**Aalborg Universitet**

AALBORG UNIVERSITY
DENMARK

# Stable 1-Norm Error Minimization Based Linear Predictors for Speech Modeling

Giacobello, Daniele; Christensen, Mads Græsbøll; Jensen, Tobias Lindstrøm; Murthi, Manohar N.; Jensen, Søren Holdt; Moonen, Marc

*Published in:*
I E E E Transactions on Audio, Speech and Language Processing

*DOI (link to publication from Publisher):*
10.1109/TASLP.2014.2311324

*Publication date:*
2014

*Document Version*
Accepted author manuscript, peer reviewed version

Link to publication from Aalborg University

*Citation for published version (APA):*
Giacobello, D., Christensen, M. G., Jensen, T. L., Murthi, M. N., Jensen, S. H., & Moonen, M. (2014). Stable 1-Norm Error Minimization Based Linear Predictors for Speech Modeling. *I E E E Transactions on Audio, Speech and Language Processing*, *22*(5), 912-922. https://doi.org/10.1109/TASLP.2014.2311324

**General rights**
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal -

**Take down policy**
If you believe that this document breaches copyright please contact us at vbn@aub.aau.dk providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from vbn.aau.dk on: March 29, 2024

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2014.2311324



# Stable 1-norm error minimization based linear predictors for speech modeling

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

***Abstract*—In linear prediction of speech, the 1-norm error minimization criterion has been shown to provide a valid alternative to the 2-norm minimization criterion. However, unlike 2-norm minimization, 1-norm minimization does not guarantee the stability of the corresponding all-pole filter and can generate saturations when this is used to synthesize speech. In this paper, we introduce two new methods to obtain intrinsically stable predictors with the 1-norm minimization. The first method is based on constraining the roots of the predictor to lie within the unit circle by reducing the numerical range of the shift operator associated with the particular prediction problem considered. The second method uses the alternative Cauchy bound to impose a convex constraint on the predictor in the 1-norm error minimization. These methods are compared with two existing methods: the Burg method, based on the 1-norm minimization of the forward and backward prediction error, and the iteratively reweighted 2-norm minimization known to converge to the 1-norm minimization with an appropriate selection of weights. The evaluation gives proof of the effectiveness of the new methods, performing as well as unconstrained 1-norm based linear prediction for modeling and coding of speech.**

## I. Introduction

Linear Prediction (LP) is widely used in a diverse range of speech modeling based algorithms (e.g., coding and recognition [1]). The traditional approach is to find the prediction coefficients by minimizing the 2-norm of the prediction error, i.e., the difference between the predicted and observed signal. This works well when the excitation signal is i.i.d. Gaussian [2]; however, when this assumption is not satisfied, problems arise. This is the case for voiced speech where the pitch excitation is sparse and pulse-like. In this case, an alternative approach based on the 1-norm minimization of the prediction error has shown to offer a better modeling thanks to its ability to decouple the pitch excitation from the vocal tract transfer function [3].

The improved modeling of 1-norm minimization also ha shown to be beneficial in speech coding. In particular, when seeing the 1-norm as a convex relaxation of the 0-norm, the minimization process offers a residual that is *sparser*, providing tighter coupling between the multiple stages of time-domain speech coders and thereby enabling a more efficient coding [4]–[6]. Nevertheless, unlike those obtained through 2-norm minimization, the predictors obtained through 1-norm minimization are not intrinsically stable [7], [8] and, in applications such as coding, having unstable filters may generate saturations in the synthesized speech. In particular, as noted in [3] for a large set of data, the percentage of unstable filters in voiced speech is around 10%.

The predictor stability problem in 1-norm LP has been tackled already in [8] by introducing the Burg method for all-pole parameters estimation based on 1-norm minimization of the forward and backward error. In this approach, however, the sparsity is not preserved [3]. In this paper, we will introduce two novel methods to obtain intrinsically stable predictors with the 1-norm minimization. The first method is based on modifying the shift operator that generates the observation matrix from the analyzed speech segment [9], reducing the numerical range of this matrix [10]. This allows us to restrict the zeros of the predictor polynomial to lie within the unit circle. A similar approach has been used in weighted LP [11], [12] to modify the weighting function to guarantee stable solutions. The second method uses the alternative Cauchy bound [13], [14] to impose a constraint on the predictor in the 1-norm error minimization.

The paper is organized as follows. In Section II, we provide a brief review of LP based on the $p$-norm. In Section III, the core of the paper, we introduce our two new methods to obtain intrinsically stable predictors with the 1-norm minimization and also review the existing ones. In Section IV, we compare the spectral modeling and coding performances of the resulting predictors. In Section V, we provide a complexity analysis and possible efficient solutions for the method presented, as initially introduced in [15] for the 1-norm LP. Finally, Section VI concludes the paper.

Copyright (c) 2013 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org.

Daniele Giacobello is with Beats Electronics, LLC, Santa Monica, CA 90404, USA (e-mail: giacobello@ieee.org).

Mads Græsbøll Christensen is with the Audio Analysis Lab, Department of Architecture, Design, and Media Technology, Aalborg Universitet, 9220 Aalborg, Denmark (email: mgc@create.aau.dk).

Tobias Lindstrøm Jensen and Søren Holdt Jensen are with the Department of Electronic Systems, Aalborg Universitet, 9220 Aalborg, Denmark (e-mail: {tlj,shj}@es.aau.dk).

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

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

The work of Daniele Giacobello was supported by the Marie Curie EST-SIGNAL Doctoral Fellowship, contract no. MEST-CT-2005-021175 and was carried out at the Department of Electronic Systems, Aalborg Universitet, 9220 Aalborg, Denmark.

Copyright (c) 2014 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2014.2311324

## II. Fundamentals of Linear Prediction

The problem considered in this paper is based on the following speech production model, where a sample of speech $x(n)$ at time $n$ is written as a linear combination of $K$ past samples:

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

where $\{a_k\}$ are the coefficients of the predictor

$$A(z) = 1 + \sum_{k=1}^{K} a_k z^{-k}, \tag{2}$$

and $e(n)$ is the driving noise process (also referred to as prediction residual or excitation). The speech production model (1) in matrix form becomes:

$$\mathbf{x} = \mathbf{X}\mathbf{a} + \mathbf{e}. \tag{3}$$

The problem considered in this paper is 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 [16]:

$$\hat{\mathbf{a}} = \underset{\mathbf{a}}{\operatorname{argmin}} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_p^p, \tag{4}$$

where

$$\mathbf{x} = \begin{bmatrix} x(N_1) & \ldots & x(N_2) \end{bmatrix}^T, \tag{5}$$

$$\mathbf{X} = \begin{bmatrix} x(N_1-1) & \cdots & x(N_1-K) \\ \vdots & & \vdots \\ x(N_2-1) & \cdots & x(N_2-K) \end{bmatrix}, \tag{6}$$

and $\|\cdot\|_p$ is the p-norm defined as $\|\mathbf{x}\|_p = (\sum_{n=1}^{N} |x(n)|^p)^{\frac{1}{p}}$ for $p \geq 1$. The starting and ending points $N_1$ and $N_2$ can be chosen in various ways assuming that $x(n) = 0$ for $n < 1$ and $n > N$ [17]. We will consider the case $N_1 = 1$ and $N_2 = N + K$, which for $p = 2$ is equivalent to the autocorrelation method:

$$\hat{\mathbf{a}} = \underset{\mathbf{a}}{\operatorname{argmin}} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_2^2 = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{x}, \tag{7}$$

where $\mathbf{R} = \mathbf{X}^T\mathbf{X}$ is the autocorrelation matrix.

The case we consider here is when $p = 1$, which corresponds to minimizing the sum of absolute values:

$$\hat{\mathbf{a}} = \underset{\mathbf{a}}{\operatorname{argmin}} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_1. \tag{8}$$

This formulation is relevant particularly in LP of voiced speech signals where the prediction residual is usually modeled by an impulse train. The 1-norm, intended as a convex relaxation of the 0-norm, will offer an approximate solution to the minimization of the cardinality, i.e., the *sparsest* prediction residual:

$$\hat{\mathbf{a}} = \underset{\mathbf{a}}{\operatorname{argmin}} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_0. \tag{9}$$

This translates into the ability of the predictor to preserve the structure of the underlying sparse pulse-like excitation. The spectral envelope will benefit from this by avoiding the over-emphasis on peaks generated in the effort to cancel the voiced speech harmonics [3], [8].

The 1-norm minimization criterion is also equivalent to the Maximum-Likelihood (ML) estimator when the prediction error is assumed to be i.i.d. Laplacian:

$$\hat{\mathbf{a}}_{\mathbf{ML}} = \underset{\mathbf{a}}{\operatorname{argmax}}\, f(\mathbf{x}|\mathbf{a}) = \underset{\mathbf{a}}{\operatorname{argmax}} \{\exp(-\|\mathbf{x} - \mathbf{X}\mathbf{a}\|_1)\}. \tag{10}$$

A multivariate Laplacian distribution can be seen as generated by an autoregressive (AR) model excited by a sequence of i.i.d. univariate Laplacian samples [18]–[20]. However, a rigorous proof cannot be obtained since the Laplacian distribution does not have a closed form solution [21]. This conjecture statistically justifies the use of the 1-norm in modeling the excitation, given that it is well known that a multivariate Laplacian distribution offers a better model for a speech signal segment [22].

The minimization problem in (8) does not allow for a closed form solution and so a linear programming formulation is required [16]. In particular, interior point methods [23] have been proven to solve the minimization problem efficiently [15].

## III. Methods for obtaining stable predictors

### A. Reducing the numerical range of the shift operator

First of all, we consider a known general framework for linear prediction, successfully implemented in [11] and [12] for the analysis of voiced speech. The columns of the matrix obtained concatenating $\mathbf{x}$ and $\mathbf{X}$, as defined in (6)

$$[\mathbf{x}|\mathbf{X}] = [\mathbf{x}_0\ \mathbf{x}_1\ \ldots\ \mathbf{x}_K] \in \mathbb{R}^{(N+K)\times(K+1)} \tag{11}$$

can be generated via the formula:

$$\mathbf{x}_{k+1} = \mathbf{B}\mathbf{x}_k, \tag{12}$$

where

$$\mathbf{x}_0 = [x_1\ x_2\ \ldots\ x_N\ 0\ \ldots\ 0]^T \in \mathbb{R}^{N+K}, \tag{13}$$

and $\mathbf{B}$ is a a noncirculant shift matrix of size $(N+K)\times(N+K)$:

$$\mathbf{B} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}. \tag{14}$$

Applied to $\mathbf{x}$, $\mathbf{B}$ shifts the elements down by one position and eliminates the last element. In other words, $\mathbf{B}$ is a nilpotent operator of power $n = N + K$, i.e., $\mathbf{B}^{N+K} = 0$.

Let us now consider the $p$-norm LP problem (4), where the column $[\mathbf{x}|\mathbf{X}]$ are constructed using the formula in (12) where $\mathbf{B}$ is generalized to any matrix in $\mathbb{R}^{(N+K)\times(N+K)}$. It has been shown that, in this case, the roots $\{z_i\}$ of the monic polynomial solution to the $p$-norm LP problem (4) belong to the numerical range $\eta_p(\mathbf{B})$ of the matrix $\mathbf{B}$, which, in turn, belongs to an open circular disk $\rho(\mathbf{B})$ of radius $2\|\mathbf{B}\|_2$ and center in the origin [9]. It is then clear that the roots of the predictor, obtained by solving (8), with $\mathbf{B}$ as defined in (14), will be contained in a closed circle of radius $2\|\mathbf{B}\|_2 = 2$. This

Copyright (c) 2014 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2014.2311324

result can be generalized for any shift matrix $\mathbf{B}$ with nonzero entries different from the unity:

$$\mathbf{B} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ B_{2,1} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & 0 & B_{N+K,N+K-1} & 0 \end{bmatrix}. \quad (15)$$

In this case, the radius of the circle $\rho(\mathbf{B})$ that contains the numerical range $\eta_1(\mathbf{B})$ is calculated as:

$$2\|\mathbf{B}\|_2 = 2\max|B_{i+1,i}|. \quad (16)$$

We will then change the nonzero values of $\mathbf{B}$ (and subsequently the construction of $[\mathbf{x}|\mathbf{X}]$) in order to reduce the radius of the circle containing $\eta_1(\mathbf{B})$ to be equal or less than one, therefore guaranteeing the stability of the linear predictor. In particular, having $\max|B_{i,j}| \leq 1/2$ will be sufficient for stability. We can also consider a more general formulation of the LP scheme, where we apply a weighting vector $\mathbf{w} \in \mathbb{R}_+^{N+K}$ on the analyzed speech signal segment. The effect of the weighting can be moved into the shift matrix and the analyzed speech segment by defining:

$$\tilde{\mathbf{B}} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ w_2/w_1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & 0 & w_{N+K}/w_{N+K-1} & 0 \end{bmatrix}, \quad (17)$$

and

$$\tilde{\mathbf{x}}_0 = [w_1x_1 \; w_2x_2 \; w_Nx_N \; 0 \; \ldots \; 0]^T. \quad (18)$$

Constructing all the other columns of the new matrix $[\tilde{\mathbf{x}}|\tilde{\mathbf{X}}]$ using the relation in (12), the minimization problem (8) then becomes:

$$\min_{\mathbf{a}} \|\tilde{\mathbf{x}} - \tilde{\mathbf{X}}\mathbf{a}\|_1. \quad (19)$$

According to (16), the circle containing the numerical range of $[\tilde{\mathbf{x}}|\tilde{\mathbf{X}}]$ and, in turn, the roots of the predictor will have radius:

$$\rho(\tilde{\mathbf{B}}) = 2\max\frac{w_{n+1}}{w_n}. \quad (20)$$

We can then construct a weighting vector that stabilizes the predictor. In [11] and [12], the weighting vector is chosen based on the short-time energy (STE):

$$w_n = \sqrt{\sum_{i=0}^{M-1} x_{n-i-1}^2} \quad (21)$$

where $M$ is the length of the STE window. The STE window tends to more heavily weight those parts of the speech signal that consist of samples of large magnitude, providing a robust signal selection especially for the analysis of voiced speech. In order to achieve intrinsically stable solutions, we can then simply define the entries of the matrix $\tilde{\mathbf{B}}$ in (17) as:

$$\tilde{B}_{i+1,i} = \begin{cases} (w_{i+1}/w_i) & \text{if} \quad (w_{i+1}/w_i) \leq 1/2, \\ 1/2 & \text{if} \quad (w_{i+1}/w_i) > 1/2. \end{cases} \quad (22)$$

Finally, we can solve our modified 1-norm problem in (19) obtaining an intrinsically stable predictor. Clearly, the window, and thus the weights, can be chosen *ad libitum*; we will use the STE windowing that provides important signal selection properties to retrieve the underlying spiky structure of the speech signal, as done in [12].

**Algorithm 1** Iteratively Reweighted 2-norm Minimization

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. $\hat{\mathbf{a}}^i \leftarrow \text{argmin}_{\mathbf{a}} \|\mathbf{W}^i(\mathbf{x} - \mathbf{X}\mathbf{a})\|_2^2$
2. $\mathbf{W}^{i+1} \leftarrow \text{diag}\left(\left|\mathbf{x} - \mathbf{X}\hat{\mathbf{a}}^i\right| + \epsilon\right)^{-1/2}$
3. $i \leftarrow i + 1$
**end while**

### B. Constrained 1-norm minimization

We will now consider another method to constrain the roots of the predictor within the unit circle. Let us consider the univariate polynomial $A(z)$ in (2). According to [24], the alternative Cauchy bound states that all zeros of (2) lie in the disk:

$$|z| \leq \lambda, \quad \text{where} \quad \lambda = \max\left\{1, \sum_{k=1}^{K}|a_k|\right\}. \quad (23)$$

This bound, a refinement of the famous Cauchy bound [13], gives precious hints on how to modify the formulation of (8) to guarantee a stable predictor. In particular, we can rewrite the problem as:

$$\begin{aligned} \underset{\mathbf{a}}{\text{minimize}} \quad & \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_1 \\ \text{subject to} \quad & \|\mathbf{a}\|_1 < 1 \end{aligned} \quad (24)$$

where the constraint $\|\mathbf{a}\|_1 < 1$, according to (23), provides a sufficient (not necessary) condition for the zeros of (2) to belong to the open unit disk, and can be easily incorporated in the linear program to solve (8) [16].

### C. Iteratively Reweighted 2-norm minimization [25]

Now let us consider some previously proposed methods for obtaining stable solutions. A known method to obtain a stable predictor based on 1-norm minimization is based on iteratively reweighted 2-norm minimization [25]. The algorithm is shown in Algorithm 1. It is guaranteed to output a stable predictor since the only difference to the original formulation is the projection in the weighted domain by the matrix $\mathbf{W}^i$, leaving $\mathbf{x}$ and $\mathbf{X}$ untouched, as discussed in Section III-A. In [25], a proof that $\|\hat{\mathbf{r}}^{i+1}\|_2 \leq \|\hat{\mathbf{r}}^i\|_2$ (where $\hat{\mathbf{r}}^i = \mathbf{x} - \mathbf{X}\hat{\mathbf{a}}^i$) is provided, meaning that this is a descent algorithm. In Algorithm 1, the halting criterion can be chosen as either a maximum number of iterations or as a convergence criterion. The parameter $\epsilon > 0$ is used to avoid problems when a component of $\hat{\mathbf{r}}$ goes to zero. The weighting with the square root of the inverse of the amplitude of the residual increases the influence of the small values in the residual while the influence of the large residual values decreases, which is consistent with the Laplacian probability density functions (8).

Copyright (c) 2014 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2014.2311324

**Algorithm 2** 1-norm Burg Method

Inputs: speech segment $\mathbf{x}$
Outputs: reflection coefficients $\{k_i\}$
Initialize forward $\mathbf{f}_0 = \mathbf{x}$ and backward $\mathbf{b}_0 = \mathbf{x}$ error
**for** $i = 1, \ldots, K$ **do**
1. $k_i \leftarrow \text{argmin}_{k_i} \|\mathbf{f}_{i-1} + k_i \mathbf{b}_{i-1}\|_1 + \|k_i \mathbf{f}_{i-1} + \mathbf{b}_{i-1}\|_1$
update forward error
2. $f_i(n) \leftarrow f_{i-1}(n) + k_i b_{i-1}(n-1)$
update backward error
3. $b_i(n) \leftarrow k_i f(n)_{i-1} + b_{i-1}(n-1)$
**end for**

### D. Burg method based on 1-norm minimization [8]

The Burg method based on 1-norm minimization was first proposed in [8]. This method stands as a generalization of the Burg method where the reflection coefficients of the lattice filter are obtained by minimizing the 1-norm of the forward and backward prediction error instead of the 2-norm. The algorithm is shown in Algorithm 2. Once the $K$ reflection coefficients are found, the prediction polynomial and the prediction error can be easily calculated. This method is also guaranteed to provide a stable predictor since all the reflection coefficients obtained have amplitude less than one. A simple proof is shown in [8]. This method is, however, suboptimal due to the decoupling of the main $K$-dimensional minimization problem (8) in $K$ one-dimensional minimization sub-problems.

## IV. Performance Analysis

In this section, we analyze and compare the performance of the stable predictors obtained with the methods presented in the previous section with traditional 2-norm LP and 1-norm LP. An overview of the methods compared and the acronyms used through the section are shown in Table I. In the case of 1-norm LP, a stability check takes place once the predictor is obtained and the stabilization is performed through pole reflection when the predictor is unstable. Notice that pole reflection is the only way to obtain an amplitude response for the stabilized predictor that is exactly the same as the one of the unstable predictor. In all other methods, no stability check has to be performed.

### A. Modeling Performance

In this section, we analyze the modeling performance of the predictors in the case of voiced speech. The experimental analysis was done on 5,000 segments of length $N = 40$ (5 ms) of clean voiced speech coming from several different speakers with different characteristics (gender, age, pitch, regional accent) taken from the TIMIT database, downsampled to 8 kHz.

*1) Spectral Envelope Modeling:* 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 [26] and provides an approximation of the vocal tract transfer function, "cleaned" from the fine structure belonging to the pitch excitation. We then calculated the log spectral distortion (SD) between our reference envelope $S_{int}(\omega)$ and the estimated model of the all-pole model corresponding to the inverse of the predictor $S(\omega, \mathbf{a})$ as:

$$\text{SD}_m = \sqrt{\frac{1}{2\pi} \int_{-\pi}^{\pi} [10 \log_{10} S_{int}(\omega) - 10 \log_{10} S(\omega, \mathbf{a})]^2 \, d\omega}. \quad (25)$$

In general, the linear predictors obtained through 1-norm minimization provide smoother all-pole models of the vocal tract and are therefore more robust to quantization. We also compared the log spectral distortion between our reference envelope $S_{int}(\omega)$ and the quantized LP model $S(\omega, \hat{\mathbf{a}})$:

$$\text{SD}_q = \sqrt{\frac{1}{2\pi} \int_{-\pi}^{\pi} [10 \log_{10} S_{int}(\omega) - 10 \log_{10} S(\omega, \hat{\mathbf{a}})]^2 \, d\omega}. \quad (26)$$

The quantizer used is the one presented in [27], with the number of bits fixed at 20 for the different prediction orders. A critical analysis of the results in Table II shows how 1-norm based LP (**LP1**) offers substantially better modeling of

TABLE I
Description of the different prediction methods compared in our evaluation.

| METHOD | DESCRIPTION |
|---|---|
| **LP2** | Traditional 2-norm minimization (7) with 10Hz bandwidth expansion ($\gamma = 0.996$) and Hamming windowing. |
| **LP1** | Unconstrained 1-norm minimization (8). Stability is imposed by pole reflection if unstable. No windowing is performed. |
| **STW** | Stable 1-norm minimization through reduction of the numerical range of the shift operator (19). The weigths in (17) and (18) are chosen from the STE (21). |
| **CT1** | Constrained 1-norm minimization as shown in (24). No windowing is performed. |
| **BU1** | Burg method based on the 1-norm minimization of forward and backward error (as shown in Algorithm 2). No windowing is performed. |
| **RW2** | Reweighted 2-norm minimization (as shown in Algorithm 1). No bandwidth expansion is performed. No windowing is performed. |

TABLE II
Average spectral distortion for the considered methods in the unquantized case $\text{SD}_m$ and quantized case $\text{SD}_q$ for different prediction orders $K$. A 95% confidence intervals is given for each value.

| METHOD | $K$ | $\text{SD}_m$ | $\text{SD}_q$ |
|---|---|---|---|
| **LP2** | 10 | 1.97±0.03 | 2.95±0.09 |
| | 12 | 1.98±0.05 | 2.72±0.12 |
| **LP1** | 10 | 1.78±0.01 | 2.53±0.02 |
| | 12 | 1.61±0.01 | 2.31±0.04 |
| **STW** | 10 | 1.71±0.02 | 2.47±0.01 |
| | 12 | 1.52±0.01 | 2.19±0.09 |
| **CT1** | 10 | 1.88±0.01 | 2.64±0.01 |
| | 12 | 1.65±0.01 | 2.22±0.01 |
| **BU1** | 10 | 1.91±0.06 | 2.71±0.09 |
| | 12 | 1.84±0.11 | 2.59±0.10 |
| **RW2** | 10 | 1.83±0.01 | 2.51±0.02 |
| | 12 | 1.69±0.03 | 2.37±0.05 |

Copyright (c) 2014 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2014.2311324

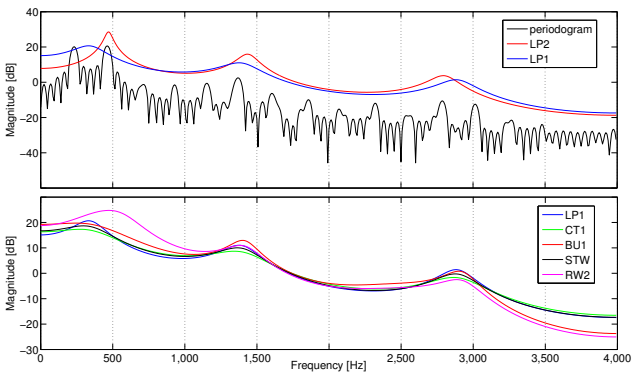



Fig. 1. Figures of typical properties of the spectrum for the methods considered using prediction order $K = 10$. The top pane illustrates an example of the spectral difference between **LP1** and **LP2** for a voiced segment of speech for which the spectrogram is shown. The bottom pane demonstrating the differences between **LP1** and the different approaches for intrinsically stable solution listed in Table I.

the envelope than traditional 2-norm LP (**LP2**). All the other methods achieve a performance similar to **LP1**, but **STW** offers even better modeling performance, thanks also to the choice of weights. It should be noted that **CT1** increases its performances considerably from order $K = 10$ to $K = 12$. This is a direct consequence of the stringent constraint on the prediction coefficients ($\|\mathbf{a}\|_1 < 1$), requiring a larger $K$ to model all the spectral information as well as the other methods.

Examples of the spectral envelopes for the different methods are shown in Figure 1. The top pane clearly illustrates some of the shortcomings of the 2-norm minimization approach, as the overemphasis on peaks of the underlying pitch excitation causes the envelope to have a sharper contour than desired with poles close to the unit circle. The bottom pane shows the similarity in performance of the stable methods presented. It can be seen that **BU1** and **RW2** have a slightly different behavior. The Burg method, in particular, can suffer from spectral line-splitting when there is a mismatch between the true order of the system and the order of the Burg algorithm [28]. In this particular case, a pole is not located around the first formant. Instead, two poles are estimated around it, which makes for a fading spectral lobe. A similar effect happens in **RW2**, where the spectral lobe around the first formant is unusually wide. This is an effect of the ill-conditioning of the autocorrelation matrix $(\mathbf{X}^T\mathbf{X})^{-1}$ in Step 1 of Algorithm 1 after a few reweighting iterations. The eigenvalues might tend to cluster towards certain locations of the autocorrelation matrix, thus generating a higher spread in the eigenvalues and a peakier behavior in the envelope. This can also be seen in the difference between **LP1** and **RW2** in the higher frequencies [29].

*2) Shift Invariance:* Linear predictors obtained with the 1-norm minimization criterion are well documented to be robust to small shifts of the analysis window [3]. In speech analysis, this is a desirable property, since speech, and voiced speech in particular, is assumed to be short-term stationary. The shortcomings of the **LP2** method are a direct consequence of the coupling between the vocal tract transfer function and the underlying pitch excitation that standard LP introduces in the estimate [30]. To analyze the invariance of the LP methods to window shifts, we took the same 5,000 frames of clean voiced speech mentioned above and we expanded them to the left and to the right with 20 samples, giving a total length $N = 80$. In each frame of length $N = 80$ we defined a $M = 40$ samples rectangular window for all methods and we shifted the window by $s = 1, 2, 5, 10, 20$ samples. The average log spectral difference of the $10^{th}$ order LP estimate between $S_0(\omega)$ and $S_s(\omega)$:

$$\mathrm{SD}_s = \sqrt{\frac{1}{2\pi}\int_{-\pi}^{\pi} \left[10\log_{10} S_0(\omega) - 10\log_{10} S_s(\omega, \mathbf{a})\right]^2 d\omega}, \tag{27}$$

was analyzed. The average differences obtained for the methods in Table I are shown in Table III. The results obtained indicate clearly that the 1-norm based predictors' robustness to small shifts in the analyzed window is still maintained. While the decay in performance for increasing shift in the analysis window is comparable for all methods, the stable predictors still retain better performance. Also in this 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.

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

| METHOD | $\mathrm{SD}_1$ | $\mathrm{SD}_2$ | $\mathrm{SD}_5$ | $\mathrm{SD}_{10}$ | $\mathrm{SD}_{20}$ |
|---|---|---|---|---|---|
| **LP2** | 0.138 | 0.159 | 0.233 | 0.478 | 1.342 |
| **LP1** | 0.005 | 0.010 | 0.017 | 0.021 | 0.039 |
| **STW** | 0.003 | 0.008 | 0.015 | 0.023 | 0.032 |
| **CT1** | 0.007 | 0.013 | 0.024 | 0.036 | 0.082 |
| **BU1** | 0.015 | 0.077 | 0.135 | 0.191 | 0.401 |
| **RW2** | 0.006 | 0.059 | 0.161 | 0.199 | 0.515 |

*3) Pitch Independence:* The ability of the linear predictors obtained with the 1-norm minimization criterion 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. In this experiment, the envelope was calculated using a $10^{th}$ order predictor obtained with **LP1**. The underlying pitch excitation is then modeled with an impulse train with different spacing. We then filtered this synthetic pitch excitation through the obtained LP filter and analyzed the synthetic speech applying the different LP methods in Table I. The analysis is divided into three subsets: high-pitched $T_p \in [16, 35]$ ($f_0 \in [228\mathrm{Hz}, 500\mathrm{Hz}]$), mid pitched $T_p \in [36, 71]$ ($f_0 \in [113\mathrm{Hz}, 222\mathrm{Hz}]$), and low pitched $T_p \in [72, 120]$ ($f_0 \in [67\mathrm{Hz}, 111\mathrm{Hz}]$). The shortcomings of **LP2** 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 stable methods are much less affected by the underlying pitch excitation, which results in an improved spectral modeling.

Copyright (c) 2014 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2014.2311324

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 |
|---|---|---|---|
| **LP2** | 0.81±0.14 | 1.12±0.29 | 1.32±0.59 |
| **LP1** | 0.05±0.00 | 0.11±0.00 | 0.14±0.01 |
| **STW** | 0.03±0.01 | 0.04±0.01 | 0.03±0.03 |
| **CT1** | 0.09±0.01 | 0.08±0.03 | 0.19±0.04 |
| **BU1** | 0.23±0.02 | 0.16±0.10 | 0.28±0.09 |
| **RW2** | 0.11±0.03 | 0.12±0.07 | 0.27±0.09 |

### B. Coding Performance

The second objective is to adopt the presented methods in the speech coding context. The experimental analysis was conducted on about one hour of clean speech (both voiced and unvoiced) coming from several different speakers with different characteristics (gender, age, pitch, regional accent) taken from the TIMIT database, re-sampled at 8 kHz. We propose three different experiments to evaluate the coding performance of the presented methods. In the first and second experiment we evaluate the sparsity of the residual induced by the choice of the prediction method and the consequent improvement in coding efficiency. In the third experiment, we explore the noise robustness of the LP estimators based on the 1-norm criterion, as noted and analyzed in [3] for coding and [50] for general modeling purposes.

*1) Experiment 1:* A $10^{th}$ order predictive analysis was first done on a segment of speech of $N = 40$. Then a multipulse encoding procedure [31] was performed to code $T$ pulses in the residual, with $T = 5$ and $T = 10$. Multipulse encoding was used to obtain a sparse residual, rather than a pseudo-random one, as obtained through algebraic codes, to better match the characteristics of the output of the 1-norm minimization. In Table V, we present the results in terms of segmental SNR and number of bits necessary to encode the prediction vector $\hat{\mathbf{a}}$ within the well-known 1 dB distortion [32] using the method presented in [27]. Since most of the residual samples are not identically zero[1], as an addition to the coding results, we provide three measures widely known throughout the literature to be more robust and effective than the 0-norm in measuring sparsity [34]. Considering $\mathbf{r}_i = \mathbf{x}_i - \mathbf{X}_i\mathbf{a}_i$, the linear prediction residual of a given unquantized predictor for the $i$-th considered segment of speech, we calculated the Hoyer criterion [35]:

$$\xi_i^H(\mathbf{r}_i) = \frac{N}{N-\sqrt{N}}\left(1 - \frac{\|\mathbf{r}_i\|_1}{\sqrt{N}\|\mathbf{r}_i\|_2}\right), \qquad (28)$$

the $pq$-mean [36]:

$$\xi_i^{pq}(\mathbf{r}_i) = \frac{1}{N^{\frac{1}{p}-\frac{1}{q}}}\left(\frac{\|\mathbf{r}_i\|_p}{\|\mathbf{r}_i\|_q}\right), \; 1 \le p < q, \qquad (29)$$

and the Gini index [37]:

$$\xi_i^G(\mathbf{r}_i) = 1 - 2\sum_{n=1}^{N}\frac{\hat{r}_{n,i}}{\|\mathbf{r}_i\|_1}\left(\frac{N-n+\frac{1}{2}}{N}\right), \qquad (30)$$

[1]Except for **LP1** where at least $K$ values will be zero, we cannot estimate *a priori* the number of zeros that will result in the optimization [33].

TABLE V
COMPARISON BETWEEN THE CONSIDERED PREDICTORS $\hat{\mathbf{a}}$ TRANSPARENTLY ENCODED WITH $B$ BITS ADOPTED IN A MULTIPULSE ENCODING SCHEME WITH $T$ PULSES. RESULTS ARE GIVEN IN TERMS OF SEGMENTAL SNR WITH 95% CONFIDENCE INTERVAL. THE AVERAGE SPARSITY OF THE RESIDUAL CALCULATED WITH THE HOYER MEASURE $\xi^H(\cdot)$, THE $pq$-MEAN $\xi^{pq}(\cdot)$, AND THE GINI INDEX $\xi^G(\cdot)$ IS ALSO SHOWN.

| METHOD | $T$ | $B(\hat{\mathbf{a}})$ | SSNR | $\xi^H(\cdot)$ | $\xi^{pq}(\cdot)$ | $\xi^G(\cdot)$ |
|---|---|---|---|---|---|---|
| **LP2** | 5 | 19 | 14.1±3.2 | 0.33 | 0.25 | 0.16 |
| | 10 | 19 | 19.1±2.9 | | | |
| **LP1** | 5 | 18 | 15.3±2.1 | 0.57 | 0.73 | 0.81 |
| | 10 | 18 | 20.1±1.7 | | | |
| **STW** | 5 | 17 | 14.9±1.6 | 0.51 | 0.63 | 0.75 |
| | 10 | 17 | 20.6±0.9 | | | |
| **CT1** | 5 | 15 | 13.9±1.9 | 0.49 | 0.61 | 0.72 |
| | 10 | 15 | 19.2±1.5 | | | |
| **BU1** | 5 | 19 | 14.2±0.9 | 0.47 | 0.59 | 0.71 |
| | 10 | 19 | 19.4±0.4 | | | |
| **RW2** | 5 | 21 | 15.2±1.2 | 0.55 | 0.67 | 0.74 |
| | 10 | 21 | 20.9±1.7 | | | |

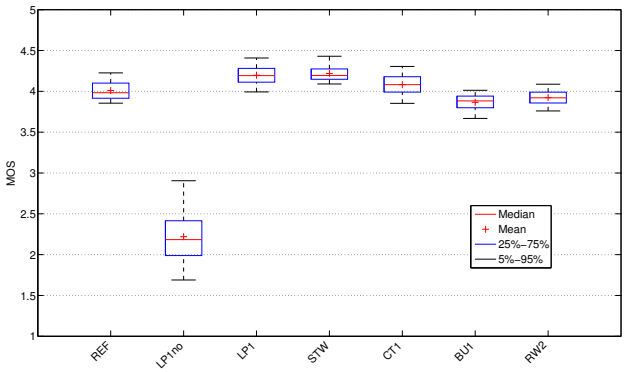



Fig. 2. Box plot of Mean Opinion Score obtained with POLQA calculated for the LP methods presented in Table I when implemented in the AMR-NB speech codec. **LP1no** represented the non-stabilized version of **LP1**. The reference is the traditional AMR with LP analysis performed with **LP2**.

where $\hat{\mathbf{r}}_i$ is the version of $\mathbf{r}_i$ indexed in non-decreasing order ($\hat{r}_{n,i} \le \hat{r}_{n+1,i}$). We then averaged over all the $T$ segments of speech for a given prediction method and determined the different sparsity measures $\xi(\cdot) = \frac{1}{T}\sum_{i=1}^{T}\xi_i(\mathbf{r}_i)$, where $\xi \in (0,1)$ ($\xi \to 1$ only for $\mathbf{r} \to \mathbf{0}$, the null vector). For $\xi_i^{pq}(\cdot)$, we used $p = 2$ and $q = 4$, which is a normalized version of the kurtosis, a well known measure for the peakedness of a distribution [38], and therefore another good measure for sparsity [39].

The best coding performance was achieved by **RW2**, consistently with the "guidance" in the reweighting algorithm based on the square root of the inverse of the residual amplitude. However, it also required a larger number of bits to transparently encode the predictor. As mentioned in the introduction, **BU1** does not preserve the sparsity of the residual and the coding characteristics of the 1-norm, giving similar performance to the **LP2**. The methods we have introduced seem to offer a good coding performance. The very smooth spectrum obtained with **CT1** allows considerably fewer bits than any other method to achieve transparent coding of the prediction coefficients, achieving a performance comparable to **LP2** and **BU1**. **STW** performs slightly worse than **RW2**,

Copyright (c) 2014 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2014.2311324

but with a significant saving in the bit budget of the predictor. The sparsity measures $\xi(\cdot)$ follow the experimental results pretty closely, confirming the validity of the SNR resulting from MPE encoding as a measure of sparsity. The Gini index offers the best discriminative properties for sparsity measures, as noted in [34].

*2) Experiment 2:* In the second experiment, we evaluate the performance of the different methods in an actual speech coder. In particular, we substitute the linear predictive analysis stage in the Adaptive Multi-Rate Narrow-Band (AMR-NB) encoder in its 12.2 kbps configuration [40], with the presented methods[2]. The main advantage of AMR-NB is that it is a multimodal coder, working on different rates from 12.2 kbps to 4.75 kbps, with the possibility of changing rate during the voice transmission by interacting with the channel coder [41]. The AMR codec is based on the Algebraic Code-Excited Linear Prediction (ACELP) paradigm [42], the most used approach for speech coding even for most recently developed codecs, e.g., OPUS and SILK [43] and thus the results presented in this section easily generalize to any speech codec based on the ACELP paradigm.

In our experiment, we replaced the calculation of the LP coefficients, obtained through efficiently solving (7), with the LP methods listed in Table I. Notice that **LP2** in this case is the unmodified AMR encoding. The AMR encoder uses 38 bits per frame (20 ms) to encode 2 sets of Line Spectral Frequencies (LSFs) calculated on the first and third subframe (5 ms). To demonstrate how the instability of **LP1** can be detrimental to the synthesized speech, thus justifying the need for the stabilized methods, we also included the non-stabilized 1-norm LP (**LP1no**). Given that the AMR transforms the LP coefficients in LSFs to quantize the LP coefficients, to obtain the results for **LP1no** we coded the stable predictor **LP1** and then reproduced the instability at the decoder by reflecting the poles with original magnitude greater than one. This was done because the LSF have an inherent stability control, the *interlacing property*; thus LSF cannot be calculated if the predictor is unstable [44].

We calculated the Mean Opinion Score (MOS) [45] obtained with POLQA [46], the latest ITU-T standard for objective speech quality assessment (the successor of PESQ [47]). We considered roughly 500 samples of clean speech of length 5 seconds (around 40 minutes).

The results presented in Figure 2 show a significant overlap of the 5-95 percentile regions and the interquartile ranges of **STW** and **LP1**, suggesting no statistical difference between the two distributions. The mean and median values of the scores obtained with **CT1** and **STW** are both significantly higher than the AMR with **LP2**. Notice that the instability of **LP1no** generates saturations in the decoded speech, thus greatly degrading the codec performance. The instability, while on average around 3% for the analyzed set (since it includes both voiced and unvoiced speech), also corrupts the state of several future decoded frames, hence the net difference in performance with its stabilized version. The low performance of **RW2** compared to the other methods seems to bring different conclusions from what is presented in Table V. However, it should be noted that **RW2** requires more bits for quantization, thus suggesting that the 19 bits allocated in the AMR payload are probably not sufficient for transparent quantization. This presents a practical problem for the comparison of the different methods, as we would need to optimize and apply different bit-rates for the different methods. For example, a better trade-off between the bits necessary to encode the predictor and the residual should be addressed. However, the experiment done and its results outline, without loss of generality, higher modeling properties of the proposed predictors over the state-of-the-art methodology.

[2]While AMR-NB is arguably a "old" codec, it is still widely used in speech communications, especially given the delay in adoption of wideband codec caused by the increased usage of bandwidth hungry smartphones and the explosion of cellular phone usage in emerging markets.

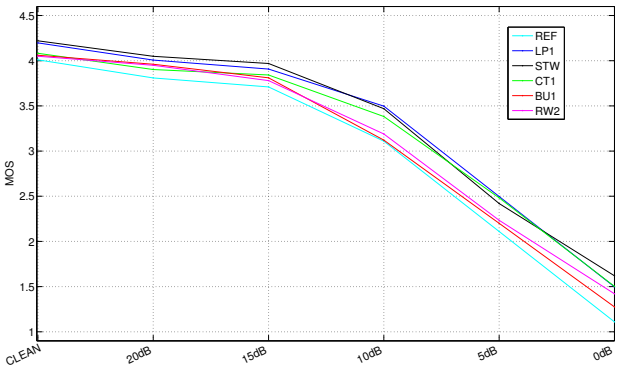



Fig. 3. Mean Opinion Score obtained with POLQA calculated for the LP methods presented in Table I when implemented in the AMR-NB speech codec for varying SNRs (white noise).

*3) Experiment 3:* In the third experiment, we evaluate the setup of the previous experiment for noise robustness. It is well known throughout the literature that **LP1** is more robust than **LP2** for analysis and coding of speech, e.g., [12], [48]. In this section, we compared the stable solutions for different SNRs. We considered white noise only, as it is more appropriate for this type of experiment and does not corrupt the spectral features of the speech signal. Noise and speech were mixed at SNRs ranging from -5 to 25 dB following the ITU-T Recommendation P. 835 [49] where the reference signal was always scaled to an ideal average active level of approximately -26 dBov to avoid clipping in the mixed signals. The MOS scores calculated using POLQA are the median values calculated over all the utterances. Notice that the clean condition scores are the median values of Figure 2. The results clearly show the improvement in quality and a slightly slower decay for the **STW**, **LP1**, and **CT1**. It is interesting to see that **STW** achieves an improvement of 0.5 in MOS over traditional AMR and 0.1 over **LP1**. This suggests that the windowing performed in **STW** helps in noisy conditions, as noted in [50], and can actually improve the performance of **LP1**.

## V. Solutions and Complexity

The higher complexity burden of linear prediction based on the 1-norm minimization is well known throughout the literature [3]. In particular, the traditional linear prediction solution **LP2** in (7) can be solved using $\mathcal{O}((N+K)K)$

Copyright (c) 2014 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2014.2311324

floating point operations using the Levinson-Durbin algorithm [2]. This is one order lower than solving one iteration for the 1-norm problem using interior-point methods. However, efficient solutions based on hand-tailoring solvers to the given problem have proven to be successful in bringing the complexity down to acceptable levels [15]. In this section, we introduce efficient solvers for the methods considered in the paper and analyze their complexity.

### A. 1-norm error minimization

It is well-known that solving the least 1-norm problem (8) using interior points methods corresponds to solving a small number of weighted least-squares problems with the same coefficient matrix and weights that change each iteration [16, §11.8.2]. The number of iterations is, in the worst case, also a function of $K$, but this is often neglected because in practice the dependency is very small if any [51, §14.1]. Via direct methods this then has the complexity $\mathcal{O}(K^2(N+K))$. This can be exploited in hand-tailored algorithms to produce fast solvers, e.g., [15]. If a stable solution is necessary, a stability check performed by determining the roots of the LP polynomial significantly adds on the complexity. Furthermore, no simple modification can be made to the efficient process of estimating the LP coefficients and the following calculation of the LSF [52] [53] in a speech coder, as they are all based on the assumption of the roots of the polynomial being contained in the unit circle. This also justifies the need for intrinsically stable solutions.

### B. Reducing the numerical range of the shift operator

Complexity-wise, solving (19) has the same complexity of **LP1**. However, the calculation of the STE weighting (21) and the shift matrix (17) to generate (19), increases the complexity to $\mathcal{O}(K^2(N+K)+NK)$. Nonetheless, this is still a significantly lower increase in computational overhead compared to stabilizing through pole reflection, as mentioned in Section V-A.

### C. Constrained 1-norm minimization

Solving the constrained 1-norm minimization can be achieved by solving a number of linear systems with the coefficient matrix:

$$X^T D_1 X + D_2 + \nu d_3 d_3^T, \tag{31}$$

where $D_1 \in \mathbb{R}^{N+K\times N+K}$, $D_2 \in \mathbb{R}^{K\times K}$, $\nu \in \mathbb{R}^1$ and $d_3 \in \mathbb{R}^{K\times 1}$ changes in each iteration ($D_1$ and $D_2$ are diagonal matrices). This is the same as what is required for solving the unconstrained minimization problem, see [15, Eqn. (2)], plus a rank-1 term. The complexity for forming and solving these linear systems of equations are then $\mathcal{O}(K^3+K^2(N+K))$. This is the same complexity as in [15] and similar practical performance is expected.

### D. Burg method based on 1-norm minimization

The optimization step 1 in Algorithm 2 can be computed in closed form. For notation, let us consider $\boldsymbol{\alpha} = [\mathbf{f};\mathbf{b}]$ and $\boldsymbol{\beta} = [\mathbf{b};\mathbf{f}]$. The convex problem can then be written as $\min_k \|k\boldsymbol{\alpha}+\boldsymbol{\beta}\|_1$ with the optimality condition:

$$\sum_{i=1}^{M} \alpha_i \partial \|k^\star \alpha_i + \beta_i\|_1 \ni 0 \tag{32}$$

where $M = 2(N+K)$ and

$$\partial\|x\|_1 = \begin{cases} [-1;1] & \text{if} \quad x = 0 \\ -1 & \text{if} \quad x < 0 \\ 1 & \text{if} \quad x > 0 \end{cases} \tag{33}$$

is the subgradient of $\|x\|_1$. The optimum can be computed by considering that the left-hand side of (33) only changes at points where there is a shift in the sign of $\partial\|\cdot\|_1$. Then a solution satisfies $k^\star \in \mathbb{K} = \{-\beta_i/\alpha_i \,|\, \alpha_i \neq 0, i = 1,\ldots,M\}$. It is then possible to test the candidates $k \in \mathbb{K}$ and evaluate either the optimality condition or compute and compare the objectives with a total complexity of $\mathcal{O}(M^2)$. Another algorithm first computes $v_i = -\beta_i/\alpha_i$ (if $\alpha_i = 0$ this term can be removed from the optimization problem). If any $v_i = v_j$, with $i \neq j$, remove element $i$ and scale $\alpha_j \leftarrow 2\alpha_j$ and $\beta_j \leftarrow 2\beta_j$ (such that all $v_i$ are unique). Then sort the values such that

$$v_{\mathcal{I}(j)} < v_{\mathcal{I}(j+1)}. \tag{34}$$

Let

$$F(j) = \sum_{i=1, i\neq\mathcal{I}(j)}^{M} \alpha_i \text{sgn}(v_{\mathcal{I}(j)}\alpha_i + \beta_i)\,. \tag{35}$$

Notice that at $k = v_{\mathcal{I}(j)} \in \mathbb{K}$

$$\sum_{i=1}^{M} \alpha_i \partial\|v_{\mathcal{I}(j)}\alpha_i + \beta_i\| = F(j) + \alpha_{\mathcal{I}(j)}\partial\|0\|. \tag{36}$$

The optimality criteria at a candidate point in $\mathbb{K}$ is then

$$|F(j^\star)| \leq |\alpha_{\mathcal{I}(j^\star)}|, \quad k^\star = v_{\mathcal{I}}(j^\star) \in \mathbb{K}\,. \tag{37}$$

Instead of evaluating $F(j)$ for all $j$ via (37), it is possible to use the recursive formula

$$\begin{aligned} F(j+1) = F(j) &- \alpha_{\mathcal{I}(j+1)}\text{sgn}(v_{\mathcal{I}(j)}\alpha_{\mathcal{I}(j+1)} + \beta_{\mathcal{I}(j+1)}) \\ &+ \alpha_{\mathcal{I}(j)}\text{sgn}(v_{\mathcal{I}(j+1)}\alpha_{\mathcal{I}(j)} + \beta_{\mathcal{I}(j)}), \end{aligned} \tag{38}$$

for $j = 1,\ldots,M-1$. It is then possible to evaluate and check the optimality condition $|F(j)| \leq |\alpha_{\mathcal{I}(j)}|$ for all $j = 1,\ldots,M$ with complexity $\mathcal{O}(M\log M)$. Since this is a linear program, general algorithms for solving linear programs are also applicable, but the method listed above is preferable due to its simplicity. The method is summarized in Algorithm 3. The total complexity of the 1-norm Burg method is then $\mathcal{O}(K(N+K)\log(N+K))$.

### E. Iteratively Reweighted 2-norm minimization

The iteratively reweighted 2-norm minimization has the per iteration floating-point complexity $\mathcal{O}(K^2(N+K))$ using direct methods. Empirically we have observed that 4-5 reweighting schemes are sufficient to reach convergence, thus the complexity is in the same order as **LP1**.

Copyright (c) 2014 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2014.2311324

**Algorithm 3** Solving The Subproblem In 1-norm Burgs Method (Algorithm 2)

Inputs: speech segment $\mathbf{f}, \mathbf{b} \in \mathbb{R}^{M/2\times 1}$
Output: $k^\star$
$\boldsymbol{\alpha} = [\mathbf{f}; \mathbf{b}]$, $\boldsymbol{\beta} = [\mathbf{b}; \mathbf{f}]$
$v_i = -\beta_i/\alpha_i$ (assume unique $v_i$ and $\alpha_i \neq 0$)
Calculate an index table $\mathcal{I}$ for sorting $v_i$ ascending
Calculate $F(1)$ via Equation (35)
**if** $|F(1)| \leq |\alpha_{\mathcal{I}(1)}|$ **then**
  Return $k^\star = v_{\mathcal{I}(1)}$
**end if**
**for** $j = 2, \ldots, M$ **do**
  Calculate $F(j)$ via Equation (38)
  **if** $|F(j)| \leq |\alpha_{\mathcal{I}(j)}|$ **then**
    Return $k^\star = v_{\mathcal{I}(j)}$
  **end if**
**end for**

## VI. Conclusions

In this paper, we have presented two new methods for intrinsically finding stable predictors based on 1-norm error minimization. The methods introduced, one based on the reduction of the numerical range of the shift operator and one based on constrained 1-norm minimization, have both shown to offer a valid alternative to the original 1-norm linear prediction, preserving the properties of the 1-norm error minimization criterion. In particular, the experimental analysis has shown that both methods offer attractive modeling and coding performance without any significant increase in complexity. The two methods have also been shown to offer slightly better modeling performance compared to the Burg method based on the 1-norm minimization and the 2-norm reweighted minimization method. For all the considered methods, a thorough experimental analysis has shown that the properties that make 1-norm based linear prediction appealing for both analysis and coding of speech are preserved without too much degradation. These properties, shift invariance and pitch invariance, derive from the more efficient decoupling between the pitch harmonics and the spectral envelope and an overall better modeling of the speech production process. Furthermore, the application of the proposed predictors by modifying the linear prediction step to currently deployed state-of-the-art codecs, showed improved quality for clean conditions and a slower decaying of performance for decreasing SNR.

## 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, 1975.
[3] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen "Sparse linear prediction and its applications to speech processing," *IEEE Trans. on Speech, Audio and Language Processing*, vol. 20, no. 5, pp. 1644–1657, 2012.
[4] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen, "Speech coding based on sparse linear prediction," *Proc. European Signal Processing Conf.*, pp. 2524–2528, 2009.
[5] 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 Processing Conf.*, pp. 234–238, 2010.
[6] 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 Signal Processing Letters*, vol. 17, no. 1, pp. 103–106, 2010.
[7] W. F. G. Mecklenbrauker, "Remarks on the minimum phase property of optimal prediction error filters and some related questions," *IEEE Signal Processing Letters*, vol. 5, no. 4, pp. 87–88, 1998.
[8] E. Denoël and J.-P. Solvay, "Linear prediction of speech with a least absolute error criterion," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 33, no. 6, pp. 1397–1403, 1985.
[9] L. Knockaert, "Stability of linear predictors and numerical range of shift operators in normed spaces," *IEEE Trans. on Information Theory*, vol. 38, no. 5, pp. 1483–1486, 1992.
[10] Q. F. Stout, "The numerical range of a weighted shift," *Proceedings American Mathematical Society*, no. 88, pp. 495–502, 1983.
[11] C. Ma, Y. Kamp, and L. F. Willems, "Robust signal selection for linear prediction analysis of voiced speech," *Speech Communication*, vol. 12, no. 1, pp. 69–81, 1993.
[12] C. Magi, J. Pohjalainen, T. Bäckström, and P. Alku, "Stabilized weighted linear prediction," *Speech Communication*, vol. 51, no. 5, pp. 401–411, 2009.
[13] A. L. Cauchy, *Exercise de mathematique*, Oeuvres 2, vol. 19, 1829.
[14] M. Marden, *Geometry of polynomials*, Mathematical Surveys and Monographs, American Mathematical Society, 1966.
[15] T. L. Jensen, D. Giacobello, M. G. Christensen, S. H. Jensen, and M. Moonen, "Real-time implementation of sparse linear prediction for speech processing," *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pp. 8184–8188, 2013.
[16] S. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge University Press, 2004.
[17] P. Stoica and R. Moses, *Spectral analysis of signals*, Pearson Prentice Hall, 2005.
[18] S. Kotz, N. Balakrishnan, N. L. Johnson, *Continuous multivariate distributions, Volume 1, Models and applications*, 2nd edition, Wiley, 2000.
[19] A. W. Marshall and I. Olkin, "Maximum likelihood characterizations of distributions," *Statistica Sinica*, vol. 3, pp. 157–171, 1993.
[20] C. Fernandez, J. Osiewalski, and M. F. J. Steel, "Modeling and inference with $v$-spherical distributions," *J. American Statistical Association*, vol. 90, pp. 1331–1340, 1995.
[21] T. Eltoft, T. Kim, and T. Lee, "On the multivariate Laplace distribution," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 300–303, 2006.
[22] S. Gazor and W. Zhang, "Speech probability distribution," *IEEE Signal Processing Letters*, vol. 10, no. 7, pp. 204–207, 2003.
[23] S. J. Wright, *Primal-Dual Interior-Point methods*, SIAM, 1997.
[24] H. P. Hirst and W. T. Macey, "Bounding the roots of polynomials," *The College Mathematics Journal*, vol. 18, no. 4, pp. 292–295, 1997.
[25] Y. Li, "A globally convergent method for $L_p$ problems," *SIAM J. Optimization*, vol. 3, no. 3, pp. 609–629, 1993.
[26] L. A. Ekman, W. B. Kleijn, and M. N. Murthi, "Regularized Linear Prediction of Speech," *IEEE Trans. on Speech, Audio and Language Processing*, vol. 16, no. 1, pp. 65–73, 2008.
[27] A. D. Subramaniam and B. D. Rao, "PDF optimized parametric vector quantization of speech line spectral frequencies," *IEEE*

Copyright (c) 2014 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2014.2311324

*Trans. on Speech and Audio Processing*, vol. 11, no. 2, pp. 87–89, 2003.

[28] L. Marple, "A new autoregressive spectrum analysis algorithm," *IEEE Trans. on Acoustic, Speech and Signal Processing*, vol. 28, no. 4, pp. 441–454, 1980.

[29] D. Messerschmitt, *Autocorrelation matrix eigenvalues and the power spectrum*, EECS Department, University of California, Berkeley, Technical Report No. UCB/EECS-2006-90, 2006.

[30] C. H. Lee, "On robust linear prediction of speech," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 36, no. 5, pp. 642–650, 1988.

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

[32] W. B. Kleijn and A. Ozerov, "Rate distribution between model and signal," *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 243–246, 2007.

[33] J. Cadzow, "Minimum $\ell_1$, $\ell_2$, and $\ell_\infty$ norm approximate solutions to an overdetermined system of linear equations," *Digital Signal Processing*, vol. 12, no. 4, pp. 524–560, 2002.

[34] N. Hurley and S. Rickard, "Comparing measures of sparsity," *IEEE Trans. on Information Theory*, vol. 55, no. 10, pp. 4723–4741, 2009.

[35] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *J. of Machine Learning Research*, vol. 5, pp. 1457–1469, 2004.

[36] A. Bronstein, M. Bronstein, M. Zibulevsky, and Y. Y. Zeevi, "Sparse ICA for blind separation of transmitted and reflected images," *Int. J. Imag. Syst. Technol.,* vol. 15, no. 1, pp. 84–91, 2005.

[37] C. Gini, "Measurement of inequality of incomes," *The Economic Journal*, vol. 31, pp. 124–126, 1921.

[38] 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. on Acoustics, Speech, and Signal Processing*, pp. 4650–4653, 2010.

[39] B. A. Olshausen and D. J. Field, "Sparse coding of sensory inputs," *Current Opinion in Neurobiology*. vol. 14, no. 4, pp. 481–487, 2004.

[40] *Mandatory Speech Codec speech processing functions; Adaptive Multi-Rate (AMR) speech codec; Transcoding functions*, 3GPP, TS 26.090, Rel. 11, 2012.

[41] E. Ekudden, R. Hagen, I. Johansson, and J. Svedberg, "The adaptive multi-rate speech coder," *Proc. IEEE Workshop on Speech Coding*, pp. 117–119, 1999.

[42] J-P. Adoul, P. Mabilleau, M. Delprat, and S. Morissette, "Fast CELP coding based on algebraic codes," *PProc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 12, pp. 1957–1960, 1987.

[43] K. Vos, K. V. Sorensen, S. S. Jensen, J.-M. Valin, "Voice Coding with Opus," *Proc. 135th Audio Engineering Society Conv.*, 2013.

[44] T. Bäckström and C. Magi, "Properties of line spectrum pair polynomials - A review," *Signal Processing*, vol. 86, no. 11, pp. 3286–3298, 2006.

[45] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. on Speech, Audio and Language Processing*, vol. 16, no. 1, pp. 229–238, 2008.

[46] *Perceptual objective listening quality assessment (POLQA)*, ITU-T, Rec. P. 863, 2011.

[47] *Perceptual evaluation of speech quality (PESQ)*, ITU-T, Rec. P. 862, 2001.

[48] R. Saeidi, J. Pohjalainen, T. Kinnunen, and P. Alku, "Temporally weighted linear prediction features for tackling additive noise in speaker verification," *IEEE Signal Processing Letters*, vol. 17, no. 6, pp. 599–602, 2010.

[49] *Subjective Test Methodology for Evaluating Speech Communication Systems that Include Noise Suppression Algorithms*, ITU-T, Rec. P. 835, 2003.

[50] C. Magi, T. Bäckström, and P. Alku, "Objective and subjective evaluation of seven selected all-pole modeling methods in processing of noise corrupted speech," *Proc. IEEE Nordic Signal Processing Symposium*, 2006.

[51] J. Nocedal and S. J. Wright, *Numerical Optimization*, Second Edition, Springer, 2006.

[52] P. Kabal and R. P. Ramachandran, "The computation of line spectral frequencies using Chebyshev polynomials," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 34, no. 6, pp. 1419–1426, 1986.

[53] F. Soong and B. H. Juang, "Line spectrum pair (LSP) and speech data compression," *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 9, pp. 37–40, 1984.



**Daniele Giacobello** (S'07, M'10) received his B.Sc. (Laurea) and M.Sc. (Laurea Magistrale) degrees in Telecommunications Engineering from Politecnico di Milano, Italy, in 2003 and 2006, respectively, and his Ph.D. degree in Electrical and Electronic Engineering from Aalborg University, Denmark, in 2010. Before joining Beats Electronics, LLC as a Senior Research Engineer, he was with the Office of the CTO at Broadcom Corporation, Irvine, CA; the Department of Electronic Systems at Aalborg University; Asahi-Kasei Corporation, Atsugi, Japan; and Siemens AG, Milan, Italy. He was also a Visiting Scholar at Delft University of Technology, University of Miami, and Katholieke Universiteit, Leuven, Belgium.

Dr. Giacobello's research interests include digital signal processing theory and methods with applications to speech and audio signals, in particular sparse representation, statistical modeling, coding, and recognition. He is a recipient of the European Union Marie Curie Doctoral Fellowship and was awarded the Best Information Engineering Thesis Award by the Milan Engineers Foundation for his M.Sc. thesis work. He is EURASIP local liaison for Southern California and will co-chair the special session "Digital Audio Processing for Loudspeakers and Headphones" at the 22nd European Signal Processing Conference (EUSIPCO 2014).



**Mads Græsbøll Christensen** (S'00, M'05, SM'11) received the M.Sc. and Ph.D. degrees in 2002 and 2005, respectively, from Aalborg University (AAU) in Denmark, where he is also currently employed at the Dept. of Architecture, Design & Media Technology as Professor in Audio Processing and is head of the Audio Analysis Lab, which conducts research in audio signal processing.

He was formerly with the Dept. of Electronic Systems, Aalborg University and has held visiting positions at Philips Research Labs, ENST, UCSB, and Columbia University. He has published more than 100 papers in peer-reviewed conference proceedings and journals as well as 2 research monographs. His research interests include digital signal processing theory and methods with application to speech and audio, in particular parametric analysis, modeling, enhancement, separation, and coding.

Prof. Christensen has received several awards, including an ICASSP Student Paper Award, the Spar Nord Foundation's Research Prize for his Ph.D. thesis, a Danish Independent Research Council Young Researcher's Award, and the Statoil Prize, as well as grants from the Danish Independent Research Council and the Villum Foundation's Young Investigator Programme. He is an Associate Editor for IEEE Transactions on Audio, Speech, and Language Processing and has previously served as an Associate Editor for IEEE Signal Processing Letters.

Copyright (c) 2014 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2014.2311324



**Tobias Lindstrøm Jensen** received the M.Sc. degree and Ph.D. degree in Electrical Engineering from Aalborg University in 2007 and 2011, respectively. Tobias L. Jensen is currently affiliated with the Department of Electronic Systems at Aalborg University. In 2007, he worked at Wipro-NewLogic Technologies in Sophia-Antipolis, France. In 2009 he was a visiting research scholar at University of California, Los Angeles (UCLA). His research interests includes optimization, signal and image processing, signal processing for communication, inverse problems and estimation.



**Manohar N. Murthi** received his B.S. degree in electrical engineering and computer science from the University of California, Berkeley, in 1990, and his M.S. and Ph.D. degrees in electrical engineering (communication theory and systems) from the University of California, San Diego, CA, in 1992 and 1999, respectively.

He has previously worked at Qualcomm in San Diego, CA, KTH (Royal Institute of Technology), Stockholm, Sweden, and Global IP Sound in San Francisco, CA. In September 2002 he joined the Department of Electrical and Computer Engineering, University of Miami, Coral Gables, FL, where he is an Associate Professor. His research interests are in the general areas of signal and data modeling, compression, fusion and learning, and networking. He is a recipient of a National Science Foundation CAREER Award.



**Søren Holdt Jensen** (S'87, M'88, SM'00) received the M.Sc. degree in electrical engineering from Aalborg University, Aalborg, Denmark, in 1988, and the Ph.D. degree in signal processing from the Technical University of Denmark, Lyngby, Denmark, in 1995. Before joining the Department of Electronic Systems of Aalborg University, he was with the Telecommunications Laboratory of Telecom Denmark, Ltd, Copenhagen, Denmark; the Electronics Institute of the Technical University of Denmark; the Scientific Computing Group of Danish Computing Center for Research and Education (UNI•C), Lyngby; the Electrical Engineering Department of Katholieke Universiteit Leuven, Leuven, Belgium; and the Center for PersonKommunikation (CPK) of Aalborg University. He is Full Professor and heading a research team working in the area of numerical algorithms, optimization, and signal processing for speech and audio processing, image and video processing, multimedia technologies, and digital communications. Prof. Jensen was an Associate Editor for the IEEE Transactions on Signal Processing, Elsevier Signal Processing and EURASIP Journal on Advances in Signal Processing, and is currently Associate Editor for the IEEE Transactions on Audio, Speech and Language Processing. He is a recipient of an European Community Marie Curie Fellowship, former Chairman of the IEEE Denmark Section and the IEEE Denmark Sections Signal Processing Chapter. He is member of the Danish Academy of Technical Sciences and was in January 2011 appointed as member of the Danish Council for Independent Research—Technology and Production Sciences by the Danish Minister for Science, Technology and Innovation.



**Marc Moonen** (M'94, SM'06, F'07) is a Full Professor at the Electrical Engineering Department of KU Leuven, where he is heading a research team working in the area of numerical algorithms and signal processing for digital communications, wireless communications, DSL and audio signal processing.

He received the 1994 KU Leuven Research Council Award, the 1997 Alcatel Bell (Belgium) Award (with Piet Vandaele), the 2004 Alcatel Bell (Belgium) Award (with Raphael Cendrillon), and was a 1997 Laureate of the Belgium Royal Academy of Science. He received a journal best paper award from the IEEE Transactions on Signal Processing (with Geert Leus) and from Elsevier Signal Processing (with Simon Doclo).

He was chairman of the IEEE Benelux Signal Processing Chapter (1998-2002), a member of the IEEE Signal Processing Society Technical Committee on Signal Processing for Communications, and President of EURASIP (European Association for Signal Processing, 2007-2008 and 2011-2012).

He has served as Editor-in-Chief for the EURASIP Journal on Applied Signal Processing (2003-2005), and has been a member of the editorial board of IEEE Transactions on Circuits and Systems II, IEEE Signal Processing Magazine, Integration-the VLSI Journal, EURASIP Journal on Wireless Communications and Networking, and Signal Processing. He is currently a member of the editorial board of EURASIP Journal on Applied Signal Processing and Area Editor for Feature Articles in IEEE Signal Processing Magazine.

Copyright (c) 2014 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.