Fast Algorithms for High-Order Sparse Linear Prediction with Applications to Speech Processing

Jensen, Tobias Lindstrøm; Giacobello, Daniele; van Waterschoot, Toon; Christensen, Mads Græsbøll

Published in:
Speech Communication

DOI (link to publication from Publisher):
10.1016/j.specom.2015.09.013

Publication date:
2016

Document Version
Accepted author manuscript, peer reviewed version

Fast Algorithms for High-Order Sparse Linear Prediction with Applications to Speech Processing

Tobias Lindstrøm Jensen[a,1,*], Daniele Giacobello[a], Toon van Waterschoot[b], Mads Græsbøll Christensen[c]

[a] *Signal and Information Processing, Department of Electronic Systems, Aalborg University, Denmark*
[b] *Center for Dynamical Systems Signal Processing and Data Analytics (STADIUS), Department of Electrical Engineering (ESAT), KU Leuven, Belgium*
[c] *Audio Analysis Lab, AD:MT, Aalborg University, Denmark*

Abstract

In speech processing applications, imposing sparsity constraints on high-order linear prediction coefficients and prediction residuals has proven successful in overcoming some of the limitation of conventional linear predictive modeling. However, this modeling scheme, named sparse linear prediction, is generally formulated as a linear programming problem that comes at the expenses of a much higher computational burden compared to the conventional approach. In this paper, we propose to solve the optimization problem by combining splitting methods with two approaches: the Douglas-Rachford method and the alternating direction method of multipliers. These methods allow to obtain solutions with a higher computational efficiency, orders of magnitude faster than with general purpose software based on interior-point methods. Furthermore, computational savings are achieved by solving the sparse linear prediction problem with lower accuracy than in previous work. In the experimental analysis, we clearly show that a solution with lower accuracy can achieve approximately the same performance as a high accuracy solution both objectively, in terms of prediction gain, as well as with perceptually relevant measures, when evaluated in a speech reconstruction application.

Keywords: sparse linear prediction, speech and audio processing, spectral

*Corresponding author

Email addresses: `tlj@es.aau.dk` (Tobias Lindstrøm Jensen), `giacobello@ieee.org` (Daniele Giacobello), `toon.vanwaterschoot@esat.kuleuven.be` (Toon van Waterschoot), `mgc@create.aau.dk` (Mads Græsbøll Christensen)

[1] The work of T.L. Jensen is supported by the Danish Council for Independent Research, Technology and Production Sciences. Grant no. 4005-00122.

modeling, convex optimization, linear programming, real-time optimization, speech reconstruction, packet loss concealment.

---

1. Introduction

Linear prediction (LP) is a well understood technique for the analysis, modeling, and coding of speech signals [1]. The widespread use of LP of speech can be attributed to its correspondence to the source-filter model of speech production [2, 3]. An emitted speech sound can be modeled as a combination of the excitation process (the air flow) and the filtering process (vocal tract effect). The vocal tract can, to a large extent, be modeled as a slow varying low-order all-pole filter, while the air flow can be modeled by a white noise sequence, for unvoiced sounds, or an impulse train generated by periodic vibrations of the vocal chords pulses, for voiced sounds [4].

In speech analysis, the purpose of the all-pole model obtained through LP is to construct a spectral envelope that models the behavior of the vocal tract. For a segment of unvoiced speech, considering the excitation of the all-pole filter as white noise, the envelope is the same as its power spectrum of and the LP model coincides theoretically with the autoregressive (AR) model [5]. However, for a segment of voiced speech, the connection is more complex. The power spectrum of the voiced speech signal has a clear harmonic structure that can be approximated more effectively as a line spectrum [6]. The line frequencies are located at the multiples of the pitch frequency and their amplitude are given by the shape of the spectral envelope.

The all-pole coefficients are usually identified by minimizing the mean-squared (2-norm) error of the difference between the observed signal and the predicted signal [7]. In the source-filter model, this approach yields the LP all-pole filter, thus the prediction error (the residual signal) represents the source. Unvoiced speech lends itself readily to the principles of the 2-norm error criterion as a means of estimating the model parameters [2]. Furthermore, the 2-norm is consistent with an i.i.d. Gaussian interpretation of the prediction residual [8, 9]. The quality of the 2-norm based LP all-pole model in the context of voiced speech, which is approximately two-thirds of speech, is questionable and, theoretically, not well-founded. In particular, the all-pole spectrum does not

provide a good spectral envelope and sampling the spectrum at the line frequencies does not provide a good approximation of their amplitudes [10]. In general, the shortcomings of LP in spectral envelope modeling can be traced back to the 2-norm minimization[2]. In particular, analyzing the goodness of fit between a given harmonic line spectrum and its LP model, as done in [2], a major flaw can be derived. The LP tries to cancel the input voiced speech harmonics causing the resultant all-pole model to have poles close to the unit circle. Consequently, the LP spectrum tends to overestimate the spectral powers at the formants, providing a sharper contour than the original vocal tract response. 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 (notably [13, 10, 14]) while some others are based on changing the statistical assumptions made on the prediction error in the minimization process (notably [15, 16]). Many other formulations for finding the parameter of the all-pole model exist, a special mention is for methods that include perceptual knowledge into the estimation process (e.g., [17, 18]), or account for the non-linearities in the speech production model, e.g., [19].

Despite the wealth of alternative methods introduced to overcome the deficiencies of the 2-norm criterion, traditional usage of LP methods is, however, still confined to modeling only the spectral envelope (the vocal tract transfer function), i.e., 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 and the spectrum will still show a clear harmonic structure. The usual approach is then to employ a cascaded structure where, after LP is initially applied to determine the short-term

[2]To the authors' knowledge, the "original sin" behind the use of the 2-norm in LP, comes from its first application in speech coding, trying to reduce the entropy of speech for more efficient encoding than simple differential pulse code modulation [11]. The fundamental theorem of predictive quantization [12] 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.

prediction coefficients, a long-term predictor is determined to model the harmonic behavior of the spectrum [4]. Such a structure is arguably suboptimal since it ignores the interaction between the two different stages [20, 21]. This is known in the literature and early contributions have outlined gains in performance in jointly estimating the two filters (the work in [22] is perhaps the most successful attempt). The combination of the two filters determines a high-order linear predictor with a pretty evident sparse characteristics.

In recent work [23, 24], a more general framework for LP was presented with several benefit by introducing sparsity in the LP minimization framework. This was renamed sparse linear prediction (SpLP). In particular, while reintroducing well-known methods to seek a short-term predictor that produces a residual that is sparse rather than minimum variance (e.g., [16, 25]), the idea of employing high-order SpLP (HOSpLP) to model the cascade of short-term and long-term predictors was also introduced [26, 27]. The application of HOSpLP was originally introduced for speech processing purposes, however its formulation is intimately related to the regularization of ill-conditioned problems and to the precise modeling of long-term redundancies, thus it quickly found applications in diverse fields, such as radar [28], geology [29], video packet-loss concealment [30], and general signal representations [31, 32].

The SpLP problem can be posed as a linear programming problem, a special case of convex optimization. In order to be deployed in real-time applications, it requires its convex optimization core to be embedded directly in the algorithm that runs online and where strict real-time constraints apply. While convex optimization problems can be efficiently solved, both in theory, with worst-case polynomial complexity [33], and in practice, such as [34], it is rarely limited in its implementation by real-time constraints as discussed in [35]. The real-time implementation of such schemes calls for application-tailored optimization methods able to solve instances of the optimization problem at hand in a fast and reliable way (see, e.g., [36, 37, 38] for application in signal processing). Convex optimization solvers are usually based on iterative approaches, which is in contrast to traditional methods relaying on closed-form solutions. This is also the case for LP and SpLP. The former, has a closed-form solution that, e.g., can be calculated via the Levinson-Durbin algorithm with time complexity $\mathcal{O}(N^2)$,

with N being the prediction order. The latter, when solved with, e.g., interior-point methods [39, 37], has a time complexity of $\mathcal{O}(T^2N+T^3)$ or $\mathcal{O}(N^2T+N^3)$ depending on the setup where T is the frame size. Thus, considering the high-order case where $N \approx T$, the sparse solution is at least a order of magnitude more costly to achieve.

To reduce the complexity of solving the SpLP problem, we turn our attention to two other methods, specifically the Douglas-Rachford (DR) method and the alternating direction method of multipliers (ADMM). These two methods applied with a splitting technique that have become popular in recent years for problems requiring only moderate accuracy, see e.g. the treatment in [40, 41]. The DR method originates from [42, 43], and have recently found applications in signal processing problems, e.g., [44, 45]. Similar, the ADMM method originates from [46, 47] and have also found applications in signal processing, e.g., [48, 49]. Interestingly, there are known connections among proximal methods, Bregman iterative regularization, and the DR and ADMM algorithms. Specifically, ADMM can be understood as the DR method applied to the dual problem [50, 51, 52].

In this paper, we will show how to reduce the per-iteration time complexity for an iterative solver for the SpLP problem employing splitting methods rather than interior-point methods. The splitting methods require solving a subproblem involving a symmetric positive definite Toeplitz coefficient matrix. By exploiting this particular structure, the time complexity is quadratic ($\mathcal{O}(N^2+T^2)$) for the initialization step but linearithmic ($\mathcal{O}(N \log N+T \log T)$) for all the subsequent iterations. In order to evaluate the approximate solutions, firstly, we assess their performance via prediction gain and, secondly, assess the performance in terms of perceptual objective quality measures in a speech reconstruction framework. Despite solving the SpLP problem to a lower accuracy compared to interior-point methods, the results show that the solutions still achieve similar performance when employed in typical speech processing applications.

The paper is organized as follows. In Section 2, we present principles of linear prediction including conventional methods and sparse linear prediction. The proposed methods for solving the sparse linear prediction problem are presented in Section 3, and the computational costs are assessed by timings the methods

in Section 4. Finally, we present experimental results of the performance of the presented predictors both in objective terms by analyzing their prediction gains in Section 5.1 and in perceptual terms by employing them to reconstruct missing data in a speech reconstruction framework in Section 5.2. In Section 6, we provide additional discussions and conclude our work.

2. Principles of linear prediction

Linear prediction is based on the following model, where a stationary set of samples of speech $x[t]$, for $t = 1, \ldots, T$, are written as a linear combination of N past samples [1]

$$x[t] = \sum_{n=1}^{N} \alpha_n x[t-n] + r[t], \tag{1}$$

where $\{\alpha_n\}$ are the prediction coefficients and $r[t]$ is the prediction error. The optimization problem is then to find the prediction coefficient vector $\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_N]^T$ so that the prediction error in this interval is minimized [5]. If we rewrite this model for the segment of T samples considered in matrix form

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

the optimization problem becomes

$$\underset{\alpha}{\text{minimize}} \quad \|x - X\alpha\|_p^p \tag{3}$$

where $\|\cdot\|_p$ is the p-norm defined as $\|\mathbf{x}\|_p = (\sum_{n=1}^{N} |x(n)|^p)^{\frac{1}{p}}$ for $p \geq 1$ and

$$[x|X] = \begin{bmatrix} x[T_1] & \cdots & x[T_1 - n] \\ \vdots & \ddots & \vdots \\ x[T_2] & \cdots & x[T_2 - n] \end{bmatrix}. \tag{4}$$

Assuming $x[t] = 0$ for $t < 1$ and $t > T$, the indexes T_1 and T_2 can be chosen in various ways which lead to different types of solutions with different properties [5].

2.1. Conventional linear prediction

In a common case, the starting and ending point of the window used to determine the set of equation in (3) are chosen as $T_1 = 1$ and $T_2 = T + N$

and the 2-norm is minimized ($p = 2$). This leads to the conventional linear prediction problem

$$\underset{\alpha}{\text{minimize}} \quad \|x - X\alpha\|_2^2 \tag{5}$$

with $x \in \mathbb{R}^{M\times 1}, X \in \mathbb{R}^{M\times N}, \alpha \in \mathbb{R}^{N\times 1}$, $M = T + N$, and an analytic solution satisfying the normal equation

$$X^T X\alpha = X^T x\,. \tag{6}$$

This approach is also known as the Yule-Walker method for autoregressive (AR) spectral estimation or the autocorrelation method [5]. By exploiting the Toeplitz structure of the autocorrelation matrix $R = X^T X$, the system can be solved efficiently with, e.g., the Levinson–Durbin algorithm in $\mathcal{O}(N^2)$ [1].

2.2. Long-term prediction

Generally, linear prediction models only short-term redundancies of speech, thus is often used in combination with a single-tap or multi-tap long-term predictor [22]. The speech model for the long-term predictor is

$$d[t] = \sum_{k=0}^{K} \phi_k d[t - T_\mathrm{p} - k] + r[t], \tag{7}$$

where, similarly to (1), $\{\phi_k\}$ are the prediction coefficients and $r[t]$ is the prediction error. The major difference of the model, other than the relatively small number of taps K employed gathered around the pitch period T_p is that the optimization problem is generally done on the (usually weighted) output of the short-term filter $d[t]$. Considering the matrix form of (7)

$$d = D\phi + r, \tag{8}$$

we can obtain the coefficients by solving the optimization problem

$$\underset{\phi}{\text{minimize}} \quad \|d - D\phi\|_2^2\,. \tag{9}$$

The method obviously requires an estimate of T_p which can be found with a variety of methods all with different complexity and accuracy trade-offs [6].

2.3. Combining short-term and long-term prediction

If we consider the cascade of the long-term and short-term predictor, it is not hard to see that a sparse high-order filter is obtained. This was noted already in [22], and used to find short-term and long-term predictors jointly. In particular, the sparse filter will have two well distinguished nonzero regions: the first N taps will correspond to the short-term predictor α while the taps after the pitch period T_{p} will correspond to the convolution between the short-term and long-term coefficients $\varrho = \alpha * \phi$.

This gives us the opportunity of seeing the estimation problem as

$$\underset{\upsilon}{\text{minimize}} \quad \|x - U\upsilon\|_2^2 \tag{10}$$

where U is a column-wise partition of X accounting only for the nonzero elements of the combined filter $\upsilon = \left[\alpha^T \ \varrho^T\right]^T$. Again, applying the same traditional linear prediction approach shown in (5), we obtain the normal equations

$$U^T U \upsilon = U^T x \,, \tag{11}$$

where $U^T U$ retains a Toeplitz structure and its size is much smaller than $X^T X$, as only a fraction of the elements in x is used to estimate the combined predictor υ.

2.4. High-order prediction

As mention earlier, LP of speech is well known mostly for its modeling capabilities of short-term redundancies, which corresponds to a model of the envelope of the spectrum. When the LP order increases, the model starts encompassing more and more details of the spectrum, thus allowing for a more complete frequency representation. This follows directly from the theory, where for $N \to \infty$ the spectrum of the predictor matches the one of the signal [2]. This means that also in the voiced speech case, we can model the signal using just a high-order LP model without worrying about, e.g., pitch estimation. However, calculating a high-order predictor generally results in a ill-conditioned problem as the model order N approaches the length of the frame T [53]. The ill-conditioning can be easily tracked back to the autocorrelation matrix $R = X^T X$. In particular, the eigenvalue spread of the autocorrelation matrix corresponds to the ratio between

maximum and minimum value of the power spectrum of the speech segment analyzed, therefore, except for few exceptions, a high order in LP for speech makes R easily ill-conditioned and can cause large variance of the estimated model parameters, leading to spurious peaks in the signal spectral estimate [54]. While regularization is possible to reduce the eigenvalue spread and thus conditioning, this also corresponds to adding a noise floor, affecting the accuracy of the solution.

2.5. High-order sparse linear prediction

In our recent work [24], we generalized the optimization problem in (3) by adding a regularization criterion on the solution vector

$$\underset{\alpha}{\text{minimize}} \quad \|x - X\alpha\|_p^p + \gamma\|\alpha\|_l^l \,. \tag{12}$$

Considering the similarities with conventional high-order LP, imposing sparsity on the high-order predictor while retaining a 2-norm criterion on the prediction error, the problem can be seen as a more educated regularization approach that accounts for the sparsity of the predictor resulting from modeling short-term and long-term redundancies [26]

$$\underset{\alpha}{\text{minimize}} \quad \|x - X\alpha\|_2^2 + \gamma\|\alpha\|_1 \,. \tag{13}$$

However, when imposing sparsity on both the residual vector and high-order predictor, gains can been obtained both in terms of modeling and coding performance [27]

$$\underset{\alpha}{\text{minimize}} \quad \|x - X\alpha\|_1 + \gamma\|\alpha\|_1 \,. \tag{14}$$

For problem (14) we denote a solution $\alpha^\star$, objective $f(\alpha) = \|x - X\alpha\|_1 + \gamma\|\alpha\|_1$ and optimal objective $f^\star = f(\alpha^\star)$. The regularization term γ in (14) can be seen as related to the prior knowledge of the prediction coefficients vector α. While sparsity is often measured by the cardinality $\mathbf{card}(x) = |\{\, i \,|\, x_i \neq 0 \,\}|$, we use the more computational tractable 1-norm $\|\cdot\|_1$, which is known throughout the sparse recovery literature (see, e.g., [55]) to perform well as a relaxation of the cardinality measure with equivalence in certain cases.

3. Solving the sparse linear prediction problem

The objective function in (14), as well as the terms that compose it, is convex but not differentiable, thus proximal gradients method are not directly applicable [56, 57, 58]. The optimization can be solved as a general linear programming problem using interior-point methods [39, 37]. However, solving (14) using interior-point methods introduces certain diagonal matrices D_1, D_2 into the problem, such that when using the augmented form approach, it is required to solve a linear system of equations with the coefficient matrix $C = X^T D_1 X + D_2$ where D_1, D_2 change at each iteration. This makes it difficult to exploit the structure in X and $X^T X$ using direct method (see, e.g., [45]). In [37], the coefficient matrix C is explicit formed and solved via a Cholesky factorization followed by triangular solves with a per iteration complexity of $\mathcal{O}(M^2 N + M^3)$ or $\mathcal{O}(N^2 M + N^3)$ depending on the setup [37].

In the following, it is showed how to exploit the DR and the ADMM methods to reduce the per-iteration complexity. Specifically, with the use of splitting combined with ADMM and DR methods, an auxiliary symmetric positive definite Toeplitz system arises that can be solved with a total per-iteration complexity of $\mathcal{O}(N \log N + M \log M)$ (forming right-hand side and solving the system). The algorithm for solving the system is described in the subsequent Section 3.3.

The saving in per-iteration complexity of the DR and ADMM methods reflects into a slower convergence than interior-point methods. However, when tailored to speech processing applications, a low-accuracy solution of the problem (14) obtained via the DR or ADMM method is sufficient for practical purposes, as we will show in Section 5.

3.1. Douglas-Rachford

In the following we will rewrite the SpLP into a form applicable to the DR algorithm. We will write the problem in (14) as

$$\underset{\alpha}{\text{minimize}} \quad f_1(\alpha) + f_2(X\alpha) \tag{15}$$

where $f_1(u) = \gamma \|u\|_1$ and $f_2(u) = \|x - u\|_1$. Introducing the variable splitting $h(u_1, u_2) = f_1(u_1) + f_2(u_2)$ the problem can be reformulated as the optimization

problem

$$\begin{aligned} \underset{u_1,u_2}{\text{minimize}} \quad & h(u_1,u_2) \\ \text{subject to} \quad & u_2 = Xu_1\,. \end{aligned} \tag{16}$$

Before we proceed, we define two functions. The proximal mapping of a convex function f is given by (see, e.g., [59] or [60] for a more recent treatment)

$$\mathbf{prox}_f(x) = \underset{u}{\operatorname{argmin}} \left(f(u) + \frac{1}{2}\|u-x\|_2^2 \right)\,. \tag{17}$$

The Euclidean projection of x onto a set $\mathbb{C}$ is $\mathcal{P}_{\mathbb{C}}(x)$ given by

$$\mathcal{P}_{\mathbb{C}}(x) = \underset{y\in\mathbb{C}}{\operatorname{argmin}} \|x-y\|_2^2\,. \tag{18}$$

Using the indicator function $\mathcal{I}_{\mathbb{C}}$ of the set $\mathbb{C}$, we obtain

$$\mathcal{P}_{\mathbb{C}}(x) = \mathbf{prox}_{\mathcal{I}_{\mathbb{C}}}(x)\,. \tag{19}$$

Let $\mathbb{Q} = \{\, [u_1,u_2]^T \mid u_2 = Xu_1 \,\}$, the Douglas-Rachford splitting method applied to problems of the form

$$\begin{aligned} \underset{u\in\mathbb{R}^U}{\text{minimize}} \quad & h(u) \\ \text{subject to} \quad & u \in \mathbb{Q} \end{aligned} \tag{20}$$

can be written in a number of equivalent forms, including [45, 51, 61]

$$u^{(k+1)} = \mathbf{prox}_{th}(z^{(k)}) \tag{21}$$

$$y^{(k+1)} = \mathcal{P}_{\mathbb{Q}}(2u^{(k+1)} - z^{(k)}) \tag{22}$$

$$z^{(k+1)} = z^{(k)} + \eta(y^{(k+1)} - u^{(k+1)}) \tag{23}$$

with the iterates $u^{(k)}, z^{(k)}, y^{(k)} \in \mathbb{R}^{U\times 1}$. Here $\eta \in \mathbb{R}$ is a relaxation parameter $0 < \eta < 2$, $t \in \mathbb{R}, t > 0$ is a step-size parameter, or, equivalently a scaling of the problem.

The individual steps for solving the sparse linear prediction problem using Douglas-Rachford splitting are described in the following. For the update in (21), notice that $\mathbf{prox}_{th}(u_1,u_2) = [\mathbf{prox}_{tf_1}(u_1), \mathbf{prox}_{tf_2}(u_2)]^T$, i.e., it is separable since the binding is done in the constraints. A classical result is that with $f_1(u_1) = \gamma\|u_1\|_1$

$$\mathbf{prox}_{tf_1}(v) = \mathcal{S}_{t\gamma}(v) \tag{24}$$

where S is the soft-thresholding function given by

$$(\mathcal{S}_t(v))_i = \operatorname{sign}(v_i)\max(|v_i| - t, 0) \tag{25}$$

and $(\cdot)_i$ denotes the ith element. For $f_2(u) = \|x - u\|_1$ we have

$$\begin{aligned}
\mathbf{prox}_{tf_2}(v) &= \operatorname*{argmin}_u \left(t\|x-u\|_1 + \frac{1}{2}\|u-v\|_2^2 \right) \\
&= x - \operatorname*{argmin}_w \left(t\|w\|_1 + \frac{1}{2}\|w-(x-v)\|_2^2 \right) \\
&= x - \mathbf{prox}_{t\|\cdot\|_1}(x-v) \\
&= x - \mathcal{S}_t(x-v)\,. \qquad (26)
\end{aligned}$$

So, $\mathbf{prox}_{th}(u_1, u_2) = [\mathbf{prox}_{tf_1}(u_1), \mathbf{prox}_{tf_2}(u_2)]^T = [\mathcal{S}_t(u_1), x - \mathcal{S}_t(x - u_2)]^T$ can be calculated with complexity $\mathcal{O}(M + N)$. The projection $\mathcal{P}_{\mathbb{Q}}(v)$ is given by

$$\begin{aligned}
\mathcal{P}_{\mathbb{Q}}(v) &= \operatorname*{argmin}_{u\in\mathbb{Q}} \|u-v\|_2^2 \qquad (27) \\
&= \operatorname*{argmin}_{u_2=Xu_1} \|u_1-v_1\|_2^2 + \|u_2-v_2\|_2^2\,. \qquad (28)
\end{aligned}$$

This is a convex quadratic problem with the KKT conditions

$$\begin{aligned}
u_1 - v_1 - X^T\nu &= 0 \qquad (29) \\
u_2 - v_2 + \nu &= 0 \qquad (30) \\
u_2 &= Xu_1 \qquad (31)
\end{aligned}$$

from which we obtain

$$\begin{aligned}
0 &= u_1 - v_1 - X^T\nu \qquad (32) \\
&= u_1 - v_1 - X^T(v_2 - u_2) \qquad (33) \\
&= u_1 - v_1 - X^Tv_2 + X^TXu_1\,. \qquad (34)
\end{aligned}$$

Hence we obtain the linear systems $(I + X^TX)u_1 = v_1 + X^Tv_2$ and $u_2 = Xu_1$. The projection is then

$$\mathcal{P}_{\mathbb{Q}}(v) = \begin{bmatrix} I \\ X \end{bmatrix} (I + X^TX)^{-1}(v_1 + X^Tv_2)\,. \qquad (35)$$

To compute (35) we need to solve a linear system of equations with coefficient matrix $I + X^TX$ and varying right-hand sides $(v_1 + X^Tv_2)$. The coefficient

matrix is positive definite, symmetric and Toeplitz. Specifically, let

$$R = X^T X = \begin{bmatrix} r_0 & r_1 & \cdots & r_{N-1} \\ r_1 & r_0 & \cdots & r_{N-2} \\ \vdots & & \ddots & \vdots \\ r_{N-1} & r_{N-2} & \cdots & r_0 \end{bmatrix}. \tag{36}$$

and

$$t_0 = 1 + r_0 \tag{37}$$

$$t_k = r_k, \quad k = 1, \ldots, N-1 \tag{38}$$

then the positive definite, symmetric Toeplitz matrix is

$$I + X^T X = \begin{bmatrix} t_0 & t_1 & \cdots & t_{N-1} \\ t_1 & t_0 & \cdots & t_{N-2} \\ \vdots & & \ddots & \vdots \\ t_{N-1} & t_{N-2} & \cdots & t_0 \end{bmatrix}. \tag{39}$$

It is well known that linear systems with a coefficient matrix given as (39) can be solved efficiently. We will discuss different methods in Section 3.3.

3.2. Alternating Direction Method of Multipliers

A first step in deriving an ADMM algorithm consists in reformulating the problem in (14) as a basis pursuit problem, following the procedure in [49]. To this end, we first rewrite the unconstrained problem in (14) as an equality constrained problem

$$\begin{aligned} \underset{\alpha, e}{\text{minimize}} \quad & \|e\|_1 + \gamma \|\alpha\|_1 \\ \text{subject to} \quad & X\alpha + e = x \end{aligned} \tag{40}$$

where $e = x - X\alpha$ represents the linear prediction residual. Next, we perform a change of variables by stacking a scaled version of the linear prediction coefficient vector and the linear prediction residual in a new parameter vector

$$\tilde{z} = \begin{bmatrix} \gamma\alpha \\ e \end{bmatrix}. \tag{41}$$

This allows to reformulate the problem in (40) using $\gamma\|\alpha\|_1 = \|\gamma\alpha\|_1$ in terms of a single parameter vector as follows [49]

$$\begin{aligned} \underset{\tilde{z}}{\text{minimize}} \quad & \|\tilde{z}\|_1 \\ \text{subject to} \quad & \tilde{X}\tilde{z} = \tilde{x} \end{aligned} \tag{42}$$

where

$$\tilde{X} = \begin{bmatrix} X & \gamma I \end{bmatrix} \tag{43}$$

$$\tilde{x} = \gamma x\,. \tag{44}$$

In a second step, we write the basis pursuit problem in (42) in the ADMM form as explained in [40]. To this end, we define the set $\mathbb{U} = \{\tilde{z} \in \mathbb{R}^{m+n} \,|\, \tilde{X}\tilde{z} = \tilde{x}\}$ and introduce an variable $\tilde{y} \in \mathbb{R}^{m+n}$ such that the basis pursuit problem can be split over $\tilde{z}$ and $\tilde{y}$,

$$\begin{aligned} \underset{\tilde{z},\tilde{y}}{\text{minimize}} \quad & \mathcal{I}_{\mathbb{U}}(\tilde{z}) + \|\tilde{y}\|_1 \\ \text{subject to} \quad & \tilde{z} - \tilde{y} = 0 \end{aligned} \,. \tag{45}$$

This problem formulation readily brings us to an ADMM algorithm defined by the iterations [40]

$$\tilde{z}^{(k+1)} = \mathcal{P}_{\mathbb{U}}(\tilde{y}^{(k)} - \tilde{u}^{(k)}) \tag{46}$$

$$\tilde{y}^{(k+1)} = \mathcal{S}_{1/\rho}(\tilde{z}^{(k+1)} + \tilde{u}^{(k)}) \tag{47}$$

$$\tilde{u}^{(k+1)} = \tilde{u}^{(k)} + \tilde{z}^{(k+1)} - \tilde{y}^{(k+1)}\,. \tag{48}$$

Variables $\tilde{z}^{(k)}$ and $\tilde{y}^{(k)}$ denote iterates of the primal variables, $\tilde{u}^{(k)}$ denotes the scaled dual variable, and $\rho > 0$ is the augmented Lagrangian parameter. Similarities can then be seen with the DR algorithm in (21)–(23).

We will now focus on the $\tilde{z}$-update in (46), which is the most involved step of the algorithm. The $\tilde{z}$-update consists in the projection of the point $\tilde{y}^{(k)} - \tilde{u}^{(k)}$ onto the convex set $\mathbb{U}$ with the following KKT conditions

$$\tilde{z} + (\tilde{y}^{(k)} - \tilde{u}^{(k)}) - \tilde{X}^T\lambda = 0 \tag{49}$$

$$\tilde{X}\tilde{z} = \tilde{x}\,. \tag{50}$$

It is instructive to rewrite these KKT conditions in terms of the original parameter vectors α and e by substituting the variable definitions (41), (43), and (44)

in the KKT system (49)-(50)

$$\gamma\alpha - (\tilde{y}_1^{(k)} - \tilde{u}_1^{(k)}) - X^T\lambda = 0 \tag{51}$$

$$e - (\tilde{y}_2^{(k)} - \tilde{u}_2^{(k)}) - \gamma\lambda = 0 \tag{52}$$

$$X\alpha + e = x \tag{53}$$

where

$$\tilde{y}^{(k)} = \begin{bmatrix} \tilde{y}_1^{(k)} \\ \tilde{y}_2^{(k)} \end{bmatrix}, \ \tilde{u}^{(k)} = \begin{bmatrix} \tilde{u}_1^{(k)} \\ \tilde{u}_2^{(k)} \end{bmatrix}, \tag{54}$$

have been partitioned similarly to $\tilde{z}$ in (41). From the KKT conditions we obtain

$$\begin{aligned} 0 &= \gamma\alpha - (\tilde{y}_1^{(k)} - \tilde{u}_1^{(k)}) - X^T\lambda && (55) \\ &= \gamma\alpha - (\tilde{y}_1^{(k)} - \tilde{u}_1^{(k)}) \\ &\quad - \frac{1}{\gamma}X^T(e - (\tilde{y}_2^{(k)} - \tilde{u}_2^{(k)})) && (56) \\ &= \gamma\alpha - (\tilde{y}_1^{(k)} - \tilde{u}_1^{(k)}) \\ &\quad - \frac{1}{\gamma}X^T(x - X\alpha - (\tilde{y}_2^{(k)} - \tilde{u}_2^{(k)}))\,. && (57) \end{aligned}$$

This results in the following system to be solved for the linear prediction coefficient vector α

$$\begin{aligned} (X^TX + \gamma^2 I)\alpha &= X^T + \gamma(\tilde{y}_1^{(k)} - \tilde{u}_1^{(k)}) - X^T(\tilde{y}_2^{(k)} - \tilde{u}_2^{(k)}) \\ &= X^Tx - \begin{bmatrix} -\gamma I & X^T \end{bmatrix} (\tilde{y}^{(k)} - \tilde{u}^{(k)})\,. \end{aligned} \tag{58}$$

The solution $\alpha^{(k+1)}$ to this system in the $(k+1)$th ADMM iteration can be expressed as the sum of an iteration-independent term and an iteration-dependent term,

$$\begin{aligned} \alpha^{(k+1)} &= \underbrace{(X^TX + \gamma^2 I)^{-1}X^Tx}_{\triangleq\, \alpha_{\gamma,2}} \\ &\quad - \underbrace{(X^TX + \gamma^2 I)^{-1}\begin{bmatrix} -\gamma I & X^T \end{bmatrix}}_{\triangleq \begin{bmatrix} -\gamma I \\ X \end{bmatrix}^+} (\tilde{y}^{(k)} - \tilde{u}^{(k)}) \end{aligned} \tag{59}$$

where $(\cdot)^+$ denotes the Moore-Penrose pseudo-inverse. The iteration-independent term $\alpha_{\gamma,2}$ is the solution to the ℓ_2-regularized linear prediction problem in (12)

with $p = l = 2$. This system may, e.g., be solved by by applying the Levinson-Durbin algorithm to the modified autocorrelation sequence $\{r_0 + \gamma^2, r_1, \ldots, r_N\}$ with r_k as defined in (36) for $k = 0, \ldots, N-1$ and similarly for $k = N$. The iteration-dependent term in (59) can be computed by solving a positive definite, symmetric Toeplitz system similar to the system solved in the DR method.

The ADMM iterations (46)-(48) can hence be rewritten as follows

$$\alpha^{(k+1)} = \alpha_{\gamma,2} - \begin{bmatrix} -\gamma I \\ X \end{bmatrix}^{+} (\tilde{y}^{(k)} - \tilde{u}^{(k)}) \tag{60}$$

$$e^{(k+1)} = x - X\alpha^{(k+1)} \tag{61}$$

$$\tilde{y}^{(k+1)} = \mathcal{S}_{1/\rho} \left(\begin{bmatrix} \gamma\alpha^{(k+1)} \\ e^{(k+1)} \end{bmatrix} + \tilde{u}^{(k)} \right) \tag{62}$$

$$\tilde{u}^{(k+1)} = \tilde{u}^{(k)} + \begin{bmatrix} \gamma\alpha^{(k+1)} \\ e^{(k+1)} \end{bmatrix} - \tilde{y}^{(k+1)} . \tag{63}$$

Note that with $\tilde{y}^{(0)} - \tilde{u}^{(0)} = 0$, we have $\alpha^{(1)} = \alpha_{\gamma,2}$, and the ADMM algorithm can then be interpreted as iterative "sparsification" of the ℓ_2-regularized "classical" linear prediction solution.

3.3. Solving a positive definite symmetric Toeplitz system

A classical algorithm for solving a (positive definite) symmetric Toeplitz system is the Levinson algorithm with time complexity $\mathcal{O}(N^2)$ and space complexity $\mathcal{O}(N)$ [62, 63]. Algorithms like the Levinson algorithm with time complexity $\mathcal{O}(N^2)$ are called fast algorithms, but there also exist superfast algorithms with time complexity $\mathcal{O}(N \log^2 N)$, see [64, 65]. These algorithms also have the advantage that the first solution can be obtained in $\mathcal{O}(N \log^2 N)$ and any other solution with the same coefficient matrix but different right-hand side is possible with linearithmic time complexity $\mathcal{O}(N \log N)$. There are also algorithms where there is a one time penalty of $\mathcal{O}(N^2)$ and again all subsequent solutions with same coefficient matrix but a different right-hand side only requires $\mathcal{O}(N \log N)$ [66, 65]. The constant in the first step of a superfast algorithm is large and hence there is a break-even point in the number of operations at approximately $N = 256$ for N as a base 2 number [65]. Since the experiments in Section 4 and 5 uses $N < 256$ and that the $\mathcal{O}(N^2)$ algorithm in the first step is much simpler, we use the algorithm in [66], see also [65].

The inverse of a Toeplitz matrix $\bar{T}$ can be described by the Gohberg-Semencul formula

$$\delta_{N-1}\bar{T}^{-1} = \bar{T}_1\bar{T}_1^T - \bar{T}_0^T\bar{T}_0 \tag{64}$$

where

$$\bar{T}_0 = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ \rho_0 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ \rho_{N-2} & \cdots & \rho_0 & 0 \end{bmatrix}, \tag{65}$$

and

$$\bar{T}_1 = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ \rho_{N-2} & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ \rho_0 & \cdots & \rho_{N-2} & 1 \end{bmatrix}, \tag{66}$$

and

$$\delta_{N-1}\bar{T}^{-1} = \begin{bmatrix} 1 & \rho_{N-2} & \cdots & \rho_0 \\ \rho_{N-2} & & & \\ \vdots & & S & \\ \rho_0 & & & \end{bmatrix} \tag{67}$$

and $S \in \mathbb{R}^{(N-1)\times(N-1)}$ denotes the remaining submatrix. The variables δ_{N-1} and $\rho_0, \ldots, \rho_{N-2}$ can be obtained using the Szegö recursion in $\mathcal{O}(N^2)$ operations as a one time-cost per system. The solution to the system $\bar{T}x = b$ is then given by

$$x = \bar{T}^{-1}b = \frac{1}{\delta_{N-1}}\left(\bar{T}_1\bar{T}_1^T b - \bar{T}_0^T\bar{T}_0 b\right) . \tag{68}$$

Since $\bar{T}_0, \bar{T}_1$ are Toeplitz, a product like $T_0 b$ can be evaluated via FFTs/IFFTs, see [66] for an algorithm for fast evaluation of (68) in $\mathcal{O}(N \log N)$ operations. So, all subsequent solutions with the same coefficient matrix are available in $\mathcal{O}(N \log N)$ operations. Recall that the coefficient matrix is $X^T X + I = R + I$ for the DR algorithm and $X^T X + \gamma^2 I = R + \gamma^2 I$ for the ADMM algorithm. From the perspective of signal processing, the coefficient matrices $X^T X + I = R + I$ and $X^T X + \gamma^2 I = R + \gamma^2 I$ are updated for each frame. It is only during each call of the DR and ADMM algorithms that the coefficient matrix is fixed. For each call the computation of δ_{N-1} and $\rho_0, \ldots, \rho_{N-2}$ with appropriate discrete Fourier transforms (DFTs) of the latter sequence can be seen as an initial caching to

make subsequent iterations cheaper (as in, e.g., [48]). Note that the diagonal term provides regularization to the solver.

4. Implementation and empirical computation time

The DR and ADMM algorithms for solving the SpLP problem using the Levinson algorithm to solve the symmetric positive definite Toeplitz system, are denoted **DR-L** and **ADMM-L**. The DR and ADMM algorithm that are using the method in [66] to solve the symmetric positive definite Toeplitz system are denoted **DR-GS** and **ADMM-GS**.

The algorithms[3] were implemented in C++ using Intel Math Kernel Library (MKL) [67] for BLAS level 1 routines. The application of matrix-vector product with X and X^T was implemented as FFT filtering using the FFTW3 library [68]. The time from call of the solver to return was measured using the POSIX function `gettimeofday` (for the C++ implementations). The timing was measured over 100 executions of each frame to average out possible system processes (note that each frame was then static and the solvers then run with the exact same input). The setting $\gamma = 0.12$ was found in Section 5 and fixed for all simulations. The simulations were executed on an Intel(R) Dual Core(TM) i5-2410M CPU at 2.3 GHz with Ubuntu Linux kernel 3.2.0-32-generic, MKL 10.3 and Matlab 7.13.0.564. The algorithms implemented with C++ were compiled using gcc-4.8 and the `-Os -march=native` optimization option. We compared the implementation with the general purpose software **Mosek** 7.0 [34] and **CVX+SeDuMi** 1.21 [69, 70]. This problem was too large to use CVXGEN [71]. Algorithms **Cprimal** and **Cprimal(s/d)** are presented in [37] and were C++ implementations of interior-point methods.

Analytically motivated selections of parameters ρ, t requires that one of the functions applied to the Douglas-Rachford setup are smooth and/or strongly convex, see [72]. Since this is not the case, we empirically found $\eta = 1.8$ and $t = 0.1$ to be useful. For the ADMM based algorithms we empirically found $\rho = 100$ to be useful.

[3]MATLAB and C++ implementations are available from `http://kom.aau.dk/~tlj/`

We chose the following stopping criteria. Algorithms **ADMM-L** and **ADMM-GS** were stopped if

$$\frac{1}{M+N}\|\tilde{z}^{(k+1)} - \tilde{y}^{(k+1)}\|_2^2 \leq \epsilon \tag{69}$$

$$\frac{\rho}{M+N}\|\tilde{y}^{(k+1)} - \tilde{y}^{(k)}\|_2^2 \leq \epsilon\,. \tag{70}$$

This reflected the primal and dual residual and was (up to a scaling of $\rho = 100$) the absolute criteria in [40]. The algorithms **DR-L** and **DR-GS** were stopped if

$$\frac{1}{N+M}\|z^{(k)} - z^{(k-1)}\|_2^2 \leq \epsilon\,. \tag{71}$$

For both the DR and ADMM based algorithms we selected $\epsilon = 10^{-6}$ and also stopped if a maximum of $k = 100$ iterations were reached. Such a maximum of allowed iterations is useful to bound the worst-case execution time.

Here we presented results for $T = 320$, $N = 250$ ($M = 570$) processed on a 2 s speech signal sampled at 16 kHz taken from the testing database. The results are shown in Table 1.

Methods	Timings
CVX+SeDuMi	**2467.29** / 1327.29/3619.74
Mosek	**224.71** / 145.54/307.60
Cprimal	**92.70** / 55.24/180.46
Cprimal(s/d)	**63.66** / 33.59/112.09
DR-L	**6.62** / 0.65/10.11
DR-GS	**2.28** / 0.61/3.26
ADMM-L	**2.99** / 0.65/5.14
ADMM-GS	**1.29** / 0.61/1.92

Table 1: Timing in milliseconds. Format: **Average**/min/max. The setting is $T = 320$, $N = 250$ ($M = 570$).

From Table 1 we observed that the splittings methods are orders faster than general purpose software and one order faster than hand-tailored interior-point

methods. It was also clear that using the algorithm [66] for solving the auxiliary symmetric positive definite Toeplitz system was faster that the classical Levinson-Durbin algorithm for these dimensions (otherwise the methods were equivalent in that the same iterations are generated). The ADMM based methods converged faster to the used stopping criteria and this was the reason for being faster than the DR based methods. Specifically, the ADMM based algorithms used on average 13.5 iterations, while the DR based algorithms used 35.3 iterations.

The splitting methods solved the problem to a low accuracy. Specifically, using the metrics

$$m_{\mathbf{DR}} = \frac{f_{\mathbf{DR}} - f^\star}{f^\star}, \qquad m_{\mathbf{ADMM}} = \frac{f_{\mathbf{ADMM}} - f^\star}{f^\star} \tag{72}$$

we observed than on average $m_{\mathbf{DR}}$ and $m_{\mathbf{ADMM}}$ is 0.14 and 0.12, respectively, i.e. approximately only a 10^{-1} sub-optimal solution ($f^\star$ is approximated via the solution from **Mosek**). The impact of this low accuracy was then assessed in the experimental analysis, presented in Section 5. An example of the different convergence behaviors of the DR and ADMM algorithms is illustrated in Figure 1 with the metric $\frac{f(\alpha^{(k)})-f^\star}{f^\star}$ using a single frame from the simulations in Table 1. Notice that the **ADMM-L** algorithm converges faster the few first iterations. The endpoint of the graphs illustrates where the stopping criteria was activated and stopped the iterative algorithms.

Algorithms **CVX+SeDuMi**, **Mosek**, **Cprimal**, **Cprimal(s/d)** are all primal-dual interior-point methods, and the accuracy of these methods is all higher than using the DR and ADMM algorithms. However, the DR and ADMM algorithms have the advantage that some of the elements in an approximate solution $\hat{\alpha} \approx \alpha^\star$ are exactly 0 due to the soft-thresholding function (25) applied at (21) and (62). On the other-hand, interior-point methods reformulate the problem as a constrained problem and approach the solution from the interior but never exactly reach the bound of the constraints of the reformulated problem, such that the small magnitude elements in $\hat{\alpha}_{\mathrm{IP}} \approx \alpha^\star$ are small but not exactly 0.

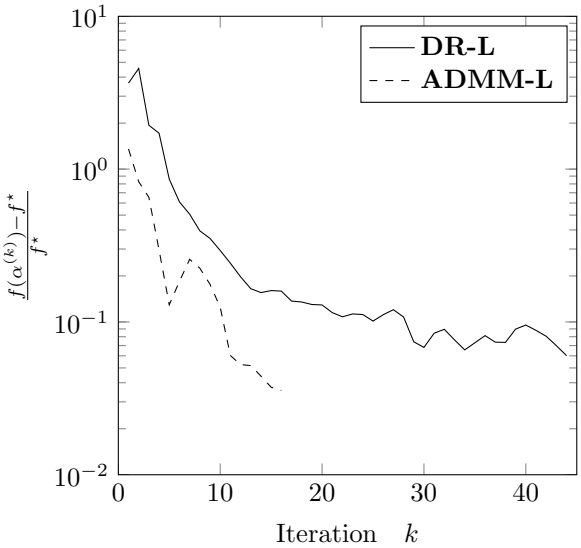

Figure 1: Example of the convergence behavior of the algorithms **DR-L** and **ADMM-L**. The endpoints of the graphs illustrates where the stopping criteria has become active and stopped the iterative algorithm.

5. Experimental analysis

In this section, we outline some of the properties of the different LP models introduced in Section 2 and, in particular the SpLP method of Section 2.5 equipped with the solvers proposed in Section 3. Firstly, we present their objective spectral modeling properties by analyzing their prediction gain. Secondly, we provide a more practical example of their modeling properties in a speech reconstruction scenario where we also evaluate the goodness of the approximate solutions of Section 3 through perceptual objective quality measures. The algorithms compared are outlined in Table 2. The splitting methods solvers in algorithms **HOSpLPdr** and **HOSpLPadmm** used the same stopping criteria as outlined in Section 4. The TIMIT database [73] was chosen for the analysis because of its manageable size with a sufficiently large number of speakers for testing speech processing algorithms accuracy. Since our algorithms windows are in the order of few milliseconds, we can reasonably assume that the conclusions of the experimental analysis done with the TIMIT dataset extent to a larger class of speakers.

Method	Description
LTP	Combined 20 tap short-term LP and 1 tap long-term LP calculated separately.
LTP3	Combined 20 tap short-term LP and 3 tap long-term LP calculated separately.
LTP3j	Combined 20 tap short-term LP and 3 tap long-term LP calculated jointly using (11).
HOLP	High-order LP with 2-norm criterion calculated with high-order (Section 2.4).
HOSpLPip	High-order sparse LP (14) obtained with interior point solution.
HOSpLPdr	High-order sparse LP (14) obtained with DR algorithm.
HOSpLPadmm	High-order sparse LP (14) obtained with ADMM algorithm.

Table 2: Prediction methods used for comparison.

5.1. Prediction gain

The vowel and semivowel phones [75] from the TIMIT database (sampled at 16 kHz) were processed, belonging to 3,696 sentences from 462 speakers. We chose the ones of duration of at least 640 samples (40 ms) for a total of about 40,000 voiced speech frames.

The methods compared are presented in Table 2. In **LTP**, **LTP3**, and **LTP3j** the short-term predictor had 20 coefficients. For the purposes of comparison, the autocorrelation method was used for both the short-term and long-term predictors. The value of the pitch lag in **LTP** and **LTP3** was chosen such that the prediction gain was maximized. This was accomplished by an exhaustive search over the allowable range $T_{\mathrm{p}} \in [34, 231]$, as used in AMR-WB [76], effectively covering pitch frequencies belonging to the range $[69, 470]$ Hz. For **LTP3j**, the pitch lag found in **LTP3** was used in the optimization. Note that,

while a plethora of methods exist for pitch estimation (see, e.g., [6]), we chose the exhaustive search to guarantee the prediction performance not to be biased by a possible erroneous estimation of the pitch lag.

In the methods **HOLP**, **HOSpLP**, **HOSpLPdr** and **HOSpLPadmm** the order of the filter was fixed to 250, allowing to cover the above-mentioned pitch lag range and a related bulk of nonzero coefficients clustered around the maximum allowable pitch lag. In order to find an appropriate value of the regularization parameter, we analyzed a set of voiced speech frames (different from the one used in the experiments) and determined the point of maximum curvature on the curve $(\|\alpha\|_1, \|x - X\alpha\|_1)$, a modified version of the L-curve [77], appropriate to determine trade-offs between the sparsity of the predictor and the sparsity of the residual. The regularization parameter was then chosen fixed, $\gamma = 0.12$.

For a fair comparison, after calculating **HOSpLP**, **HOSpLPdr** and **HOSpLPadmm**, only the 21 largest values were retained, this kept the actual sparsity of the methods the same as the simple **LTP**. Note that the actual number of nonzero samples for **LTP** would be 40 considering the convolution of short-term and long-term predictors in the analysis.

The average prediction gains are shown in Table 3. The 95% confidence intervals showed a clear distinction between the various method as well as a certain consistency in performance. Clearly, **HOLP** was the best performing method. This behavior can be explained from Parsevals theorem and the power spectrum matching properties of the all-pole spectrum obtained with 2-norm LP that could approximate the power spectrum of a signal with arbitrarily small error [2]. The performance of **HOLP** then determined the upper bound of performance.

The methods **HOSpLP**, **HOSpLPdr** and **HOSpLPadmm** behaved, in statistical terms, identically, thus providing further proof of the reliability of the proposed fast solution, even though the latter two calculated a much less accurate solution to the problem (14). In general, the sparseness criterion helped providing a net reduction in the number of nonzero samples while obtaining just slightly lower performance. The **LTP**, **LTP3**, and **LTP3j** methods provided proof of the gain in prediction gain given by the joint estimation of short-term and long-term coefficients provided by the high-order sparse model which

METHOD	**card**(·)	T	
		320	640
LTP	21	17.3±0.8	14.2±1.0
LTP3	23	22.3±0.8	19.9±0.9
LTP3j	43	24.2±0.6	22.6±0.8
HOLP	250	32.4±0.6	31.3±0.7
HOSpLPip	21	28.6±1.1	27.8±1.4
HOSpLPdr	21	28.5±1.4	27.6±1.6
HOSpLPadmm	21	28.3±1.7	27.2±1.6

Table 3: Average prediction gains for segments of different length T. A 95% confidence interval is shown. The number of nonzero elements, **card**(·), is shown for comparison.

achieves more than 10 dB gain compared to traditional **LTP**.

A proof of concept example is shown in Figure 2 and Figure 3 for a 640 samples segment of the vowel /a/ uttered by a female speaker where the predictors and related frequency magnitudes are shown. Visually, the dissimilarities between **LTP3j** and the sparse methods, which behave very similarly, come mostly from the lower order short-term predictor necessary to model the envelope (around 10 taps versus 20) and the larger cluster of taps around T_{p} needed to model the pitch redundancies. This allowed the sparse methods not to just have a more parsimonious representation, but also allowed for general better modeling thanks to the sparse criteria imposed both on the predictor and residual.

5.2. Speech Reconstruction

Considering speech as a slowly-varying process, we can verify the performance of the different predictors presented in Table 2 in a statistical cross-validation framework. In particular, for a given speech segment, we can determine how accurately the predictors perform in reconstructing a set of unknown speech samples given the known samples and their prediction model. The prob-

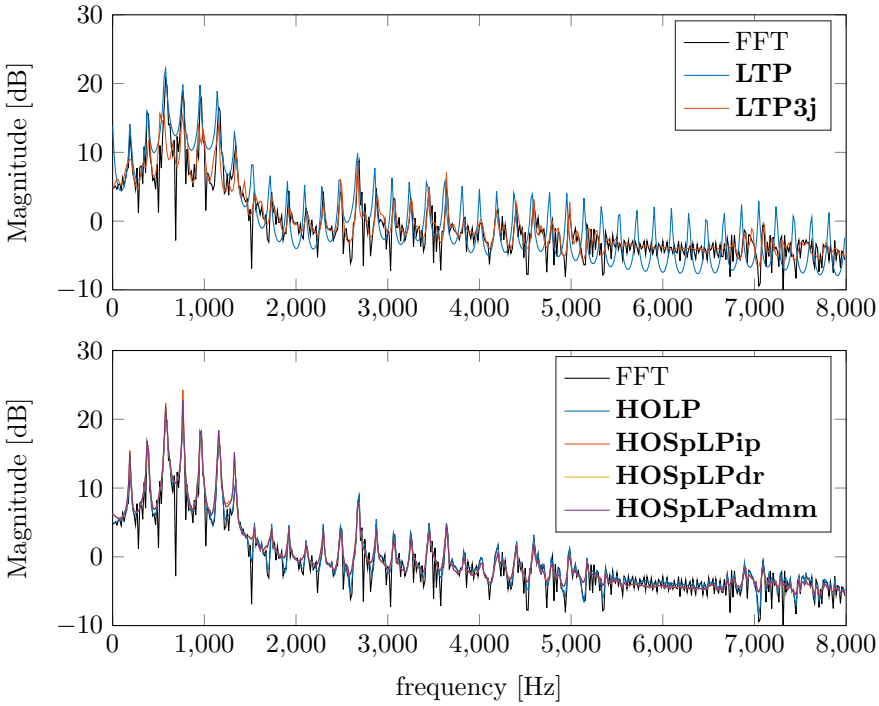

Figure 2: Magnitude of the frequency response of the different methods proposed. A 640 samples segment of the voiced speech (vowel /a/ uttered by a female speaker) is used for the analysis.

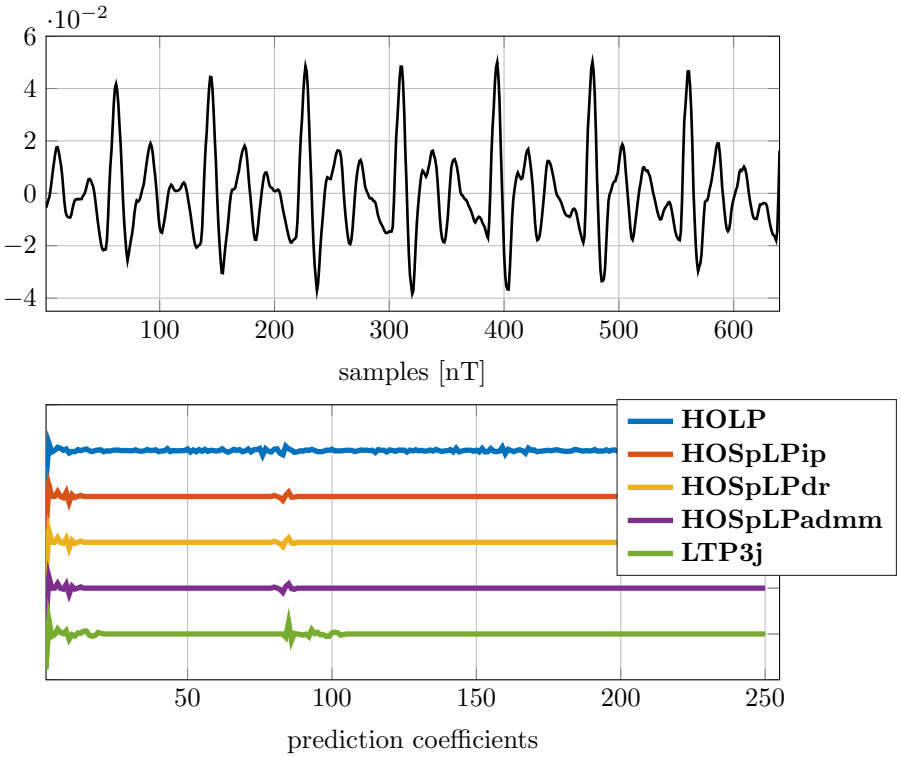

Figure 3: A 640 samples segment of the voiced speech (vowel /a/ uttered by a female speaker) and the calculated predictors for the different methods proposed.

lem can then be rewritten as a maximum a posteriori (MAP) estimator

$$\underset{x_{\mathrm{u}}}{\text{maximize}} \quad p(x_{\mathrm{u}}|x_{\mathrm{k}}, \alpha) \tag{73}$$

where x_{k} represents the known samples and x_{u} represents the unknowns samples of a given speech segment of length T and the predictor α is calculated using only the known samples. It is therefore imperative that α is effective in modeling the underlying statistics without underfitting or overfitting in order to have the best estimate of x_{u}. This problem is well-known in the statistical audio processing literature as AR model-based speech reconstruction [74]. Considering the segment of T speech samples x as partitioned in terms of known and unknown samples

$$x = Kx_{\mathrm{k}} + Ux_{\mathrm{u}}, \tag{74}$$

where U and K are $T \times T$ "rearrangements" matrices that form a columnwise partition of the identity matrix I, and if we consider the data samples x as drawn by an AR process with parameters α, we can rewrite the interpolation error as

$$e = A(Kx_{\mathrm{k}} + Ux_{\mathrm{u}}) \tag{75}$$

where A is the so-called analysis matrix obtained with α [74], thus (75) is just another way to rewrite the system of equations in (2). If we fit the interpolation error into an i.i.d. Gaussian process, which is reasonable given the limited knowledge of the reconstruction process, we obtain

$$\begin{aligned} p(x_{\mathrm{u}}|x_{\mathrm{k}}, \alpha) &\propto \exp(-\|e\|_2^2) \\ &= \exp(-\|A(Kx_{\mathrm{k}} + Ux_{\mathrm{u}})\|_2^2). \end{aligned} \tag{76}$$

Maximizing the argument of (76), we obtain

$$x_{\mathrm{u}} = -\left(A_{\mathrm{u}}^T A_{\mathrm{u}}\right)^{-1} A_{\mathrm{u}}^T A_{\mathrm{k}} x_{\mathrm{k}} \tag{77}$$

where $A_{\mathrm{u}} = AU$ and $A_{\mathrm{k}} = AK$. This solution is then equivalent to minimizing the mean-square error of e of (75). Note that the basic formulation given in (73) assumes that the AR parameters are known *a priori*. In practice, there are ways to obtain a robust estimate during the detection stage [74], however, in our case we will limit ourselves to estimating α over the known speech samples. Considering the reconstruction equations (77), it can be seen that traditional LP

model might fail to properly reconstruct the pitch periodicity if the estimator of the model parameters α used to generate A do not account for long-term redundancies.

We compared the different methods presented in Table 2 in the reconstruction approach presented in (73) to estimate the predictor used to generate the matrix A. Differently from Section 5.1, in this section we measured the reconstruction using the mean opinion score (MOS), as calculated through POLQA [78], to account for perceptual qualities as well.

In this experiment we targeted both voiced and unvoiced speech, in order to provide proof of the overall robustness of the sparse linear predictor introduced. This is different from Section 5.1, where we targeted uniquely voiced speech as a proof of concept. The comparison was carried out for missing segment length T_{gap} of 4, 6, 8, 10, and 20 ms, respectively, 64, 96, 128, 160, and 320 samples at 16 kHz. We process 1000 sentences coming from several different speakers with different characteristics (gender, age, pitch, regional accent) taken from the TIMIT database. We applied a robust speech activity detector to avoid applying the reconstruction and calculating statistics over silence. For each file, the losses were produced every 150 ms, the 40 ms (640 samples) before the loss are used to generate the known vector x_{k}, while the varying length gap forms the unknown vector x_{u}. The predictor was calculated with the methods presented in Table 2 on the known vector x_{k}, the unknown segment was then reconstructed using (77). For comparison, we have also added a traditional low-order method **sLP** of order 20.

The results shown in Table 4 gave a different perspective on the performance of the different predictors. While the prediction gain results of Table 3 showed **HOLP** to perform significantly better, in terms of perceptual quality of the reconstructed signal the higher order did not mean higher quality, unless it has actually a clear meaning of representing short-term and long-term redundancies of the speech signal. Thus, **HOSpLP**, **HOSpLPdr** and **HOSpLPadmm** performed better being more accurate and avoiding overfitting the data. It was interesting to notice that **LTP3** and **LTP3j** also performed fairly closely to **HOLP**. The **sLP** method performance were close to the other 2-norm based methods for the smallest gap size, however they decay quite rapidly as the gap

size increased. Finally, we noted that, differently from other interpolation approaches that involve a Estimation-Maximization (EM) approach to enhance the estimation of the AR model and reconstructed signal, while a slight increase in mean-squared error was achieved, no improvement in MOS was actually seen.

METHOD	T_{GAP} [ms]				
	4	6	8	10	20
sLP	3.92±0.09	3.15±0.15	2.96±0.16	2.30±0.18	1.71±0.22
LTP	4.13±0.07	3.44±0.14	3.17±0.12	2.71±0.09	2.45±0.13
LTP3	4.17±0.07	3.53±0.09	3.22±0.13	2.92±0.12	2.63±0.09
LTPj	4.12±0.05	3.63±0.12	3.31±0.12	3.00±0.11	2.75±0.16
HOLP	4.27±0.04	3.55±0.06	3.34±0.08	2.91±0.09	2.61±0.11
HOSpLPip	4.34±0.03	3.75±0.05	3.56±0.08	3.27±0.09	3.12±0.15
HOSpLPdr	4.34±0.02	3.74±0.08	3.55±0.07	3.27±0.11	3.12±0.12
HOSpLPadmm	4.31±0.04	3.69±0.07	3.54±0.07	3.24±0.08	3.11±0.11

Table 4: Average MOS for speech reconstruction with different gap size losses. The 40 ms before the loss are known and used in the reconstruction framework (77). A 95% confidence interval is shown.

6. Discussion and Conclusions

We presented algorithms suitable for finding approximate solutions to the high-order sparse linear prediction problem in speech applications. In particular, we pointed out that a lower accuracy and slower convergence did not affect the overall performance of the predictors when applied in realistic applications that required both objective and subjective quality metrics to be met. The resilience to this approximation could be explained from the actual solution needed by the problem being actually different from the "true" 1-norm solution found by the interior point method. We are indeed looking for a residual and predictor that are "small" and with sparse structures and not particularly for the 1-norm solution. Thus, further work can include better sparse approximations,

rather than seeking more accurate convergence methodologies, as done with, e.g., interior point methods.

Acknowledgement

The authors would like to thank Prof. Søren Holdt Jensen for pointing them to superfast solvers for Toeplitz systems.

References

[1] P. P. Vaidyanathan, The theory of linear prediction, Synthesis Lectures on Signal Processing, Morgan & Claypool Publishers, 2009.

[2] J. Makhoul, Linear prediction: a tutorial review, Proceedings of the IEEE, 63 (4), pp. 561–580, 1975.

[3] T. Bäckström, Linear predictive modelling of speech: constraints and line spectrum pair decomposition, Doctoral dissertation, Helsinki University of Technology, 2004.

[4] J. H. L. Hansen, J. G. Proakis, J. R. Deller Jr, Discrete-time processing of speech signals, Prentice Hall, 1987.

[5] P. Stoica, R. L. Moses, Spectral analysis of signals, Prentice Hall, 2005.

[6] M. G. Christensen, A. Jakobsson, Multi-pitch estimation, Synthesis Lectures on Speech & Audio Processing, Morgan & Claypool Publishers, 2009.

[7] B. S. Atal, S. L. Hanauer, Speech analysis and synthesis by linear prediction of the speech wave, The Journal of the Acoustical Society of America, 50 (2B), pp. 637–655, 1971.

[8] S. Saito, F. Itakura, Theoretical consideration of the statistical optimum recognition of the spectral density of speech, The Journal of the Acoustical Society of Japan, Jan. 1967.

[9] F. Itakura, S. Saito, A statistical method for estimation of speech spectral density and formant frequencies, Electronics and Communications in Japan, 53 (A), pp. 36–43, 1970.

[10] M. N. Murthi, B. D. Rao, All-pole modeling of speech based on the minimum variance distortionless response spectrum, IEEE Transactions on Speech and Audio Processing, 8 (3), pp. 221–239, 2000.

[11] B. S. Atal, The history of linear prediction, IEEE Signal Processing Magazine, 23 (2), pp. 154–161, 2006.

[12] A. Gersho, R. M. Gray, Vector Quantization and Signal Compression, Springer, 1992.

[13] A. El-Jaroudi, J. Makhoul, Discrete all-pole modeling, IEEE Transactions on Signal Processing, 39 (2), pp. 411–423, 1991.

[14] L. A. Ekman, W. B. Kleijn, M. N. Murthi, Regularized linear prediction of speech, IEEE Transactions on Speech, Audio, and Language Processing, 16 (1), pp. 65–73, 2008.

[15] C.-H. Lee, On robust linear prediction of speech, IEEE Transactions on Acoustics, Speech and Signal Processing, 36 (5), pp. 642–650, 1988.

[16] E. Denoel, J.-P. Solvay, Linear prediction of speech with a least absolute error criterion, IEEE Transactions on Acoustics, Speech and Signal Processing, 33 (6), pp. 1397–1403, 1985.

[17] H. Hermansky, Perceptual linear predictive (PLP) analysis of speech, The Journal of the Acoustical Society of America, 87 (4), pp. 1738–1752, 1990.

[18] C. Magi, J. Pohjalainen, T. Bäckström, P. Alku, Stabilised weighted linear prediction, Speech Communication, 51 (5), pp. 401–411, 2009.

[19] J. Thyssen, H. Nielsen, S. D. Hansen, Non-linear short-term prediction in speech coding, in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. I/185–I/188, 1994.

[20] H. Kameoka, N. Ono, S. Sagayama, Speech spectrum modeling for joint estimation of spectral envelope and fundamental frequency, IEEE Transactions on Speech, Audio, and Language Processing, 18 (6), pp. 1507–1516, 2010.

[21] S. Bensaid, D. Slock, Blind audio source separation exploiting periodicity and spectral envelopes, in Proc. of International Workshop on Acoustic Signal Enhancement, pp. 1–4, 2012.

[22] P. Kabal, R. Ramachandran, Joint optimization of linear predictors in speech, IEEE Transactions on Acoustics, Speech and Signal Processing, 37 (5), pp. 642–650, 1989.

[23] D. Giacobello, M. G. Christensen, J. Dahl, S. H. Jensen, M. Moonen, Sparse linear predictors for speech processing, in Proc. of the Annual Conference of the International Speech Communication Association, pp. 1353–1356, 2008.

[24] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, M. Moonen, Sparse linear prediction and its applications to speech processing, IEEE Transactions on Speech, Audio, and Language Processing, 20 (5), pp. 1644–1657, 2012.

[25] M. Murthi, B. Rao, Towards a synergistic multistage speech coder, in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. I/369–I/372, 1998.

[26] D. Giacobello, M. G. Christensen, J. Dahl, S. H. Jensen, M. Moonen, Joint estimation of short-term and long-term predictors in speech coders, in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, 2009, pp. 4109–4112.

[27] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, M. Moonen, Speech coding based on sparse linear prediction, in Proc. European Signal Processing Conference, pp. 2524–2528, 2009.

[28] I. Erer, K. Sarikaya, H. Bozkurt, Enhanced radar imaging via sparsity regularized 2D linear prediction, in Proc. European Signal Processing Conference, pp. 1751–1755, 2014.

[29] N. Bochud, A. Gomez, G. Rus, A. Peinado, Sparse signal model for ultrasonic nondestructive evaluation of CFRP composite plates, in Proc. IEEE

International Conference on Acoustics, Speech, and Signal Processing, pp. 2844–2847, 2013.

[30] J. Koloda, J. Østergaard, S. H. Jensen, V. E. Sánchez, A. M. Peinado, Sequential error concealment for video/images by sparse linear prediction, IEEE Transactions on Multimedia, 15 (4), pp. 957–969, 2013.

[31] D. Angelosante, G. Giannakis, N. Sidiropoulos, Sparse parametric models for robust nonstationary signal analysis: leveraging the power of sparse regression, IEEE Signal Processing Magazine, 30 (6), pp. 64–73, 2013.

[32] D. Angelosante, Sparse regressions for joint segmentation and linear prediction, in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 335–339, 2014.

[33] Y. Nesterov, A. Nemirovskii, Interior-Point Polynomial Methods in Convex Programming, SIAM, 1994.

[34] E. D. Andersen, C. Roos, T. Terlaky, On implementing a primal-dual interior-point method for conic quadratic optimization, Mathematical Programming Series B, 95 (2), pp. 249–277, 2003.

[35] J. Mattingley, S. Boyd, Real-time convex optimization in signal processing, IEEE Signal Processing Magazine, 27 (3), pp. 50–61, 2010.

[36] B. Defraene, T. van Waterschoot, H. J. Ferreau, M. Diehl, M. Moonen, Real-time perception-based clipping of audio signals using convex optimization, IEEE Transactions on Speech, Audio, and Language Processing, 20 (10) pp. 2657–2671, 2012.

[37] T. Jensen, D. Giacobello, M. Christensen, S. Jensen, M. Moonen, Real-time implementations of sparse linear prediction for speech processing, in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 8184–8188, 2013.

[38] B. Defraene, T. van Waterschoot, M. Diehl, M. Moonen, Embedded-optimization-based loudspeaker precompensation using a Hammerstein loudspeaker model, IEEE/ACM Transactions on Speech, Audio, and Language Processing, 22 (11), pp. 1648–1659, 2014.

[39] G. Alipoor, M. H. Savoji, Wide-band speech coding based on bandwidth extension and sparse linear prediction, in Proc. International Conference on Telecommunications and Signal Processing, pp. 454–459, 2012.

[40] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers, Foundations and Trends in Machine Learning, 3 (1), pp. 1–122, 2011.

[41] L. Vandenberghe, Douglas-Rachford method and ADMM, Optimization methods for large-scale systems, Lecture notes, University of California, Los Angeles (UCLA), 2013.

[42] J. Douglas, H. H. Rachford, On the numerical solution of heat conduction problems in two and three space variables, Transactions of the American Mathematical Society, 82, pp. 421–439, 1956.

[43] P. L. Lions, B. Mercier, Splitting algorithms for the sum of two nonlinear operators, SIAM Journal on Numerical Analysis, 16 (6), pp. 964–979, 1979.

[44] P. L. Combettes, J.-C. Pesquet, A Douglas–Rachford splitting approach to nonsmooth convex variational signal recovery, IEEE Journal of Selected Topics in Signal Processing, 1 (4), pp. 564–574, 2007.

[45] D. O'Connor, L. Vandenberghe, Primal-dual decomposition by operator splitting and applications to image deblurring, SIAM Journal on Imaging Sciences, 7 (3), pp. 1724–1754, 2014.

[46] R. Glowinski, A. Marroco, Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problémes de Dirichlet non linéaires, ESAIM: Mathematical Modelling and Numerical Analysis - Modélisation Mathématique et Analyse Numérique, 9 (R2), pp. 41–76, 1975.

[47] D. Gabay, B. Mercier, A dual algorithm for the solution of nonlinear variational problems via finite element approximation, Computers & Mathematics with Applications, 2 (1), pp. 17 – 40, 1976.

[48] M. Afonso, J. Bioucas-Dias, M. Figueiredo, Fast image recovery using variable splitting and constrained optimization, IEEE Transactions on Image Processing, 19 (9), pp. 2345–2356, 2010.

[49] J. Yang, Y. Zhang, Alternating direction algorithms for ℓ_1-problems in compressive sensing, SIAM Journal of Scientific Computing 33 (1), pp. 250–278, 2011.

[50] D. Gabay, Applications of the method of multipliers to variational inequalities, in M. Fortin, F. Glowinski (Eds.), Augmented Lagrangian methods: Applications to the Numerical Solution of Boundary-Value Problems, Studies in Mathematics and Its Applications, pp. 299–340, Springer-Verlag, 1983.

[51] J. Eckstein, D. Bertsekas, On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators, Mathematical Programming, 55 (1–3), pp. 293–318, 1992.

[52] W. Yin, S. Osher, D. Goldfarb, J. Darbon, Bregman iterative algorithms for ℓ_1-minimization with application to compressed sensing, SIAM Journal on Imaging Sciences, 1 (1), pp. 143–168, 2008.

[53] T. van Waterschoot, M. Moonen, Comparison of linear prediction models for audio signals, EURASIP Journal on Audio, Speech, and Music Processing, 2008.

[54] S. M. Kay, The effects of noise on the autoregressive spectral estimator, IEEE Transactions on Acoustics, Speech and Signal Processing, 27 (5), pp. 478–485, 1979.

[55] D. L. Donoho, M. Elad, Optimally sparse representation in general (nonorthogonal) dictionaries via $\ell 1$ minimization, Proceedings of the National Academy of Sciences of the United States of America, 4 (5), pp. 2197–2202, 2002.

[56] Y. Nesterov, Gradient methods for minimizing composite objective function, Center for Operations Research and Econometrics (CORE) Discussion Papers no. 2007/76, Université Catholique de Louvain, 2007.

[57] A. Beck, M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, SIAM Journal of Imaging Sciences, 2 (1), pp. 183–202, 2009.

[58] S. J. Wright, R. Nowak, M. A. T. Figueiredo, Sparse reconstruction by separable approximation, IEEE Transaction on Signal Processing, 57 (7), pp. 2479–2493, 2009.

[59] J. J. Moreau, Proximité et dualité dans un espace hilbertien, Bulletin de la Société Mathématique de France, 93, pp. 273–299, 1965.

[60] P. L. Combettes, V. R. Wajs, Signal recovery by proximal forward-backward splitting, Multiscale Modeling & Simulation, 4, pp. 1168–1200, 2005.

[61] J. Spingarn, Applications of the method of partial inverses to convex programming: decomposition, Mathematical Programming, 32 (2), pp. 199–223, 1985.

[62] N. Levinson, The Weiner RMS error criterion in filter design and prediction, Journal of Mathematics and Physics, MIT, 25, pp. 261–278, 1947.

[63] G. H. Golub, C. F. van Loan, Matrix Computations, 4th Edition, Johns Hopkins University Press, 2013.

[64] R. Bitmead, B. Anderson, Asymptotically fast solution of Toeplitz and related systems of linear equations, Linear Algebra and its Applications, 34, pp. 103–116, 1980.

[65] G. Ammar, W. Gragg, Superfast solution of real positive definite Toeplitz systems, SIAM Journal on Matrix Analysis and Applications 9 (1), pp. 61–76, 1988.

[66] J. R. Jain, An efficient algorithm for a large Toeplitz set of linear equations, IEEE Transactions on Acoustics, Speech and Signal Processing, 27 (6), pp. 612–615, 1979.

[67] Intel(R) Math Kernel Library, `https://software.intel.com/en-us/intel-mkl`

[68] M. Frigo, S. G. Johnson, The design and implementation of FFTW3, Proceedings of the IEEE, 93 (2), pp. 216–231, 2005.

[69] M. Grant, S. Boyd, CVX: Matlab software for disciplined convex programming, version 1.21, `http://cvxr.com/cvx/`, 2011.

[70] J. F. Sturm, Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones, Optimization Methods and Software, 11–12, pp. 625–653, 1999.

[71] J. Mattingley, S. Boyd, CVXGEN: A code generator for embedded convex optimization, Optimization and Engineering, 13 (1), pp. 1–27, 2012.

[72] P. Patrinos, L. Stella, A. Bemporad, Douglas-Rachford splitting: complexity estimates and accelerated variants, in Proc. IEEE Conference on Decision and Control, pp. 4234–4239, 2014.

[73] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, TIMIT acoustic-phonetic continuous speech corpus, NIST Interagency/Internal Report (NISTIR) - 4930, 1993.

[74] S. Godsill, P. Rayner, Digital audio restoration, Springer, 1998.

[75] A. K. Halberstadt, J. R. Glass, Heterogeneous acoustic measurements for phonetic classification, in Proc. European Conference on Speech Communication and Technology, pp. 401–404, 1997.

[76] B. Bessette, R. Salami, R. Lefebvre, M. Jelinek, J. Rotola-Pukkila, J. Vainio, H. Mikkola, K. Jarvinen, The adaptive multirate wideband speech codec (AMR-WB), IEEE Transactions on Speech and Audio Processing, 10 (8), pp. 620–636, 2002.

[77] P. C. Hansen, D. P. O'Leary, The use of the L-curve in the regularization of discrete ill-posed problems, SIAM Journal on Scientific Computing 14 (6), pp. 1487–1503, 1993.

[78] ITU-T, Perceptual Objective Listening Quality Assessment, P.863, 2010.