spectral envelope of the speech signal. The parameter estimation is posed as an $\ell_0$-norm minimization problem. Through experimental evaluation, we show that the proposed approach of estimating sparse residual translates to more accurate modeling of the system function compared to the least-squares based approach. The present method exploits only the sparsity of the excitation signal, we can model the system itself as sparse for further advantage.




(a) Original Signal (b) quasiLPC

(c) Sp-TVLPC (d) Ls-TVLPC

**Fig. 5**. *(a) Wideband spectrogram of the original signal. (b,c,d) Frequency response of the system function as a function of time using quasi-stationary LPC analysis, Sp-TVLPC, and Ls-TVLPC methods. Estimated formant tracks are also shown.*

## 5. REFERENCES


[1] J. I. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE,* Vol. 32, April 1975, pp. 561-582.

[2] M. G. Hall, A. V. Oppenheim and A. S. Willsky, "Time-varying parametric modeling of speech," *Signal Processing,* Vol. 5, No. 3, 1983, pp. 267-285.

[3] Panbong Ha, Souguil Ann, " Robust time-varying parametric modelling of voiced speech," *Signal Processing,* Vol. 42, No. 3, March 1995, pp. 311-317.

[4] D. Giacobello et. al., " Sparse Linear Prediction and Its Applications to Speech Processing," *Audio, Speech and Language Processing, IEEE transactions on,* Vol. 20, No. 5, July 2012, pp. 1644-1657.

[5] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[6] E. Candes, M. Wakin, and S. Boyd, " Enhancing sparsity by reweighted $\ell_1$ minimization," *J. Fourier Anal. and Appl.* Vol. 14, Issue 5-6, pp. 877-905, Decemeber 2008.

[7] D. Wipf and S, Nagarajan, " Iterative re-weighted $\ell_1$ and $\ell_2$ methods for finding sparse solutions," *IEEE J. Sel. Topics Signal Process.,* Vol. 4, No. 2, Apr. 2010, pp. 317-329.

[8] D. L. Donoho,"Compressed Sensing," *IEEE Trans. Inf. Theory,* Vol.52, No.4, pp. 1289-1306, Apr. 2006.

[9] Michael Grant and Stephen Boyd. CVX: Matlab software for disciplined convex programming, version 2.0 beta. http://cvxr.com/cvx, September 2013.

[10] J. Turner and B. Dickinson, " Linear prediction applied to time-varying all-pole signals," *Proc. 1977 IEEE Int. Conf. Acoust. Speech Signal Process.,* Hartford, CT, may 9-11, 1977, pp. 750-753.