Sparse Linear Prediction and Its Applications to Speech Processing

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

***Abstract*—The aim of this paper is to provide an overview of *Sparse Linear Prediction*, a set of speech processing tools created by introducing sparsity constraints into the linear prediction framework. These tools have shown to be effective in several issues related to modeling and coding of speech signals. For speech analysis, we provide predictors that are accurate in modeling the speech production process and overcome problems related to traditional linear prediction. In particular, the predictors obtained offer a more effective decoupling between the vocal tract transfer function and its underlying excitation, making it a very efficient method for the analysis of voiced speech. For speech coding, we provide predictors that shape the residual according to the characteristics of the sparse encoding techniques creating more straightforward coding strategies. Furthermore, encouraged by the promising application of compressed sensing in signal compression, we investigate its formulation and application to sparse linear predictive coding. The proposed estimators are all solutions to convex optimization problems, which can be solved efficiently and reliably using, e.g., interior-point methods. Extensive experimental results are provided to support the effectiveness of the proposed methods, showing the improvements over traditional linear prediction in both speech analysis and coding.**

***Index Terms*—Linear prediction, speech analysis, speech coding, sparse representation, 1-norm minimization, compressed sensing.**

I. Introduction

Linear prediction (LP) has been successfully applied in many modern speech processing systems ranging from such diverse applications as coding, analysis, synthesis and recognition (see, e.g., [1]). The speech model used in many of these applications is the source-filter model where the speech signal is generated by passing an excitation through an all-pole filter (the predictor in the feedback loop). Typically, the prediction coefficients are identified such that the 2-norm of the residual, the difference between the observed signal and the predicted signal, is minimized. This works well when the excitation signal is Gaussian and independent and identically distributed (i.i.d.) [2], consistent with the equivalent maximum likelihood approach to determine the coefficients [3]. However, when the excitation signal does not satisfy these assumptions, problems arise [2]. This is the case for voiced speech where the excitation can be considered to be a spiky excitation of a quasi-periodic nature [1]. In this case, the spectral cost function associated with the minimization of the 2-norm of the residual can be shown to suffer from certain well known problems such as overemphasis on peaks and cancellation of errors [2]. In general, the shortcomings of LP in spectral envelope modeling can be traced back to the 2-norm minimization approach: by minimizing the 2-norm, the LP filter tries to cancel the input voiced speech harmonics causing the envelope to have a sharper contour than desired with poles close to the unit circle. A wealth of methods have been proposed to mitigate these effects. Some of the proposed techniques involve a general rethinking of the spectral modeling problem (see, e.g., [4], [5], [6], and [7]) while some others are based on changing the statistical assumptions made on the prediction error in the minimization process (notably [8], [9], and [10]).

The above mentioned deficiencies of the 2-norm minimization in LP modeling have also repercussions in the speech coding scenario. In fact, while the 2-norm criterion is consistent with achieving minimal variance of the residual for efficient coding[1], sparse techniques are employed to encode the residual. Examples of this can be seen since early GSM standards with the introduction of multi-pulse excitation (MPE [12]) and regular-pulse excitation (RPE [13]) methods and, more recently, in sparse algebraic codes in code-excited linear prediction (ACELP [14]). Although the sparsity of the RPE and ACELP excitation was motivated, respectively, by psychoacoustic and by the dimensionality reduction of the excitation vector space, a better suited predictor for these two coding schemes clearly is not the one that minimizes the 2-norm, but the one that leaves the fewest non-zero pulses in

Daniele Giacobello is with the Office of the CTO, Broadcom Corporation, Irvine, CA 92617, USA (e-mail: giacobello@broadcom.com).

Mads Græsbøll Christensen is with the Department of Architecture, Design, and 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).

Søren Holdt Jensen ia with the Department of Electronic Systems, Aalborg University, 9220 Aalborg, Denmark (e-mail: shj@es.aau.dk).

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 was supported by the Marie Curie EST-SIGNAL Fellowship (http://est-signal.i3s.unice.fr), contract no. MEST-CT-2005-021175 and was carried out at the Department of Electronic Systems, Aalborg University, 9220 Aalborg, Denmark.

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

[1]The fundamental theorem of predictive quantization [11] states that the mean squared reproduction error in predictive encoding is equal to the mean squared quantization error when the residual signal is presented to the quantizer. Therefore, by minimizing the 2-norm of the residual, these variables have a minimal variance whereby the most efficient coding is achieved.

the residual, i.e., the *sparsest residual*. Early contributions (notably [9], [15], and [16]) have followed this line of thought questioning the fundamental validity of the 2-norm criterion with regards to speech coding. Despite this research effort, to the authors' best knowledge, 2-norm minimization is the only criterion used in commercial speech applications.

Traditional usage of LP is confined to modeling only the the spectral envelope capturing the short-term redundancies of speech. Hence, in the case of voiced speech, the predictor does not fully decorrelate the speech signal because of the long-term redundancies of the underlying pitch excitation. This means that the residual will still have pitch pulses present. The usual approach is then to employ a cascaded structure where LP is initially applied to determine the short-term prediction coefficients to model the spectral envelope and, subsequently, a long-term predictor is determined to model the harmonic behavior of the spectrum [1]. Such a structure is inherently suboptimal since it ignores the interaction between the two different stages. Also in this case, while early contributions have outlined gains in performance in jointly estimating the two filters (the work in [17] is arguably the most successful attempt), the common approach is to distinctly separate the two steps.

The recent developments in the field of sparse signal processing, backed up by significant improvements in convex optimization algorithms (e.g., interior point methods [18][19]), have recently encouraged the authors to explore the concept of sparsity in the LP minimization framework [20]. In particular, while reintroducing well known methods to seek a short-term predictor that produces a residual that is sparse rather than minimum variance, we have also introduced the idea of employing high order sparse predictors to model the cascade of short-term and long-term predictors, engendering a joint estimation of the two [21]. This preliminary work has led the way for the exploitation of the sparse characteristics of the high order predictor and the residual to define more efficient coding techniques. Specifically, in [22], we have demonstrated that the new model achieves a more parsimonious description of a speech segment with interesting direct applications to low bit-rate speech coding. While in these early works, the 1-norm has been reasonably chosen as a convex approximation of the so-called 0-norm[2], in [23] we have applied the reweighted 1-norm algorithm in order to produce a more focused solution to the original problem that we are trying to solve. In this work, we move forward, introducing the novelty of a compressed sensing formulation [24] in sparse LP, that will not only offer important information on how to retrieve the sparse structure of the residual, but will also help reduce the size of the minimization problem, with a clear impact on the computational complexity.

The contribution of this paper is then twofold. Firstly, we put our earlier contributions in a common framework giving an introductory overview of Sparse Linear Prediction and we also introduce its compressed sensing formulation. Secondly, we provide a detailed experimental analysis of its usefulness in modeling and coding applications transcending the well known limitations related to traditional LP.

The paper is organized as follows. In Section II, we provide a prologue that defines the mathematical formulations of the proposed sparse linear predictors. In Section III, we define the sparse linear predictors and, in Section IV, we provide their compressed sensing formulation. The results of the experimental evaluation of the analysis properties of the short-term predictors are outlined in Section V, while the experimental results of the coding properties and applications are outlined in Section VI. We provide a discussion on some of the drawbacks of sparse linear prediction in Section VII. Finally, Section VIII concludes our work.

II. Fundamentals of Linear Prediction

We consider the following speech production model, where a sample of speech $x(n)$ is written as a linear combination of K past samples:

$$x(n) = \sum_{k=1}^{K} a_k x(n-k) + r(n), \tag{1}$$

where $\{a_k\}$ are the prediction coefficients and $r(n)$ is the prediction error. In particular, we consider the optimization problem associated with finding the prediction coefficient vector $\mathbf{a} \in \mathbb{R}^K$ from a set of observed real samples $x(n)$ for $n = 1, \ldots, N$ so that the prediction error is minimized [18]. Considering the speech production model for a segment of N speech samples $x(n)$, for $n = 1, \ldots, N$, in matrix form:

$$\mathbf{x} = \mathbf{X}\mathbf{a} + \mathbf{r}, \tag{2}$$

the problem becomes:

$$\mathbf{a} = \arg\min_{\mathbf{a}} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_p^p + \gamma \|\mathbf{a}\|_k^k, \tag{3}$$

where:

$$\mathbf{x} = \begin{bmatrix} x(N_1) \\ \vdots \\ x(N_2) \end{bmatrix}, \mathbf{X} = \begin{bmatrix} x(N_1-1) & \cdots & x(N_1-K) \\ \vdots & & \vdots \\ x(N_2-1) & \cdots & x(N_2-K) \end{bmatrix}. \tag{4}$$

The p-norm operator $\|\cdot\|_p$ is defined as $\|\mathbf{x}\|_p = (\sum_{n=1}^{N} |x(n)|^p)^{\frac{1}{p}}$. The starting and ending points N_1 and N_2 can be chosen in various ways by assuming $x(n) = 0$ for $n < 1$ and $n > N$. In this paper we will use the most common choice of $N_1 = 1$ and $N_2 = N+K$, which is equivalent, when $p = 2$ and $\gamma = 0$, to the *autocorrelation method* [25]. The introduction of the regularization term γ in (3) can be seen as being related to the prior knowledge of the coefficients vector $\mathbf{a}$, problem (3) then corresponds to the *maximum a posteriori* (MAP) approach for finding $\mathbf{a}$ under the assumptions that $\mathbf{a}$ has a Generalized Gaussian Distribution [26]. In finding a sparse signal representation, there is the somewhat subtle problem of how to measure sparsity. Sparsity is often measured as the cardinality, corresponding to the so-called 0-norm $\|\cdot\|_0$. Our optimization problem (3) would then become:

$$\mathbf{a} = \arg\min_{\mathbf{a}} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_0 + \gamma \|\mathbf{a}\|_0, \tag{5}$$

[2]The 0-norm is not technically a norm since it violates the triangle inequality.

with the particular case in which we are only considering the sparsity in the residual ($\gamma = 0$):

$$\mathbf{a} = \arg\min_{\mathbf{a}} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_0. \quad (6)$$

Unfortunately, these are combinatorial problems which generally cannot be solved in polynomial time. Instead of the cardinality measure, we will then use the more tractable 1-norm $\|\cdot\|_1$, which is known throughout the sparse recovery literature (see, e.g., [27]) to perform well as a linear programming relaxation of the 0-norm. We will also consider more recent variations of the 1-norm minimization criterion such as the reweighted 1-norm [28] to enhance the sparsity measure and moving the solution closer to the original 0-norm problem (5).

III. Sparse Linear Predictors

In this section, we will define the different sparse linear predictors and show their application in the context of speech processing. In particular, we will introduce the problem of determining a short-term predictor that engenders a sparse residual and the problem of finding a high order sparse predictor that also engenders a sparse residual. Since in Section II, we have introduced the 1-norm minimization as the sparsity measure, here we will also give a brief overview of the reweighted 1-norm algorithm to enhance this sparsity measure, moving closer to the original problem (0-norm minimization).

A. Finding a Sparse Residual

We consider the problem of finding a prediction coefficient vector $\mathbf{a}$ such that the resulting residual is sparse. Having identified the 1-norm as a suitable convex relaxation of the cardinality, the cost function for this problem is a particular case of (3). By setting $p = 1$ and $\gamma = 0$ we obtain the following optimization problem:

$$\min_{\mathbf{a}} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_1. \quad (7)$$

This formulation of the LP problem has been considered since the early works on speech analysis [9] [15] [16] and becomes particularly relevant for the analysis of voiced speech. In particular, compared to the traditional 2-norm minimization, the cost function associated with the 1-norm minimization deemphasize the impact of the spiky underlying excitation associated with voiced speech on the solution $\mathbf{a}$. Thus, there is an interesting connection between recovering a sparse residual vector and applying robust statistics methods to find the predictor [8]. An example of the more accurate recovery of the voiced excitation is shown in Figure 1. The effect of putting less emphasis on the outliers of the spiky excitation associated with voiced speech will reflect on the spectral envelope that will avoid the over-emphasis on peaks generated in the effort to cancel the pitch harmonics. An example of this property is shown in Figure 2.

While the 1-norm has been shown to outperform the 2-norm in finding a more proper LP model in speech analysis, in the case of unvoiced speech both approaches seem to provide appropriate models. However, by using the 1-norm minimization, we provide a residual that is sparser. In particular in [29] it is shown that, the residual vector provided by 1-norm minimization will have at least K components equal to zero.

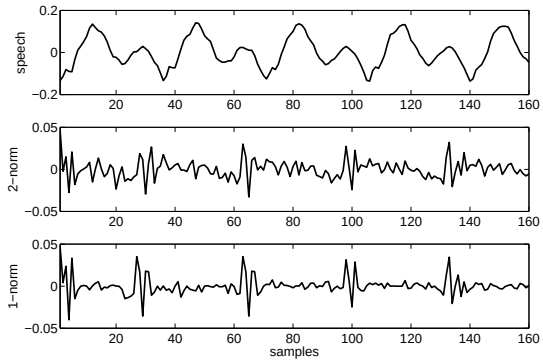


Fig. 1. An example of prediction residuals obtained by 2-norm and 1-norm error minimization. The speech segment analyzed is shown in the top box. The prediction order is $K = 10$ and the frame length is $N = 160$. It can be seen that the spiky pitch excitation is retrieved more accurately when 1-norm minimization is employed.

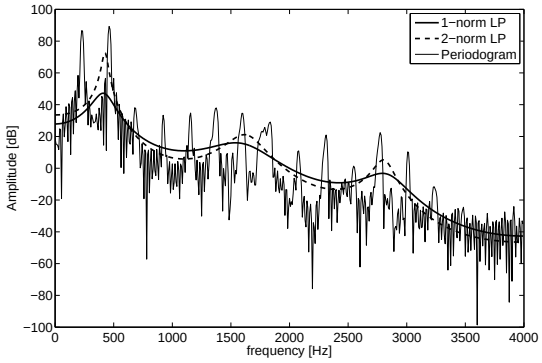


Fig. 2. An example of LP spectral model obtained by 1-norm and 2-norm error minimization for a segment of voiced speech. The prediction order is $K = 10$ and the frame length is $N = 160$. The lower emphasis on peaks in the envelope, when 1-norm minimization is employed, is a direct consequence of the ability to retrieve the spiky pitch excitation.

B. Finding a High Order Sparse Predictor

We now consider the problem of finding a high order sparse predictor that also engenders a sparse residual. This problem is particularly relevant when considering the usual modeling approach adopted in low bit-rate predictive coding for voiced speech segments. This corresponds to a cascade of a short-term linear predictor $F(z)$ and a long-term linear predictor $P(z)$ to remove respectively near-sample redundancies, due to the presence of formants, and distant-sample redundancies, due to the presence of a pitch excitation. The cascade of the predictors corresponds to the multiplication in the z-domain of the their transfer functions:

$$\begin{aligned} A(z) = F(z)P(z) &= 1 - \sum_{k=1}^{K} a_k z^{-k} \\ &= (1 - \sum_{k=1}^{N_f} f_k z^{-k})(1 - \sum_{k=1}^{N_p} g_k z^{-(T_p+k-1)}). \end{aligned} \quad (8)$$

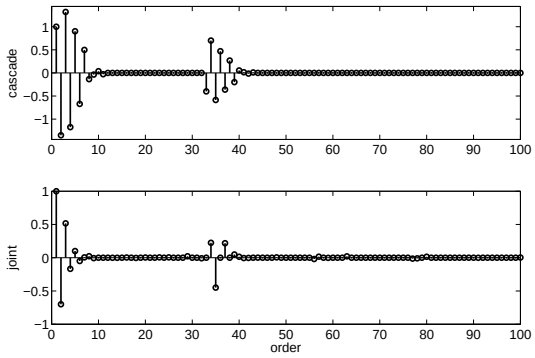


Fig. 3. An example of the high order predictor coefficient vector resulting from a cascade of long-term and short-term predictors (top box) and the solution of (9) for $\gamma = 0.1$ and order $K = 100$. The order is chosen sufficiently large to accommodate the filter cascade (8). It can be seen that the nonzero coefficient in the sparse prediction vector roughly coincide with the structure of the cascade of the two predictors.

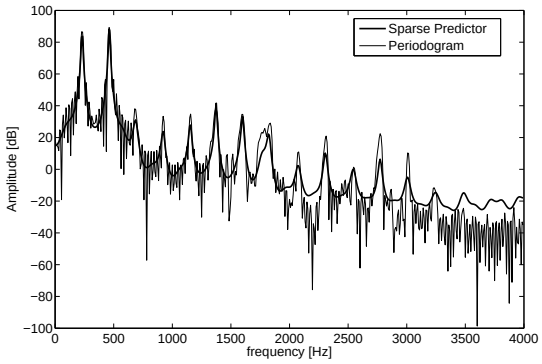


Fig. 4. Frequency response of the high order predictor of Figure 3. The order of the predictor is $K = 100$ and we consider only the nine nonzero coefficients of largest magnitude modeling the short-term and long-term predictors cascade.

The resulting prediction coefficient vector $\mathbf{a} = \{a_k\}$ of the high order polynomial $A(z)$ will therefore be highly sparse[3]. Taking this into account in our minimization process, and again considering the 1-norm as convex relaxation of the 0-norm, our original problem (5) becomes:

$$\min_{\mathbf{a}} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_1 + \gamma\|\mathbf{a}\|_1, \tag{9}$$

where the dimension of the prediction coefficient vector $\mathbf{a}$ (the order of the predictor) has to be sufficiently large to model the filter cascade ($K > N_f + T_p + N_p$) in (8). This approach, while maintaining resemblances to (7) looking for a sparse residual, is fundamentally different. While the predictor in (7) aims at modeling the spectral envelope, the purpose of the high order sparse predictor is to model the *whole* spectrum, i.e., the spectral envelope and the spectral harmonics. This can be easily achieved due to the strong ability of high order LP to resolve closely spaced sinusoids [30], [31]. Furthermore, considering the construction of the observation matrix $\mathbf{X}$, finding a high order sparse predictor is equivalent to identify which columns of $\mathbf{X}$, and in turn, which samples in $\mathbf{x}$ are important in the linear combination to predict a sample of speech (1). Thus, when a segment of voiced speech is analyzed with the predictive framework in (9), the nonzero coefficients roughly coincide with the structure in (8). An example of the predictor obtained as solution of (9) is shown in Figure 3. An example of the spectral modeling properties is shown in Figure 4.

There are mainly two problems associated with exploiting the modeling properties of the sparse high order predictor: determining an appropriate value of γ to solve (9) and using an approximate factorization to obtain again the initial formulation composed by the two predictors (8). Below we address these two issues.

1) Selection of γ: It is clear from (9) that γ controls *how sparse* the predictor should be and the trade-off between the sparsity of the predictor and the sparsity of the residual. In particular, by increasing γ, we increase the sparsity of the prediction coefficient vector, until all its entries are zero ($A(z) = 1$) for $\gamma \geq \|\mathbf{X}^T\mathbf{x}\|_\infty$ (where $\|\cdot\|_\infty$ denotes the dual norm to $\|\cdot\|_1$). More precisely, for $0 < \gamma < \|\mathbf{X}^T\mathbf{x}\|_\infty$, the solution vector $\mathbf{a}$ is a linear function of γ [32]. However, in general, the number of nonzero elements in $\mathbf{a}$ is not necessarily a monotonic function of γ.

There are obviously several ways of determining γ. In our previous work [21] [22], we have found the modified L-curve [33] as an efficient tool to find a balanced sparse representation between the two descriptions. The optimal value of γ (in the L-curve sense) is found as the point of maximum curvature of the curve $(\|\mathbf{x} - \mathbf{X}\mathbf{a}_\gamma\|_1, \|\mathbf{a}_\gamma\|_1)$. We have also observed that, in general, a constant value of γ, chosen for example as the average value of the set of γ's found with the L-curve based approach for a large set of speech frames, is an appropriate choice in the predictive problems considered. In the experimental analysis we will consider both approaches to defining γ.

2) Factorization of the high order polynomial: If γ is chosen appropriately, the considered formulation (9) results in a high order predictor $\hat{A}(z)$ with a clear structure that resembles the cascade of the short-term and long-term predictor (Figure 3). We can therefore bring $\hat{A}(z)$ to the original formulation in (8), by applying a simple and effective ad-hoc method to factorize the solution [22]. In particular, we use the first N_f coefficients of the high order predictor as the estimated coefficients of the short-term predictor:

$$\hat{F}(z) = 1 - \sum_{k=1}^{N_f} \hat{a}_k z^{-k}, \tag{10}$$

and then compute the quotient polynomial $\hat{Q}(z)$ of the division of $\hat{A}(z)$ by $\hat{F}(z)$ so that:

$$\hat{A}(z) = \hat{Q}(z)\hat{F}(z) + E(z) \approx \hat{Q}(z)\hat{F}(z), \tag{11}$$

where the deconvolution remainder $E(z)$ is considered to be negligible as most of the information of the coefficients has shown to be retained by $\hat{Q}(z)$ and $\hat{F}(z)$. From the polynomial $\hat{Q}(z)$ we can then extract the N_p taps predictor. In this paper,

[3]Traditionally, for speech sampled at 8 kHz, $N_f = 10$, $N_p = 1$, and T_p usually belongs in the range $[16, 120]$.

we will consider the most common pitch predictor where $N_p = 1$ ($P(z) = 1 - g_p z^{-T_p}$), then we merely identify the minimum value and its position in the coefficients vector of $\hat{Q}(z)$:

$$\begin{aligned} g_p &= \min\{q_k\}, \\ T_p &= \arg\min\{q_k\}. \end{aligned} \tag{12}$$

It is clear that, while heuristic, this factorization procedure is highly customizable. A different numbers of taps for both the short-term and long-term can be selected and also a voiced/unvoiced classification can be included, based on the presence or absence of long-term information, as described in [21], [22].

It should be noticed that the structure of the cascade can also be incorporated into the minimization scheme and can be potentially beneficial in reducing the size of the problem. This approach is then similar to the *One-Shot Combined Optimization* presented in [17] which is implicitly a sparse method looking for a similar high order factorisable predictor. The joint estimation in this case requires prior knowledge on the position of the pitch contributions (a pitch estimate) and the model order of both the short-term and long-term predictors. Differently from this method, in our approach, we obtain information on the model order of both short-term and long-term contribution and a pitch estimate, just by a simple post-processing the solution of (9).

C. Enhancing sparsity by reweighted 1-norm minimization

As shown throughout this section, the 1-norm is used as a convex relaxation of the 0-norm, because 0-norm minimization yields a combinatorial problem (NP-hard). We are therefore interested in adjusting the error weighting difference between the 1-norm and the 0-norm. A variety of recently introduced methods have dealt with reducing the error weighting difference between the 1-norm and the 0-norm by relying on the iterative reweighted 1-norm minimization (see, e.g., [34] and references therein). In particular, the iteratively reweighted 1-norm minimization may be used for estimating $\mathbf{a}$ and enhancing the sparsity of $\mathbf{r}$ (and $\mathbf{a}$), while keeping the problem solvable with convex tools [28] [23]. The predictor can then be seen as a solution of the following minimization problem:

$$\mathbf{a} = \arg\min_{\mathbf{a}} \lim_{p\to 0} \lim_{k\to 0} \{\|\mathbf{x} - \mathbf{X}\mathbf{a}\|_p^p + \gamma \|\mathbf{a}\|_k^k\}, \tag{13}$$

where each iteration of the reweighting process brings us closer to the 0-norm.

The mismatch between the 0-norm and the 1-norm minimization can be seen more clearly in Figure 5, where larger coefficients are penalized more heavily by the 1-norm than small ones. From an optimization point of view, when $p \leq 1$, the cost functions will have lower emphasis on large values and sharper slopes near zero compared to the $p = 1$ case. In turn, from a statistical point of view, the density functions will have heavier tails and a sharper slope near zero. This means that the minimization will encourage small values to become smaller while enhancing the amplitude of larger values. The limit case for $p = 0$ will have an infinitely sharp slope in zero and equally weighted tails. This will introduce as many zeros as possible as these are infinitely weighted. In this sense, the 0-norm can be seen as more "impartial" by penalizing every nonzero coefficient equally. It is clear that if a very small value would be weighted as much as a large value, the minimization process will try to eliminate the smaller ones and enhance the larger ones.

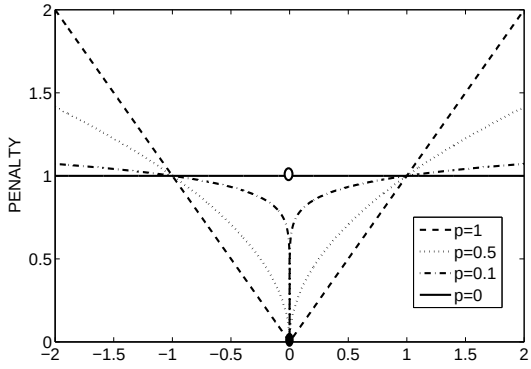


Fig. 5. Comparison between cost functions for $p \leq 1$. The 0-norm can be seen as more "democratic" than any other norm by weighting all the nonzero coefficients equally.

The algorithm to obtain a short-term predictor engendering a sparser residual, a reweighted formulation of (7), is shown in Algorithm 1. This approach, as we shall see, becomes beneficial in finding a predictor that produces a sparser residual, providing a tighter coupling between the prediction estimation and the search for the approximated sparse excitation. An example of the reweighted residual estimate is shown in Fig. 6.

When we impose sparsity both on the residual and on the high order predictor, as in (9), the algorithm is modified as shown in Algorithm 2. This formulation is relevant as it enhances the components that contain the information regarding the near-end and far-end redundancies in the high order predictor making the approximate factorization presented in III-B2 more accurate. In particular, the reweighting allows to reduce the spurious near-zero components in the high order predictor obtained (see Fig. 3) while enhancing the larger components that contain information of near-end and far-end redundancies.

It has been shown in [28] that $\|\hat{\mathbf{r}}^{i+1}\|_1 \leq \|\hat{\mathbf{r}}^{i}\|_1$, meaning that this is a descent algorithm. The halting criterion can therefore be chosen as either a maximum number of iterations or as a convergence criterion. In the experimental analysis we will give details on how many iterations are required in our setting. In both algorithms, the parameter $\epsilon > 0$ is used to provide stability when a component of $\hat{\mathbf{r}}$ goes to zero.

As a general remark, in [28] and [34], it is also shown that the reweighted 1-norm algorithm, at convergence, is equivalent to the minimization of the log-sum penalty function. This is relevant to what we are trying to achieve in (13): the log-sum cost function has a sharper slope near zero compared to the 1-norm, providing more effective sparsity inducing properties. Furthermore, since the log-sum is not convex, the iterative algorithm corresponds to minimizing a sequence of

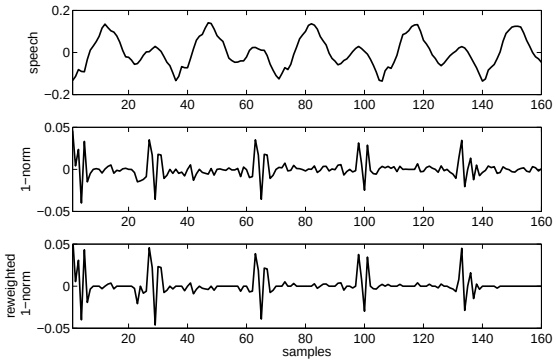


Fig. 6. An example of prediction residuals obtained through 1-norm and reweighted 1-norm error minimization using Algorithm 1. The speech segment analyzed is shown in the top box. The prediction order is $K = 10$ and the frame length is $N = 160$. Five iterations where made with $\epsilon = 0.01$.

Algorithm 1 Iteratively Reweighted 1-norm Minimization of the Residual

Inputs: speech segment $\mathbf{x}$
Outputs: predictor $\hat{\mathbf{a}}^i$, residual $\hat{\mathbf{r}}^i$
$i = 0$, initial weights $\mathbf{W}^{i=0} = \mathbf{I}$
while halting criterion false **do**
1. $\hat{\mathbf{a}}^i, \hat{\mathbf{r}}^i \leftarrow \arg\min_{\mathbf{a}} \|\mathbf{W}^i\mathbf{r}\|_1 \quad \text{s.t.} \quad \mathbf{r} = \mathbf{x} - \mathbf{X}\mathbf{a}$
2. $\mathbf{W}^{i+1} \leftarrow \text{diag}\left(\left|\hat{\mathbf{r}}^i\right| + \epsilon\right)^{-1}$
3. $i \leftarrow i + 1$
end while

linearizations of the log-sum around the previous solution estimate, providing at each step a sparser solution (until convergence).

IV. Compressed Sensing formulation for Sparse Linear Prediction

The CS formulation is particularly relevant in our sparse recovery problems: by exploiting prior knowledge about the sparsity of the signal $\mathbf{x}$ we will show that a limited number of random measures are sufficient to recover our predictors and sparse residual with high accuracy. In particular, it has been shown [24] [35] that a random projection of a high-dimensional but sparse or compressible signal vector onto a lower-dimensional space contains enough information to be able to reconstruct the signal with small or zero error. The random measures in CS literature are usually obtained by projecting the considered measurement vectors onto a lower dimensional space, using random matrices.

In recent work [36], [37], CS formulations in the context of speech analysis and coding have been formulated in order to find a sparse approximation of the residual, given the predictor. It is then interesting to extend this work to the case where we want to find directly the predictor that engenders intrinsically a sparse residual. In particular, given the sparsity level of the sparse representation that we wish to retrieve in a given domain, we can determine an efficient *shrinkage* of the minimization problem in a lower dimensional space, with a clear impact on the computational complexity.

Algorithm 2 Iteratively Reweighted 1-norm Minimization of Residual and Predictor

Inputs: speech segment $\mathbf{x}$
Outputs: predictor $\hat{\mathbf{a}}^i$, residual $\hat{\mathbf{r}}^i$
$i = 0$, initial weights $\mathbf{W}^{i=0} = \mathbf{I}$ and $\mathbf{D}^{i=0} = \mathbf{I}$
while halting criterion false **do**
1. $\hat{\mathbf{a}}^i, \hat{\mathbf{r}}^i \leftarrow \arg\min_{\mathbf{a}} \|\mathbf{W}^i\mathbf{r}\|_1 + \gamma\|\mathbf{D}^i\mathbf{a}\|_1$ s.t. $\mathbf{r} = \mathbf{x} - \mathbf{X}\mathbf{a}$
2. $\mathbf{W}^{i+1} \leftarrow \text{diag}\left(\left|\hat{\mathbf{r}}^i\right| + \epsilon\right)^{-1}$
3. $\mathbf{D}^{i+1} \leftarrow \text{diag}\left(\left|\hat{\mathbf{a}}^i\right| + \epsilon\right)^{-1}$
4. $i \leftarrow i + 1$
end while

If we wish to perform CS, two main ingredients are needed: a domain where the analyzed signal is sparse and the sparsity level of this signal T. In our case, the residual is the domain where the signal is sparse, while the linear transform that maps the original speech signal to the sparse residual is the sparse predictor. The sparsity in the residual domain is then imposed by our needs [35]. Let us now review the formulation presented in [37]:

$$\hat{\mathbf{r}} = \arg\min_{\mathbf{r}} \|\mathbf{r}\|_1 \quad \text{s.t.} \quad \mathbf{\Phi x} = \mathbf{\Phi H r} \tag{14}$$

where $\mathbf{x}$ is the $N \times 1$ analyzed segment of speech, $\mathbf{H}$ the $N \times (N+K)$ synthesis matrix, constructed from the truncated impulse response of the *known* predictor [38], $\mathbf{r}$ is the residual vector to be estimated (supposedly sparse) and $\mathbf{\Phi}$ is the sensing matrix of dimension $M \times N$. The dimensionality of the random linear projection M stems from the sparsity level T that one wishes to impose on the residual. In particular, based on empirical results, the number of projections is set equal to four times the sparsity, i.e. $M = 4T$. Furthermore, when the incoherence between the synthesis matrix and the random basis matrix $\mathbf{\Phi}$ holds ($\mu(\mathbf{\Phi}, \mathbf{H}) \approx 1$), even if $\mathbf{H}$ is not orthogonal the recovery of the sparse residual $\mathbf{r}$ is still possible and the linear program in (14) gives an accurate reconstruction of $\mathbf{x}$ with very high probability [24], [37]. As a general remark, the entries of the random matrix can be drawn from many different processes, in our case we will use a i.i.d. Gaussian process, as done in [36], [37].

To adapt CS principles to the estimation of the predictor as well, let us now consider the relation between the synthesis matrix $\mathbf{H}$ and the analysis matrix $\mathbf{A}$ where one is the pseudo-inverse of the other [39]:

$$\mathbf{A} = \mathbf{H}^{+}. \tag{15}$$

We can now replace the constraint $\mathbf{\Phi x} = \mathbf{\Phi H r}$ in (14) as

$$\mathbf{\Phi r} = \mathbf{\Phi A x}, \tag{16}$$

where $\mathbf{A}$ is the $(N+K) \times N$ analysis matrix that performs the whitening of the signal, constructed from the coefficients of the predictor $\mathbf{a}$ of order K [39], the dimension of the sensing matrix $\mathbf{\Phi}$ is now adjusted accordingly to $M \times (N+K)$. Notice that, due to the structure of $\mathbf{A}$ this can be rewritten equivalently to:

$$\mathbf{\Phi r} = \mathbf{\Phi A x} = \mathbf{\Phi}\left[\mathbf{x}|\mathbf{X}\right]\left[1, \mathbf{a}^T\right]^T, \tag{17}$$

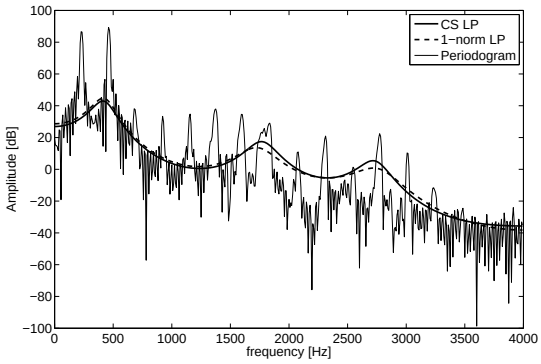


Fig. 7. An example of LP spectral model obtained through 1-norm minimization (7) and through CS based minimization (18) for a segment of voiced speech. The prediction order is $K = 10$ and the frame length is $N = 160$, for the CS formulation the dimension of the sensing matrix is $M = 80$, corresponding to the sparsity level $T = 20$.

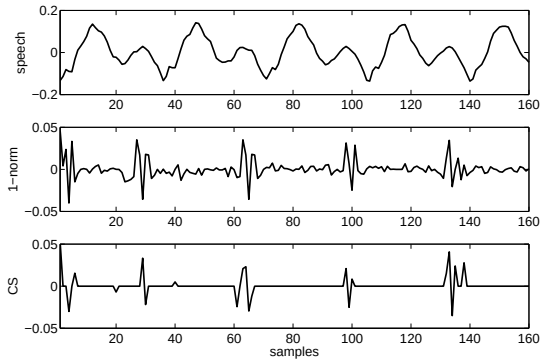


Fig. 8. An example of prediction residuals obtained through 1-norm minimization and CS recovery. The speech segment analyzed is shown in the top box. The prediction order is $K = 10$ and the frame length is $N = 160$. For the CS formulation, the imposed sparsity level is $T = 20$, corresponding to the size $M = 80$ for the sensing matrix.

where $[\mathbf{x}|\mathbf{X}]$ is the matrix obtained by stacking the vector $\mathbf{x}$ to the left of $\mathbf{X}$ in (4). The minimization problem can then be rewritten as:

$$\min_{\mathbf{a},\mathbf{r}} \|\mathbf{r}\|_1 \quad \text{s.t.} \quad \mathbf{\Phi r} = \mathbf{\Phi}(\mathbf{x} - \mathbf{Xa}). \tag{18}$$

We can now see that (18) is *equivalent* to (7), the only difference being the projection onto the random basis in the constraint. Therefore, (7) can be seen as a particular case of the formulation in (18) where $\mathbf{\Phi} = \mathbf{I}$ and $\mathbf{I}$ is a identity matrix of size $(N + K) \times (N + K)$. In this case we are then not actually performing a projection in a random subspace. The minimization constraint on the left side of (18) would become:

$$\mathbf{\Phi r} = \mathbf{\Phi}(\mathbf{x} - \mathbf{Xa}) \quad \stackrel{\mathbf{\Phi}=\mathbf{I}}{\longrightarrow} \quad \mathbf{r} = \mathbf{x} - \mathbf{Xa}. \tag{19}$$

The results obtained will then be similar to our initial formulation (7), as long as the choice of $\mathbf{\Phi}$ is appropriate. In this case, the formulation in (18) will not only provide hints on the T pulses to be selected in the residual, but also a dimensionality reduction that will simplify the calculations. This computational complexity reduction, resulting from the dimensionality reduction given by the projection onto random basis has been also observed in [40] and arises from the Johnson-Lindestrauss lemma [41]. An example of an envelope estimation using the formulation in (18) is presented in Figure 7 while the recovered sparse residual is shown in Figure 8.

Similarly, if we are looking for a high order sparse predictor, the problem (9) can be cast into a CS framework leading to:

$$\arg\min_{\mathbf{a},\mathbf{r}} \|\mathbf{r}\|_1 + \gamma\|\mathbf{a}\|_1 \quad \text{s.t.} \quad \mathbf{\Phi r} = \mathbf{\Phi}(\mathbf{x} - \mathbf{Xa}). \tag{20}$$

The formulation (9) and (20), similarly to (7) and (18), become equivalent when $\mathbf{\Phi} = \mathbf{I}$ and the minimization constraint is then (19). Both formulations (18) and (20), can also be modified to involve iterative reweighting (Algorithm 3 shows the general case for $\gamma > 0$). In [28] applications of the reweighted 1-norm minimization in a CS framework are provided.

Algorithm 3 CS Formulation of the Iteratively Reweighted 1-norm Minimization of Residual and Predictor

Inputs: speech segment $\mathbf{x}$, desired residual sparsity level T
Outputs: predictor $\hat{\mathbf{a}}^i$, residual $\hat{\mathbf{r}}^i$
$i = 0$, initial weights $\mathbf{W}^{i=0} = \mathbf{I}$ and $\mathbf{D}^{i=0} = \mathbf{I}$,
random matrix $\mathbf{\Phi}$ of size $M \times (N + K)$, $M = 4T$
while halting criterion false **do**
1. $\hat{\mathbf{a}}^i, \hat{\mathbf{r}}^i \leftarrow \arg\min_{\mathbf{a}} \|\mathbf{W}^i\mathbf{r}\|_1 + \gamma\|\mathbf{D}^i\mathbf{a}\|_1$ s.t. $\mathbf{\Phi r} = \mathbf{\Phi}(\mathbf{x} - \mathbf{Xa})$
2. $\mathbf{W}^{i+1} \leftarrow \text{diag}\left(\left|\hat{\mathbf{r}}^i\right| + \epsilon\right)^{-1}$
3. $\mathbf{D}^{i+1} \leftarrow \text{diag}\left(\left|\hat{\mathbf{a}}^i\right| + \epsilon\right)^{-1}$
4. $i \leftarrow i + 1$
end while

V. Properties of Sparse Linear Prediction

As mentioned in the introduction, many problems appearing in traditional 2-norm LP modeling of voiced speech can be traced back to the inability of the predictor to decouple the vocal tract transfer function from the pitch excitation. This results in a lower spectral modeling accuracy and a strong dependence on the placement of the analysis window. In this section we provide some experiments to illustrate how the sparse linear predictors presented in the previous sections manage to overcome these problems. As a general remark, it is well known that the $p-$norm LP estimate with $p \neq 2$ is not guaranteed to be stable [42]. Nevertheless, the results presented in this section concentrate on the spectral modeling properties of sparse LP, thus the stability of the predictor is simply imposed by pole reflection which stabilizes the filter without modifying the magnitude of the frequency response. We will provide a thorough discussion of the stability issues in the Section VII and in Section VI where the speech coding properties are analyzed and stability is critical.

The experimental analysis was done on 20,000 frames of length $N = 160$ (20 ms) of clean voiced speech coming from several different speakers with different characteristics (gender, age, pitch, regional accent) taken from the TIMIT database, downsampled at 8 kHz. The prediction methods we compare in this section are shown in Table I. The optimality of the methods **BE** and **RLP**, presented in [6], comes from the selection of the parameters which provided the lowest distortion compared with the reference envelope. For brevity

TABLE I
PREDICTION METHODS COMPARED IN THE MODELING PROPERTIES EVALUATION.

Method	Description
LP	Traditional 2-norm LP with 10Hz bandwidth expansion ($\gamma = 0.996$) and Hamming windowing.
SpLP10	1-norm LP presented in (III-A), solution of (7). Stability is imposed by pole reflection if unstable. No windowing is performed.
SpLP11	1-norm LP presented in (III-B). The order of (9) is $K = 110$ (covering accurately pitch delays in the interval $[N_f + 1, K - N_f - 1]$). γ is chosen as the point of maximum curvature in the L-curve. The short-term predictor coefficients are the first N_f coefficients of the high order polynomial. Stability is imposed by pole reflection if unstable. No windowing is performed.
BE	Optimally bandwidth expanded 2-norm LP as shown in [6]. Hamming window is used.
RLP	Optimally regularized 2-norm LP as shown in [6]. Hamming window is used.

and clarity of the presented results, we omitted the predictors obtained as solutions of the iterative reweighted algorithms presented in Section III-C and the CS formulation presented in Section IV. These methods, while presenting very similar modeling properties to **SpLP10** and **SpLP11**, produce predictors estimates with slightly higher variance, thus requiring few more bits to be encoded. Therefore, while it is hard to provide a fair comparison in terms of modeling, their properties become more interesting in the coding scenario that will thoroughly analyzed in Section VI; in particular, the differences in their bit allocation necessary for efficient coding and the information required in the residual will be analyzed.

A. Spectral Modeling

In this section, we provide results to the modeling properties of the short-term predictors. As a reference, we used the envelope obtained through a cubic spline interpolation between the harmonics peaks of the logarithmic periodogram. This method was presented in [6] and provided an approximation of the vocal tract transfer function, without the fine structure corresponding to the pitch excitation. We then calculated the log spectral distortion between our reference envelope $S_{int}(\omega)$ and the estimated predictive model $S(\omega, \mathbf{a})$ as:

$$\mathrm{SD}_m = \sqrt{\frac{1}{2\pi}\int_{-\pi}^{\pi} \left[10\log_{10} S_{int}(\omega) - 10\log_{10} S(\omega, \mathbf{a})\right]^2 d\omega}. \tag{21}$$

where the numerator gain is calculated as the variance of the residual.

The coefficients of the short-term predictors presented have also shown to be smoother and therefore they have a lower sensitivity to quantization. We also compared the log spectral distortion between our reference envelope $S_{int}(\omega)$ and the quantized predictive model $S(\omega, \hat{\mathbf{a}})$ for every predictor obtained with the presented methods. The quantizer used is the one presented in [43], with the number of bits fixed at 20 for the different prediction orders, providing in all the method presented a *transparent coding*[4]. The results are shown in Table II for different prediction orders. A critical analysis of the results showed the improved modeling properties of **SpLP11**. This was given by its ability to take into consideration the whole speech production model, thus decoupling more effectively the short-term contribution that provides the spectral envelope from the contribution given by the pitch excitation. **SpLP10** and **RLP** achieved similar performance, providing evidence supporting the generally good spectral modeling properties of the minimization problem in (7).

TABLE II
AVERAGE SPECTRAL DISTORTION FOR THE CONSIDERED METHODS IN THE UNQUANTIZED CASE (SD_m) AND QUANTIZED CASE (SD_q). A 95% CONFIDENCE INTERVAL IS GIVEN FOR EACH VALUE.

METHOD	K	SD_m	SD_q
LP	8	2.11±0.06	3.24±0.11
	10	1.97±0.03	2.95±0.09
	12	1.98±0.05	2.72±0.12
SpLP10	8	1.91±0.01	2.92±0.02
	10	1.78±0.01	2.53±0.02
	12	1.61±0.01	2.31±0.04
SpLP11	8	1.64±0.00	2.65±0.01
	10	1.69±0.00	2.37±0.01
	12	1.39±0.01	2.13±0.01
BE	8	2.04±0.03	3.11±0.08
	10	1.88±0.02	2.92±0.07
	12	1.83±0.10	2.71±0.04
RLP	8	1.89±0.02	2.93±0.04
	10	1.72±0.01	2.51±0.03
	12	1.53±0.02	2.22±0.04

B. Shift Invariance

In speech analysis, a desirable property for an estimator is to be invariant to the small shifts of the analysis window, since speech, and voiced speech in particular, is assumed to be short-term stationary. However, standard LP is well known not to be shift invariant [8]. This is a direct consequence of the coupling between the vocal tract transfer function and the underlying pitch excitation that standard LP introduces in the estimate. To analyze the invariance of the LP methods to window shifts, we took the same 20,000 frames of clean voiced speech and we expanded them to the left and to the right with 20 samples, giving a total length $N = 200$. In each frame of length $N = 200$ we defined a $M = 160$ samples boxcar window and we shifted the window by $s = 1, 2, 5, 10, 20$ samples. The average log spectral difference of the 10^{th} order AR estimate between $S_0(\omega)$ and $S_s(\omega)$ was analyzed. The average differences obtained for the methods in Table I are shown in Table III. In Figure 9, we show an example of the shift invariance property. The results obtained indicate clearly the sparse predictors robustness to small shifts in the analyzed window. While the decay in performance for increasing shift in the analysis window is comparable for all methods, the sparse predictors still retains better performance. Also in this

[4]According to [44], transparent coding of LP parameters is achieved when the two versions of coded speech, obtained using unquantized LP parameters and quantized LP parameters, are indistinguishable through listening. This is usually achieved with an average log distortion between quantized and unquantized spectra lower that 1 dB, with no outliers with log distortion greater than 4 dB and a number of outliers with 2-4 dB distortion lower than 2%.

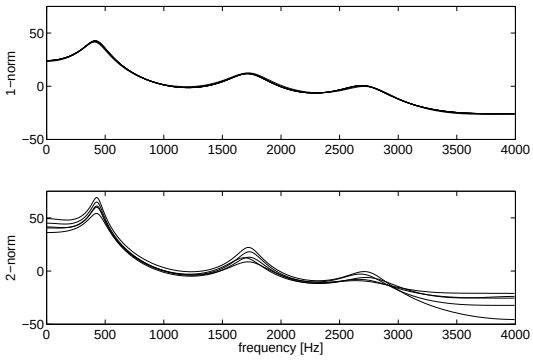


Fig. 9. An example of the shift invariance property of the sparse linear predictor (**SpLP11**) (top box), compared to traditional LP (**LP**). Ten envelopes are analyzed by shifting a the analysis window (160 samples) of $s = 1, 2, 5, 10, 20$ samples over a stationary voiced speech segment (length 200 samples).

TABLE III
AVERAGE SPECTRAL DISTORTION FOR THE CONSIDERED METHODS WITH SHIFT OF THE ANALYSIS WINDOW $s = 1, 2, 5, 10, 20$.

METHOD	SD_1	SD_2	SD_5	SD_{10}	SD_{20}
LP	0.113	0.128	0.223	0.452	1.262
SpLP10	0.003	0.003	0.011	0.017	0.032
SpLP11	0.001	0.002	0.005	0.006	0.009
BE	0.097	0.117	0.197	0.238	0.328
RLP	0.015	0.089	0.180	0.201	0.323

case, the change in the frequency response in traditional LP is clearly given by the pitch bias in the estimate of the predictor, particularly dependent on the location of the spikes of the pitch excitation. The approaches **SpLP11** and **SpLP10**, since they do not try to cancel this characteristic spiky excitation, are less dependent from its location and provide improved performance in modeling the true envelope.

C. Pitch Independence

The ability of the sparse linear predictors to decouple the pitch excitation from the vocal tract transfer function is reflected also in the ability to have estimates of the envelope that are not affected by the pitch excitation. In this experiment, we calculated the envelope using 10^{th} order regularized LP (**RLP**) and we modeled the underlying pitch excitation with an impulse train with different spacing. We then filtered this synthetic pitch excitation through the LP filter obtained and analyzed the synthetic speech applying the different LP methods in Table I. We divided the analysis into three subsets: high pitched $T_p \in [16, 35]$ ($f_0 \in [228\text{Hz}, 500\text{Hz}]$), mid pitched $T_p \in [36, 71]$ ($f_0 \in [113\text{Hz}, 222\text{Hz}]$) and low pitched $T_p \in [72, 120]$ ($f_0 \in [67\text{Hz}, 111\text{Hz}]$). The shortcomings of LP can be particularly seen in high pitched speech, as shown in the results of Table IV. Because high pitched speakers have fewer harmonics within a given frequency range, modeling of the spectral envelope is more difficult and particularly problematic for traditional LP. The sparse linear predictors are basically unaffected by the underlying pitch excitation, which results in an improved spectral modeling. In particular for **SpLP11**, since the high order structure of the initial estimate includes the pitch harmonic structure, the extracted short-term predictor is particularly robustly independent from the underlying excitation.

TABLE IV
AVERAGE SPECTRAL DISTORTION FOR THE CONSIDERED METHODS WITH DIFFERENT UNDERLYING PITCH EXCITATION. A 95% CONFIDENCE INTERVAL IS GIVEN FOR EACH VALUE.

METHOD	low	mid	high
LP	0.81±0.12	1.04±0.23	1.32±0.56
SpLP10	0.02±0.00	0.09±0.00	0.11±0.01
SpLP11	0.00±0.00	0.00±0.00	0.01±0.00
BE	0.45±0.07	0.65±0.19	0.89±0.34
RLP	0.05±0.02	0.16±0.10	0.19±0.09

VI. CODING APPLICATIONS OF SPARSE LINEAR PREDICTION

By introducing sparsity in the residual, we can reasonably assume that only a small portion of the residual samples are sufficient to reconstruct the speech signal with high accuracy. We will corroborate our intuition by providing some experiments on the coding applications of sparse linear prediction. Specifically, in VI-A, we will first give experimental proof of the sparsity inducing effectiveness of the short-term predictors in the Analysis-by-Synthesis (AbS) scheme [38]. In this case, we used a very simple excitation model coding without long-term prediction where we exploit directly the information on the location of the nonzero samples. In VI-B, we will present a simple coding procedure that exploits the properties of the combined high order sparse LP and sparse residual. As we shall see in VI-C, this approach presents interesting properties such as noise robustness for which we give both objective and subjective evaluation.

As a general remark, since the stability of the short-term predictors is not assured, we consistently performed a stability check and, if the short-term predictor was found to be unstable, we performed a pole reflection. Note that this approach necessarily modifies the time domain behavior of the residual as well as the predictor coefficients. Nevertheless, since the rate of unstable filters is low and the instability is very mild (i.e., the magnitude of the poles is only very slightly higher than one), this can be considered as an adequate solution to this problem. We will return to the stability issue in Section VII.

A. Coding Properties of the Short-Term Sparse Linear Predictor

The first experiment regards the use of the short-term predictor in speech coding. In particular we compared the use of the multipulse encoding procedure in the case of bandwidth expanded linear prediction (**LP**) with a fixed bandwidth expansion of 60 Hz (done by lag-windowing the autocorrelation function). We compared this approach with our introduced sparse linear predictors. The only difference is that, instead of performing the multipulse encoding, we performed the AbS procedure straight after selecting the T positions of the T largest samples that are located in the residual. In this experiment, we did not perform long-term prediction, focusing only on the coding properties of the sparsity inducing short-term predictors.

TABLE V
PREDICTION METHODS COMPARED IN THE CODING PROPERTIES EVALUATION.

Method	Description
LP	Traditional 2-norm LP with a fixed bandwidth expansion of 60 Hz (done by lag-windowing the autocorrelation function) and Hamming windowing.
SpLP10	1-norm LP solution of (7).
RWLP10	Reweighted 1-norm LP presented in Section III-C using Algorithm 1. Four reweighting iterations are performed (sufficient for convergence).
CSLP10	Compressed sensing formulation presented in Section IV, solution of (18). The size of the sensing matrix is given by the number of samples we want to retrieve in the residual.
RWCSLP10	Reweighed compressed sensing formulation of **CSLP10** using Algorithm 1. Four reweighting iterations are performed (sufficient for convergence).

TABLE VI
COMPARISON BETWEEN THE SPARSE PREDICTOR ESTIMATION METHODS. A 95% CONFIDENCE INTERVAL IS GIVEN FOR EACH VALUE.

METHOD	T	$\hat{\mathbf{a}}$	SSNR	MOS	t
LP	5	19	14.1±3.2	2.85±0.23	0.1±0.1
	10	19	19.1±2.9	3.01±0.16	0.9±0.3
SpLP10	5	18	15.3±2.1	2.87±0.12	1.3±0.2
	10	18	20.1±1.7	3.11±0.11	1.3±0.2
RWLP10	5	22	17.2±1.6	3.01±0.06	4.1±0.3
	10	22	21.4±1.5	3.19±0.03	4.1±0.3
CSLP10	5	19	16.9±1.9	2.97±0.04	0.4±0.0
	10	19	20.9±1.5	3.25±0.03	0.6±0.2
RWCSLP10	5	24	20.2±0.9	3.15±0.03	1.3±0.3
	10	24	24.4±0.4	3.43±0.01	1.9±0.2

We considered the formulation **SpLP10**, reweighted 1-norm **RWLP10**, and their CS formulations **CSLP10** and **RWCSLP10**. The methods compared are summarized in Table V. As mentioned in Section V, all these methods achieve similar modeling performance to **SpLP10**, although their estimate of the predictor requires a slightly larger number of bits. Here we will show this providing a comparison also in terms of bits needed for transparent quantization of the predictor. The methods **BE** and **RLP**, presented in the previous section (Table I) while offering better modeling properties than traditional LP, do not provide any significant improvement in the coding scenario, thus they will be omitted from the current experimental analysis.

We have performed the analysis on the same speech signals database considered in Section V. The frame size is $N = 40$, the 10^{th} order predictors were quantized transparently using the LSFs coding method in [43] while the T pulses are left unquantized. In the CS formulations the sensing matrix has $M = 4T$ rows; this means that just a slight reduction in the size of the problem was obtained when $T = 10$. Nevertheless we were able to obtain important information on the location of the pulses. In the reweighted schemes, the number of iterations is four, which was sufficient to reach convergence in all the analyzed frames.

In Table VI, we present the results in terms of Segmental SNR, Mean Opinion Score (obtained through PESQ evaluation) and empirical computational time t in elapsed CPU seconds for $T = 5$ and $T = 10$, and number of bits necessary to transparently encode the predictor ($\hat{\mathbf{a}}$) using LSFs [43]. The results demonstrate the effectiveness of the sparse linear predictors. These results also show that the predictors in the reweighted cases (**RWLP10** and **RWCSLP10**), need a larger number of bits for transparent quantization due to the larger variance of their estimates. This result is particularly interesting when considering the model in (2). In particular, the description of a segment of speech is distributed between its predictive model and the corresponding excitation. Thus, we can observe that the complexity of the predictor necessarily increases when the complexity of the residual decreases (less significant pulses). This also leaves open questions on the *optimal* bit distribution between the two descriptions. As a proof of concepts, the results show how only 5 bits of difference between **LP** and **RWCSLP10** in the representation of the filter result in a significant improvement in performance: only 5 pulses in the residual are necessary in **RWCSLP10** to obtain similar performance to **LP** using 10 pulses.

A critical analysis of the results leads to another interesting conclusion. In fact, while 1-norm based minimization, with or without the *shrinkage* of the problem provided by the CS formulation in (18), is computationally more costly, than 2-norm minimization, it greatly simplifies the next stage where the excitation is selected in a closed-loop AbS scheme. In particular, the empirical computational time in Table VI refers to both the LP analysis stage and the search for the MPE excitation. Since the MPE search for the location is not performed in our sparse LP methods and we exploit directly the information regarding the T pulses of largest magnitude, the AbS procedure is merely a small least square problem where we find the T pulse amplitudes. We will come back to the discussion regarding complexity in VII-B. Furthermore, it should be noted that the CS formulation improves the selection of the T largest pulses. This is remarkable since while the predictor obtained with or without the random projection is similar, the reduction of the constraints helps us find a more specific solution for the level of sparsity T that we would like to retrieve in the residual. As mentioned above, the price to pay is a slightly higher bit allocation for the predictors obtained through CS formulation.

B. Speech coding based on Sparse Linear Prediction

As a proof of concepts, we will now present a very simple coding scheme that summarizes all the previously introduced methods. We will use the method presented in Section III-B, exploiting the sparse characteristics of the high order predictor and the sparse residual. In order to reduce the number of constraints, we cast the problem in a CS formulation (20) that provides a shrinkage of the constraint according to the number of samples we wish to retrieve in the residual. Furthermore, in order to refine the initial sparse solution, we apply the reweighting algorithm. The core scheme is summarized in Algorithm 3. Differently from multistage coders, this method, with its joint estimation of a short-term and a long-term predictor and the presence of a sparse residual, provides a one-step approach to speech coding. In synthesis, given a segment of speech, a way to encode the speech signal can be as follows:

1) Define the desired level of sparsity of the residual T

TABLE VII
COMPARISON BETWEEN THE CODING PROPERTIES OF THE **AMR102** AND THE CODER BASED ON SPARSE LINEAR PREDICTION **SPLP**. A 95% CONFIDENCE INTERVAL IS GIVEN FOR EACH VALUE.

METHOD	rate	MOS	t
AMR102	10.2 kbps	4.02±0.11	0.1±0.0
SpLP	10.1 kbps	4.13±0.13	1.2±0.1

TABLE VIII
PERFORMANCES OF **AMR102** AND THE CODER BASED ON SPARSE LINEAR PREDICTION (**SPLP**) FOR DIFFERENT VALUES OF SNR (WHITE GAUSSIAN NOISE). A 95% CONFIDENCE INTERVAL IS GIVEN FOR EACH VALUE.

METHOD	clean	30dB	20dB	10dB
AMR102	4.02±0.11	3.88±0.21	3.25±0.19	2.76±0.23
SpLP	4.13±0.13	3.94±0.15	3.52±0.14	3.21±0.19

and define the sensing matrix dimensionality accordingly $M = 4T$.
2) Perform n steps of the CS reweighted minimization process (Algorithm 3).
3) Factorize the prediction coefficients into a short-term and long-term predictor using the procedure in III-B2.
4) Quantize short-term and long-term predictors.
5) Select the T positions where the values of largest magnitude are located.
6) Solve the analysis-by-synthesis equation keeping only the T nonzero positions.
7) Quantize the residual.

We have again analyzed about one hour of clean speech taken from the TIMIT database. In order to obtain comparable results, the frame length is now $N = 160$ (20 ms). The order of the high order predictor in (20) is $K = 110$ (meaning that we can cover accurately pitch delays in the interval $[N_f + 1, K - N_f - 1]$, including the usual range for the pitch frequency $[70\text{Hz}, 500\text{Hz}]$). the fixed regularization parameter is $\gamma = 0.12$ and the defined level of sparsity is $T = 20$. Four iterations of the reweighting minimization process are performed, sufficient to reach convergence in all the analyzed frames. The orders of the short-term and long-term predictors obtained from the factorization of the high order predictor are $N_f = 10$ and $N_p = 1$, respectively. 25 bits are used to transparently encode the LSF vector, 7 bits are used to quantize the pitch period T_p and 6 bits to quantize the pitch gain g_p. The stability of the overall cascade is imposed by pole reflection on the short-term predictor, and by limiting the pitch gain to be less that unity. As for the residual, the quantizer normalization factor is logarithmically encoded with 6 bits while a 8 levels uniform quantizer is used to quantize the normalized amplitudes; the signs are coded with 1 bit per each pulse. The upper bound given by the information content of the pulse location ($\log_2 \binom{160}{20}$ bits) is used as an estimate of the number of bits used for distortionless encoding of the location. No perceptual weighting is performed in our case. The total number of bits per frame used are 202, producing a 10.1 kbps rate. We will compare this method (**SpLP**) with the AMR coder in the 10.2 kbps mode (**AMR102**) [45]. The results in terms of MOS (obtained through PESQ evaluation) and empirical computation time are shown in Table VII and demonstrate similar performance but with a more straightforward approach to coding than AMR. The CS formulation also helps to generally keep the problem solvable in reasonable time.

C. Noise Robustness

This study is motivated by the ability of a sparse coder to identify more effectively the features of the residual signal that are important for its reconstruction, discarding those which probably are a result of the noise. The traditional encoding formulation, based on minimum variance analysis and residual encoding through pseudo-random sequences (i.e., algebraic codes), makes the identification of these important features basically impossible and requires, for low SNRs, noise reduction in the preprocessing. Interestingly enough, sparse LP based coding appears to be quite robust in the presence of noise. An example of the different performance in terms of MOS for different SNR under additive white Gaussian noise is given in table VIII.

D. Subjective assessment of speech quality

To further investigate the properties of our methods, we have conducted two MUSHRA listening tests [46] with 16 non-expert listeners. Ten speech clips were used in the listening test. In the first MUSHRA test we investigate what we have shown in VI-B, about the similarity in quality between the AMR coder and our method. In the second MUSHRA test the noise robustness of our method, discussed in VI-C, is proved. The test results are presented in Figure 10 where the score 100 corresponds to "Imperceptible" and the score 0 corresponds to "Very annoying" according to the 6-grade impairment scale. From the results, we can see that our method does not affect greatly the quality of the signal, given that our method is conceptually simpler and substantially less optimized compared to AMR. For example, we are not taking into account some of the main psychoacoustic criteria usually implemented in the AMR, such as the adaptive postfilter to enhance the perceptual quality of the reconstructed speech and the perceptual weighting filter employed in the analysis by synthesis search of the codebooks. Neverthless, in clean condition the average score was 89 for **AMR102**, and 82 for **SpLP**. The most significant results though, are the one related to the coding of noisy signals. In particular, we can see from Figure 10 that our method scores considerably better than the AMR showing how a sparse encoding technique can be more effective in noise robust speech coding. In fact, in noisy conditions, the average score was 62 for **AMR102**, and 75 for **SpLP**.

VII. DISCUSSION

A. Stability

In the presented applications of sparse linear predictors, the percentage of unstable filters was found to be low (around 2%) and the instability "mild[5]." This suggested the use of a simple stability check and pole reflection in our experimental analysis.

[5]The maximum absolute value for a root found in all our considered predictors is $\rho_{max} = 1.0259$.

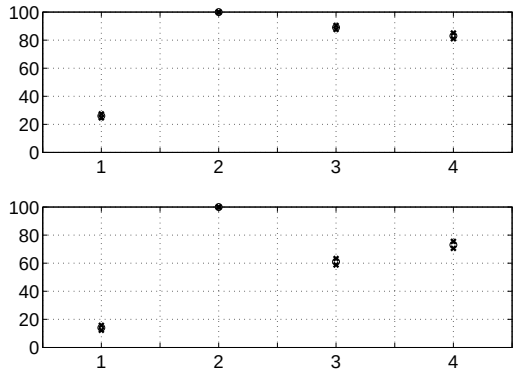


Fig. 10. MUSHRA test results. In the box above we show the results for clean speech and in the box below for speech corrupted by white noise (SNR=10dB). The four versions of the clips appear in the following order: Anchor, Hidden reference, **AMR102**, and **SpLP**. The anchor is the NATO standard 2400 bps LPC coding [47]. A 95% confidence interval is given for each value (upper and lower star).

Theorems exist to determine the maximum absolute value of the roots of a monic polynomial given the norm operator used in the minimization [42] but the bounds are generally too high to gain any real insight on how to create a intrinsic stable minimization problem, as done in [48].

The stability problem in (7) was already tackled in [9] by introducing the Burg method for prediction parameters estimation based on the least absolute forward-backward error. In this approach, however, the sparsity is not preserved. This is mostly due to the decoupling of the main K-dimensional minimization problem in K one-dimensional minimization subproblems. Therefore this method is suboptimal and produces results, as we have observed, somewhere in between those of the 2-norm and 1-norm approach. Also, the approach is only valid in (7) and not in all the other minimization schemes presented.

B. Computational Cost

As for the computational cost, finding the solution of the overdetermined system of equations in (7) using a modern interior point algorithm [19] can be shown to be equivalent to solving around 20-30 least square problems. Nevertheless, implementing this procedure in an AbS coder, as done in Section VI-A, is shown to greatly simplify the search for the sparse approximation of the residual in a closed-loop configuration, without compromising the overall quality. Furthermore, in the case of (9), the advantage is that a one step approach is taken to calculate both the short-term and the long-term predictors while the encoding of the residual is facilitated by its sparse characteristics.

The introduction of a compressed sensing formulation for the prediction problem has helped reduce dramatically the computational costs. An example of this can be seen in the coding scheme presented in VI-B. Retrieving $T = 20$ samples reduces the number of constraints of the minimization problem from 270 ($N + K$) to 80 ($M = 4T$). Since for each constraint we have a dual variable, by reducing the number of the constraints we also reduce the number of the dual variables [18]. In turn, the whole coding scheme, as shown empirically, is only about one order of magnitude more expensive than a 2-norm LP based coder, although with added improvements such as noise robustness and a fairly high conceptual simplicity.

C. Uniqueness

The minimization problems considered do not necessarily have a unique solution. In these rare cases with multiple solutions, due to the convexity of the cost function, we can immediately state that all the possible multiple solutions will still be optimal [18]. Viewing the non-uniqueness of the solution as a weakness is also arguable: in the set of possible optimal solutions we can probably find one solution that offers better properties for our modeling or coding purposes. A theorem to verify uniqueness is discussed in [50].

D. Frequency Domain Interpretation

The standard linear prediction method exhibits spectral matching properties in the frequency domain due to Parseval's theorem [2]:

$$\sum_{n=-\infty}^{\infty} |e(n)|^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} |E(e^{j\omega})|^2 d\omega. \tag{22}$$

It is also interesting to note that minimizing the squared error in the time domain and in the frequency domain leads to the same set of equations, namely the Yule-Walker equations [25]. To the best of our knowledge, the only relation existing between the time and frequency domain error using the 1-norm is the trivial Hausdorff-Young inequality [51]:

$$\sum_{n=-\infty}^{\infty} |e(n)| < \frac{1}{2\pi} \int_{-\pi}^{\pi} |E(e^{j\omega})| d\omega, \tag{23}$$

which implies that time domain minimization does not corresponds to frequency domain minimization. It is therefore difficult to say if the 1-norm based approach is always advantageous compared to the 2-norm based approach for spectral modeling, since the statistical character of the frequency errors is not clear. However, the numerical results in Tables II, III and IV clearly show better spectral modeling properties of the sparse formulation.

VIII. Conclusions

In this paper, we have given an overview of several linear predictors for speech analysis and coding obtained by introducing sparsity into the linear prediction framework. In speech analysis, the sparse linear predictors have been shown to provide a more efficient decoupling between the pitch harmonics and the spectral envelope. This translates into predictors that are not corrupted by the fine structure of the pitch excitation and offer interesting properties such as shift invariance and pitch invariance. In the context of speech coding, the sparsity of residual and of the high order predictor provides a more synergistic new approach to encode a speech segment. The sparse residual obtained allows a more compact representation, while the sparse high order predictor engenders

joint estimation of short-term and long-term predictors. A compressed sensing formulation is used to reduce the size of the minimization problem, and hence to keep the computational costs reasonable. The sparse linear prediction based robust encoding technique provided a competitive approach to speech coding with a synergistic multistage approach and a slower decaying quality for decreasing SNR.

ACKNOWLEDGMENT

The authors would like to thank Dr. Tobias Lindstrøm Jensen (Aalborg University), Dr. Shaminda Subasingha (University of Miami) and L. Anders Ekman (Royal Institute of Technology, Stockholm) for providing part of the code used in the evaluation procedures as well as useful suggestions.

REFERENCES

[1] J. H. L. Hansen, J. G. Proakis, and J. R. Deller, Jr., *Discrete-Time Processing of Speech Signals*, Prentice-Hall, 1987.

[2] J. Makhoul, "Linear Prediction: A Tutorial Review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, April 1975.

[3] F. Itakura and S. Saito, "Analysis synthesis telephony based on the maximum likelihood method,"*Rep. 6th Int. Congr. Acoustics* pp. C17–C20, Paper C-5-5, 1968.

[4] A. El-Jaroudi and J. Makhoul, "Discrete All-Pole Modeling," *IEEE Trans. Signal Processing*, vol. 39, pp. 411–423, 1991.

[5] M. N. Murthi and B. D. Rao, "All-pole modeling of speech based on the minimum variance distortionless response spectrum," *IEEE Trans. Speech and Audio Processing*, vol. 8, pp. 221–239, 2000.

[6] L. A. Ekman, W. B. Kleijn and M. N. Murthi, "Regularized Linear Prediction of Speech," *IEEE Trans. Audio, Speech, Language Processing*, vol. 16, no. 1, pp. 65–73, 2008.

[7] H. Hermansky, H. Fujisaki, Y. Sato, "Spectral envelope sampling and interpolation in linear predictive analysis of speech," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 9, pp. 53–56, 1984.

[8] C.-H. Lee, "On Robust Linear Prediction of Speech," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. 36, No. 5, pp. 642–650, 1988.

[9] E. Denoël and J.-P. Solvay, "Linear prediction of speech with a least absolute error criterion," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 33, no. 6, pp. 1397–1403, 1985.

[10] J. Schroeder, R. Yarlagadda, "Linear predictive spectral estimation via the L_1 norm," *Signal Processing*, Vol. 17, No. 1, pp. 19–29, 1989.

[11] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, Kluwer Academic Publishers, 1993.

[12] B. S. Atal and J. R. Remde, "A new model of LPC excitation for producing natural sounding speech at low bit rates," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 7, pp. 614–617, 1982.

[13] P. Kroon, E. D. F. Deprettere, and R. J. Sluyter, "Regular-pulse excitation - a novel approach to effective multipulse coding of speech", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34, pp. 1054–1063, 1986.

[14] W. C. Chu, *Speech Coding Algorithms: Foundation and Evolution of Standardized Coders*, Wiley, 2003

[15] J. Lansford and R. Yarlagadda, "Adaptive L_p approach to speech coding," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, pp 335–338, 1988.

[16] M. N. Murthi and B. D. Rao, "Towards a synergistic multistage speech coder," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, pp 369–372, 1998.

[17] P. Kabal and R. P. Ramachandran, "Joint Optimization of Linear Predictors in Speech Coders," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37(5), pp. 642–650,May 1989.

[18] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[19] S. J. Wright, *Primal-Dual Interior-Point Methods*, SIAM, 1997.

[20] 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.

[21] D. Giacobello, M. G. Christensen, J. Dahl, S. H. Jensen, and M. Moonen, "Joint estimation of short-term and long-term predictors in speech coders," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 4109–4112, 2009.

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

[23] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen, "Enhancing sparsity in linear prediction of speech by iteratively reweighted 1-norm minimization," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2010.

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

[25] P. Stoica and R. Moses, *Spectral Analysis of Signals*, Pearson Prentice Hall, 2005.

[26] S. Nadarajah, "A generalized normal distribution," *Journal of Applied Statistics*, vol. 32, no. 7, pp. 685–694, 2005.

[27] D. L. Donoho and M. Elad, "Optimally sparse representation from overcomplete dictionaries via ℓ^1-norm minimization," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 100, no. 5, pp. 2197–2202, 2002.

[28] E. J. Candés, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5, pp. 877–905, 2008.

[29] J. A. Cadzow, "Minimum ℓ_1, ℓ_1, and ℓ_∞ norm approximate solutions to an overdetermined system of linear equations," *Digital Signal Processing*, vol. 12, no. 4, pp. 524–560, 2002.

[30] P. Stoica and T. Söderström, "High Order Yule-Walker equations for estimating sinusoidal frequencies: the complete set of solutions," *Signal Processing*, vol. 20, pp. 257–263, 1990.

[31] D. Giacobello, T. van Waterschoot, M. G. Christensen, S. H. Jensen, M. Moonen, "High-Order Sparse Linear Predictors for Audio Processing," in *Proc. European Signal Proc. Conf.*, pp. 234–238, 2010.

[32] J. J. Fuchs, "On sparse representations in arbitrary redundant bases," in *IEEE Trans. on Inf. Theory*, vol. 50, no. 6, 2004.

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

[34] D. Wipf, S. Nagarajan, "Iterative reweighted ℓ_1 and ℓ_2 methods for finding sparse solutions," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 2, pp. 317–329, 2010.

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

[36] T. V. Sreenivas and W. B. Kleijn, "Compressive sensing for sparsely excited speech signals," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 4125–4128, 2009.

[37] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen, "Retrieving sparse patterns using a compressed sensing framework: applications to speech coding based on sparse linear prediction," *IEEE Sig. Proc. Letters*, vol. 17, no. 1, pp. 103–106, 2010.

[38] P. Kroon and W. B. Kleijn, "Linear-prediction based analysis-by-synthesis coding", in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal Eds., Elsevier Science B.V., ch. 3, pp. 79–119, 1995.

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

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

[41] W. B. Johnson and J. Lindenstrauss, "Extensions of Lipschitz mapping into Hilbert space," *Conf. in modern analysis and probability*, vol. 26, pp. 189–206, 1984.

[42] L. Knockaert, "Stability of linear predictors and numerical range of shift operators in normed spaces," *IEEE Trans. on Inf. Theory*, vol. 38, no. 5, pp. 1483–1486, 1992.

[43] A. D. Subramaniam, B. D. Rao, "PDF optimized parametric vector quantization of speech line spectral frequencies", *IEEE Trans. Speech and Audio Processing*, vol. 11, no. 2, 2003.

[44] K. K. Paliwal, B. S. Atal, "Efficient vector quantization of LPC parameters at 24 bits/frame," *IEEE Trans. Speech and Audio Processing*, vol. 1, no. 1, pp 3–14, 1993.

[45] "Adaptive Multi-Rate (AMR) speech codec; Transcoding functions," 3GPP TS 26.190, 2004.

[46] Method for the Subjective Assessment of Intermediate Quality Level of Coding Systems, ITU-R BS.1534-1 2003.

[47] NATO (unclassified), "Parameters and coding characteristics that must be common to assure interoperability of 2400 bps linear predictive encoded digital speech," Annex X to AC/302 (NBDS) R/2.

[48] C. Magi, J. Pohjalainen, T. Bäckström and P. Alku, "Stabilised weighted linear prediction," *Speech Communication*, vol. 51, no. 5, pp. 401–411, 2009.

[49] W. F. G. Mecklenbrauker, "Remarks on the minimum phase property of optimal prediction error filters and some related questions," *IEEE Sig. Proc. Letters*, vol. 5, no. 4, pp. 87–88, 1998.

[50] P. Bloomfield and W. Steiger, “Least absolute deviations curve-fitting”, *SIAM J. on Scientific and Statistical Computing*, vol. 1, no. 2, pp. 290–301, 1980.
[51] M. Reed and B. Simon, *Methods of Modern Mathematical Physics II: Fourier Analysis, Self-adjointness*, Academic Press, 1975.