SPARSE REGRESSIONS FOR JOINT SEGMENTATION AND LINEAR PREDICTION

Daniele Angelosante

ABB Corporate Research Center, Segelhofstrasse 1K, Baden-Daettwil, 5405, Switzerland.

ABSTRACT

Regularizing the least-squares criterion with the total number of coefficient changes, it is possible to estimate time-varying (TV) autoregressive (AR) models with piecewise-constant coefficients. Such models emerge in various applications including speech segmentation using linear predictors. To cope with the large-size optimization task, the problem is cast as a sparse regression one, and is solved by resorting to an efficient block-coordinate descent algorithm. This enables joint segmentation and linear predictor coefficients identification with linear computational complexity per iteration. Modern trends in linear prediction for speech processing also envision sparsity in the model residuals. Indeed, sparse residuals allow for an improved representation of voiced speech. So far, sparse linear coding was proposed in a stationary scenario, i.e., after speech segmentation. This paper extends joint segmentation and linear prediction coefficients identification to sparse linear coding. Fortunately, coordinate descent approaches are still applicable to carry out the optimization tasks. Numerical tests have shown the benefits of the proposed algorithm.

Index Terms— Sparse regression, Linear prediction, Convex optimization, Coordinate descent.

1. INTRODUCTION

Autoregressive (AR) models have been widely used for spectral estimation since they can approximate any process having continuous spectral density with an arbitrary precision. Furthermore, they approximate the spectrum of a given random process with few parameters [14, Chap. 3].

More specifically, in speech analysis, parametric spectral estimation via AR modeling falls under the rubric of *linear prediction* (LP). LP has been successfully applied in many state-of-the-art speech processing systems for coding, analysis, synthesis, and recognition. The model used in many of these applications is the *source-filter* where the speech signal can be generated as the output of an all-pole filter excited by a Gaussian and white (i.e., noise-like) process. As a consequence of the Gaussianity, least-squares (LS) criterion has been traditionally used for identification of LP coefficients.

While LP modeling of stationary random processes is well appreciated, a number of signals encountered in real life are non-stationary. To apply classical LS identification of LP coefficients, a first preprocessing is typically required to find the abrupt changes in the signal spectral characteristics. In speech analysis, this problem is often referred to as *speech segmentation*.

Recent advances demonstrated that regularizing the least-squares criterion with the total number of coefficient changes, it is possible to estimate TV-AR models with piecewise-constant coefficients. Indeed, in [1], it was shown that the segmentation problem can be recast as a *sparse regression* one, and, to facilitate its computation, the regularization function in [8] is relaxed with its convex approximation. The resultant joint segmentation and LP identification method is a modification of the group least-absolute shrinkage and selection operator (Lasso) [16]. With the emphasis placed on large data sets, a block-coordinate descent iteration was proposed to carry out the optimization task, which is provably convergent to the group Lasso solution with complexity per iteration that scales linearly with the data length.

Modern techniques for LP modeling envision also sparsity in the model residuals [5]. Indeed, sparse residuals enables a more effective decoupling of the vocal tract transfer function. Nevertheless, these techniques have been proposed for stationary cases, i.e., after speech segmentation. Therefore, the goal of this paper is perform *joint segmentation and sparse linear prediction*. The proposed framework for speech analysis is rooted in convex optimization and compressive sampling theory [3]. Therefore, the pursue of efficient solvers for large-size optimization problems is a fundamental task. Accounting for sparse residuals, it turns out that block-coordinate descent algorithms can still be adopted.

The remainder of the paper is structured as follows. Section 2 deals with piecewise-constant AR model estimation preliminaries and problem statement. Section 3 is devoted to joint segmentation and identification of sparse LP residual. After some manipulations, it is showed that the problem can be solved via block-coordinate descent. Numerical tests are presented in Sec. 4, and concluding remarks are summarized in Sec. 5. *Notation:* Column vectors (matrices) are denoted using lower-case (upper-case) boldface letters; calligraphic letters are reserved for sets; $(\cdot)^T$ stands for transposition, $\mathcal{N}(\mu, \sigma^2)$ denotes the Gaussian probability density function with mean μ and variance σ^2; $\otimes$ denotes the Kronecker product; $\mathbf{0}_L$ is the L-dimensional column vector with all zeros, $\mathbf{0}_{M,N}$ is the $M \times N$ matrix with all zeroes, and $\mathbf{I}_L$ is the L-dimensional identity matrix. The ℓ_p norm of $\mathbf{x} := [x_1, \ldots, x_L]^T \in \mathbb{R}^L$ is defined as $\|\mathbf{x}\|_p := (\sum_{\ell=1}^{L} |x_\ell|^p)^{\frac{1}{p}}$.

2. PRELIMINARIES AND PROBLEM STATEMENT

Let $\{y_n\}_{n=-L}^{N}$ denote the realization of an L-th order TV-AR process obeying the discrete-time input-output relationship

$$y_n = \sum_{\ell=1}^{L} a_{\ell,n} y_{n-\ell} + v_n, \quad n = 0, 1, \ldots, N \tag{1}$$

where v_n denotes the zero-mean white input noise at time n with variance $\sigma_y^2 := \mathbb{E}[v_n^2] < +\infty$, and $a_{\ell,n}$ is the ℓ-th TV-AR coefficient at time n. With $\mathbf{h}_n := [y_{n-1}, y_{n-2}, \ldots, y_{n-L}]^T \in \mathbb{R}^L$ and $\mathbf{a}_n := [a_{1,n}, a_{2,n}, \ldots, a_{L,n}]^T \in \mathbb{R}^L$, (1) can be rewritten as

$$y_n = \mathbf{h}_n^T \mathbf{a}_n + v_n, \quad n = 0, 1, \ldots, N \tag{2}$$

.

Email: daniele.angelosante@ch.abb.com

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

It is assumed that $K+1$ *abrupt changes* in the spectrum of $\{y_n\}$ occur, i.e., $\mathbf{a}_n$'s are *piecewise-constant*; that is, $\mathbf{a}_n = \boldsymbol{a}_k, \quad n_k \leq n \leq n_{k+1}-1$ for $k = 0, 1, \ldots, K$, where K denotes the number of abrupt changes in the TV-AR spectrum, and n_k the time instant of the k-th abrupt change. The interval $[n_k, n_{k+1}-1]$ is referred to as the k-th *segment*. Without loss of generality, $n_0 = 0$ and $n_{K+1} - 1 = N$.

The goal is to estimate the instants $\{n_k\}_{k=1}^K$ where the given time series $\{y_n\}$ is split into $K+1$ (stationary) segments, along with the constant AR coefficients per segment, i.e., $\{\boldsymbol{a}_k\}_{k=0}^K$. The number of abrupt changes, namely K, is not necessarily known, and, throughout the paper, different characteristics on the AR coefficients are imposed.

2.1. Optimum segmentation of TV-AR processes

Denoting with μ a positive tuning constant, the following regularization is typically adopted to estimate jointly the change points and the AR coefficients [2, 9–11], i.e.,

$$\{\widehat{\mathbf{a}}_n\}_{n=0}^N := \arg \min_{\{\mathbf{a}_n\}_{n=0}^N} \left[\frac{1}{2}\sum_{n=0}^N (y_n - \mathbf{h}_n^T \mathbf{a}_n)^2 + \mu \sum_{n=1}^N \delta_{\mathbf{0}_L}(\mathbf{a}_n - \mathbf{a}_{n-1})\right] \tag{3}$$

where $\delta_{\mathbf{0}_L}(\cdot) : \mathbb{R}^L \to \{0, 1\}$ is defined as

$$\delta_{\mathbf{0}_L}(\mathbf{a}) := \begin{cases} 0, & \text{if } \mathbf{a} = \mathbf{0}_L \\ 1, & \text{otherwise.} \end{cases} \tag{4}$$

The non-convex regularization term $\sum_{n=1}^N \delta_{\mathbf{0}_L}(\mathbf{a}_n - \mathbf{a}_{n-1})$ captures the total number of changes, and its penalization encourages piecewise-constant $\{\widehat{\mathbf{a}}_n\}_{n=0}^N$. Clearly, the larger the μ, the smaller the total number of changes. The estimator in (3) is optimal in the maximum a posteriori (MAP) sense when the change occurrences are modeled as Bernoulli random variables, and $v_n \sim \mathcal{N}(0, \sigma_y^2)$ [8].

From a practical point of view, the minimization in (3) is challenging since an exhaustive search over all possible sets of change instants has to be performed. Nevertheless, the problem can be solved with dynamic programming (DP). Despite the fact that DP approaches solve (3) in polynomial time, the computational complexity is cubic in N [12, p. 469], which limits its applicability to signal segmentation in practice. In fact, in typical speech applications, N can be very large (up to several thousands) and polynomial complexity cannot be afforded.

In the following, a convex relaxation of the problem in (3) is advocated hinging upon recent advances in sparse linear regression and compressive sampling. To this end, (3) is first reformulated as a sparse regression problem with non-convex regularization that is successively relaxed through its convex approximation. The consequent optimization rule yields sparse vector estimators which can obtained by a block-coordinate descent iteration that incurs only linear computational burden and memory storage.

Let $\mathbf{y} := [y_0, y_1, \ldots, y_N]^T \in \mathbb{R}^{N+1}$ denote the observation vector, $\mathbf{a} := [\mathbf{a}_0^T, \mathbf{a}_1^T, \ldots, \mathbf{a}_N^T]^T \in \mathbb{R}^{(N+1)L}$, $\mathbf{m}_n := [\underbrace{\mathbf{0}_L^T, \ldots, \mathbf{0}_L^T}_{n}, \mathbf{h}_n^T, \underbrace{\mathbf{0}_L^T, \ldots, \mathbf{0}_L^T}_{N-n}]^T \in \mathbb{R}^{(N+1)L}$ for $n = 0, 1, \ldots, N$, and $\mathbf{M} := [\mathbf{m}_0, \mathbf{m}_1, \ldots, \mathbf{m}_N]^T \in \mathbb{R}^{N+1 \times (N+1)L}$, such that $\sum_{n=0}^N (y_n - \mathbf{h}_n^T \mathbf{a}_n)^2 = \|\mathbf{y} - \mathbf{M}\mathbf{a}\|_2^2$.

Define the "difference" vector $\mathbf{d}_n \in \mathbb{R}^L$ as

$$\mathbf{d}_n = \begin{cases} \mathbf{a}_n, & \text{if } n = 0 \\ \mathbf{a}_n - \mathbf{a}_{n-1}, & \text{otherwise} \end{cases} \tag{5}$$

and $\mathbf{d} := [\mathbf{d}_0^T, \mathbf{d}_1^T, \ldots, \mathbf{d}_N^T]^T \in \mathbb{R}^{(N+1)L}$. Observe that $\mathbf{d}_n = \mathbf{0}_L$ for $n > 0$ if and only if there is no change in the TV-AR coefficients between time instants $n-1$ and n. Clearly, it is possible to recover $\{\mathbf{a}_n\}_{n=0}^N$ from $\{\mathbf{d}_n\}_{n=0}^N$ since $\mathbf{a}_n = \sum_{n'=0}^n \mathbf{d}_{n'}$. Let $\mathbf{T} \in \mathbb{R}^{N+1 \times N+1}$ denote a lower triangular matrix with all nonzero entries equal to one and $\mathbf{X} := \mathbf{M}(\mathbf{T} \otimes \mathbf{I}_L) \in \mathbb{R}^{(N+1) \times (N+1)L}$, having the following structure:

$$\mathbf{X} = \begin{bmatrix} \mathbf{h}_0^T & \mathbf{0}_L^T & \cdots & \cdots & \mathbf{0}_L^T \\ \mathbf{h}_1^T & \mathbf{h}_1^T & \mathbf{0}_L^T & \cdots & \mathbf{0}_L^T \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{h}_{N-1}^T & \mathbf{h}_{N-1}^T & \mathbf{h}_{N-1}^T & \mathbf{h}_{N-1}^T & \mathbf{0}_L^T \\ \mathbf{h}_N^T & \mathbf{h}_N^T & \mathbf{h}_N^T & \mathbf{h}_N^T & \mathbf{h}_N^T \end{bmatrix}. \tag{6}$$

Since $\mathbf{a} = (\mathbf{T} \otimes \mathbf{I}_L)\mathbf{d}$, an equivalent formulation of (3) in terms of $\{\mathbf{d}_n\}_{n=0}^N$ can be given as [cf. (3)]

$$\{\widehat{\mathbf{d}}_n\}_{n=0}^N := \arg \min_{\{\mathbf{d}_n\}_{n=0}^N} \left[\frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{d}\|_2^2 + \mu \sum_{n=1}^N \delta_{\mathbf{0}_L}(\mathbf{d}_n)\right]. \tag{7}$$

What makes the formulation in (7) challenging is the non-convex regularization term. The latter "pushes" most of the $\{\mathbf{d}_n\}_{n=1}^N$ vectors toward $\mathbf{0}_L$, while $\mathbf{d}_0$ is not penalized. As a consequence, the vector $\widehat{\mathbf{d}} := [\widehat{\mathbf{d}}_0^T, \widehat{\mathbf{d}}_1^T, \ldots, \widehat{\mathbf{d}}_N^T]^T$ is *group* sparse, and the non-zero group indexes correspond to the change instants of the TV-AR coefficients. Upon the basis of recent advances in model selection and compressive sampling [16], in [1] a convex relaxation of the cost in (7) was proposed.

The advocated relaxation entails the so-called group Lasso [16], and the convex problem for the identification of TV-AR models is

$$\{\widehat{\mathbf{d}}_n\}_{n=0}^N = \arg \min_{\{\mathbf{d}_n\}_{n=0}^N} \left[\frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{d}\|_2^2 + \lambda \sum_{n=1}^N \|\mathbf{d}_n\|_2\right] \tag{8}$$

where λ is a positive tuning parameter. It is known that the group Lasso regularization encourages group sparsity; that is, $\widehat{\mathbf{d}}_n = \mathbf{0}_L$ for most $n > 0$ [16], and the larger the λ, the sparser the $\widehat{\mathbf{d}}$.

Since the problem in (8) is convex, general purpose interior-point methods can in principle be used. Nevertheless, the large size prevents the usage of these methods. In [1], a block-coordinate descent approach is advocated which enables to solve the problem efficiently since it incurs in a computational complexity per iteration that scales only linearly with the data size.

3. JOINT SEGMENTATION AND SPARSE RESIDUALS

Recent advances have shown that imposing sparsity in the LP model residuals enables a more effective decoupling of the vocal tract transfer function [5]. Sparse residuals are particularly appropriate for the analysis of voiced speech. In particular, compared to the traditional ℓ_2-norm minimization, the cost function associated with the ℓ_1-norm minimization downweights the impact of the spiky excitation associated with voiced speech on the LP estimates [5].

The formulation in (8) can be modified to cope with the framework in [5] for sparse linear prediction. Substiting the ℓ_2-norm in the first term of the cost function in (8) with the ℓ_1-norm to account for sparsity in the model residuals, yields

$$\{\widehat{\mathbf{d}}_n\}_{n=0}^N = \arg \min_{\{\mathbf{d}_n\}_{n=0}^N} \left[\|\mathbf{y} - \mathbf{X}\mathbf{d}\|_1 + \lambda \sum_{n=1}^N \|\mathbf{d}_n\|_2\right]. \tag{9}$$

Solving the convex optimization problem in (9) allows for joint segmentation and sparse LP model residual. Indeed, minimizing the ℓ_1-norm of the residual term imposes that most of the entries in $\mathbf{y}-\mathbf{X}\widehat{\mathbf{d}}$ are close to zero. Unfortunately, the result in [15] cannot be directly applied to establish convergence of the coordinate descent algorithm for the cost function in (9). Indeed, the term $\|\mathbf{y}-\mathbf{X}\mathbf{d}\|_1$ is non-differentiable and non-separable coordinate-wise. Nevertheless, borrowing certain results from robust statistics, the problem in (9) can be recast in such a way that a modified block-coordinate descent method can be applied [4].

3.1. Huber cost function and sparse residuals

A generalization of the sparsity-promoting ℓ_1-norm regularization is the so-called Huber cost function defined as follows [7]:

$$\rho_\varepsilon(r) := \begin{cases} \frac{1}{2}r^2, & \text{if } |r| \leq \varepsilon \\ \varepsilon|r| - \frac{\varepsilon^2}{2}, & \text{otherwise.} \end{cases} \tag{10}$$

For large ε the Huber cost function resembles the LS cost, while for small ε, the Huber cost coincides with the ℓ_1-norm regularization and promotes sparse LP model residuals. While the ℓ_2-norm is more appropriate for unvoiced speech, the ℓ_1-norm has been used for voiced speech. Clearly, the Huber cost function aims at a balance between the two norms, and can be used for both voiced and unvoiced speech.

Substituting the ℓ_1-norm in (9) with the Huber cost yields

$$\{\widehat{\mathbf{d}}_n\}_{n=0}^N = \arg\min_{\{\mathbf{d}_n\}_{n=0}^N} \left[\sum_{n=0}^N \rho_\varepsilon(y_n - \mathbf{x}_n^T\mathbf{d}) + \lambda\sum_{n=1}^N \|\mathbf{d}_n\|_2\right] \tag{11}$$

where $\mathbf{x}_n^T$ represents the n-th row of $\mathbf{X}$. Similarly to the cost in (9), that in (11) imposes quasi-sparse model residuals for sufficiently small ε.

Proposition 1. *The problem in* (11) *is equivalent to the following convex optimization problem:*

$$[\{\widehat{\mathbf{d}}_n\}_{n=0}^N, \widehat{\mathbf{o}}] = \arg\min_{\{\mathbf{d}_n\}_{n=0}^N, \mathbf{o}} \left[\frac{1}{2}\|\mathbf{y}-\mathbf{X}\mathbf{d}-\mathbf{o}\|_2^2 + \lambda\sum_{n=1}^N \|\mathbf{d}_n\|_2 + \varepsilon\|\mathbf{o}\|_1\right] \tag{12}$$

where $\mathbf{o} := [o_0, o_1 \ldots, o_N]^T \in \mathbb{R}^{N+1}$.

Proposition 1 states that, introducing the auxiliary variables $\mathbf{o}$, the optimization in (11) can be recast as the problem in (12). Entries of the model residual that are supposed to be pushed to zero by the Huber cost in (11), corresponds to the zeros in $\mathbf{o}$. On the other hand, the non-zero elements of $\mathbf{o}$ corresponds to observations with large model residuals, i.e., the so-called outliers.

Surprisingly, the problem in (12) admits a block-coordinate descent solver since the non-differentiable part is separable coordinate-wise, and the result in [15] can be applied. Indeed, consider the objective function

$$J(\mathbf{d},\mathbf{o}) := \frac{1}{2}\|\mathbf{y}-\mathbf{X}\mathbf{d}-\mathbf{o}\|_2^2 + \lambda\sum_{n=1}^N \|\mathbf{d}_n\|_2 + \varepsilon\|\mathbf{o}\|_1. \tag{13}$$

Keeping $\mathbf{o}$ fixed, and skipping irrelevant terms, yields

Algorithm 1 Block-Coordinate Descent for sparse residual

Given $\{\mathbf{R}_{n:N}, \mathbf{r}_{n:N}\}_{n=0}^N$, $\mathbf{d}_n^{(0)} = \mathbf{0}_L$ for $n = 1, \ldots, N$ and $\mathbf{o}^{(0)} = \mathbf{0}_N$
for $i > 0$ **do**
  $\mathbf{r}^{(i-1)} = \mathbf{r} - \mathbf{X}^T\mathbf{o}^{(i-1)}$
  for $n = 0, 1, \ldots, N$ **do**
    if $n = 0$ **then**
      $\mathbf{c}_0^{(i)} = \mathbf{0}_L$
      $\mathbf{s}_0^{(i)} = \sum_{n=1}^N \mathbf{R}_{n:N}\mathbf{d}_{n-1}^{(i-1)}$
      $\mathbf{g}_0^{(i)} = \mathbf{s}_0^{(i)} - \mathbf{r}^{(i-1)}$
      $\mathbf{d}_0^{(i)} = -\mathbf{R}_{0:N}^{-1}\mathbf{g}_0^{(i)}$
    else
      $\mathbf{c}_n^{(i)} = \mathbf{c}_{n-1}^{(i)} + \mathbf{d}_{n-1}^{(i)}$
      $\mathbf{s}_n^{(i)} = \mathbf{s}_{n-1}^{(i)} - \mathbf{R}_{n:N}\mathbf{d}_n^{(i-1)}$
      $\mathbf{g}_n^{(i)} = \mathbf{R}_{n:N}\mathbf{c}_n^{(i)} + \mathbf{s}_n^{(i)} - \mathbf{r}_{n:N} - \sum_{m=n}^N \mathbf{h}_m o_m^{(i-1)}$
      if $\|\mathbf{g}_n^{(i)}\|_2 \leq \lambda$ **then**
        $\mathbf{d}_n^{(i)} = \mathbf{0}_L$
      else
        $\mathbf{d}_n^{(i)} = \arg\min_{\mathbf{d}_n \in \mathbb{R}^L} \left[\frac{1}{2}\mathbf{d}_n^T\mathbf{R}_{n:N}\mathbf{d}_n + \mathbf{d}_n^T\mathbf{g}_n^{(i)} + \lambda\|\mathbf{d}_n\|_2\right]$
      end if
    end if
  end for
  $\mathbf{e}^{(i)} = \mathbf{y} - \mathbf{X}\mathbf{d}^{(i)}$
  for $n = 0, 1, \ldots, N$ **do**
    $o_n^{(i)} = \text{shrink}(e_n^{(i)}, \varepsilon)$
  end for
end for

$$J(\mathbf{d}, \mathbf{o}^{(i-1)}) = \frac{1}{2}\mathbf{d}^T\mathbf{R}\mathbf{d} - \mathbf{d}^T\mathbf{r}^{(i-1)} + \lambda\sum_{n=1}^N \|\mathbf{d}_n\|_2 \tag{14}$$

with $\mathbf{r}^{(i-1)} = \mathbf{r} - \mathbf{X}^T\mathbf{o}^{(i-1)}$. Exploiting the results in [1] the cost in (14) can be minimized with block-coordinate descent. Unlike the cost in [1], the term $\mathbf{r}^{(i-1)}$ changes with the iteration index, and it has to be re-evaluated at the beginning of each iteration. Nevertheless, its evaluation is not expected to be time consuming since $\mathbf{o}^{(i-1)}$ has only a few non-zero elements. Therefore, (14) can be minimized over each group $\mathbf{d}_n$ using the results in [13].

Keeping $\mathbf{d}$ fixed, and discarding irrelevant terms, the cost in (13) becomes

$$J(\mathbf{d}^{(i)}, \mathbf{o}) = \frac{1}{2}\|\mathbf{e}^{(i)} - \mathbf{o}\|_2^2 + \varepsilon\|\mathbf{o}\|_1 \tag{15}$$

where $\mathbf{e}^{(i)} := \mathbf{y} - \mathbf{X}\mathbf{d}^{(i)}$. The cost in (15) can be separated coordinate-wise, i.e.,

$$J(\mathbf{d}^{(i)}, \mathbf{o}) = \sum_{n=0}^N \left[\frac{1}{2}\|e_n^{(i)} - o_n\|^2 + \varepsilon\|o_n\|\right]. \tag{16}$$

Each term in (16) can be minimized via the shrinkage operator [13], i.e., $o_n^{(i)} = \text{shrink}(e_n^{(i)}, \varepsilon)$ where

$$\text{shrink}(x, \varepsilon) := \begin{cases} 0 & \text{if } |x| < \varepsilon \\ \text{sign}(x)(|x| - \varepsilon) & \text{otherwise.} \end{cases} \tag{17}$$

Summing up, the problem in (12) admits the block-coordinate descent solver described in Alg. 1.

The ensuing proposition is a direct consequence of the results in [15].

Proposition 2. *The iterates* $\mathbf{d}^{(i)} := [\mathbf{d}_0^{(i)T}, \mathbf{d}_1^{(i)T}, \ldots, \mathbf{d}_N^{(i)T}]^T$, *and* $\mathbf{o}^{(i)}$ *obtained by* Algorithm 1 *converge to the global minimum of* (12).

4. SIMULATED TESTS

The merits of the novel approaches to identify change-points in TV-AR processes are assessed via numerical simulations using real speech data.

The dataset is depicted in Fig. 1, and it represents 0.125 s of voiced speech sampled at 8000 Hz.

The joint segmentation and LP identification method in (8), solved with block-coordinate descent method in [1] has been tested. As suggested in [1], the regularization parameter λ is selected as 10 % of the maximum λ_{max} that would entail a constant AR model. A model order $L = 20$ has been selected. The first four LP coefficients are depicted in Fig. 2. Observe that the LP coefficients change slowly (sharper change can be obtained resorting to iterative weighting to enhance sparsity as advocated in [1]). The joint segmentation and sparse LP model residual identification method in (11) was tested. The problem in (11) was solved with the block-coordinate descent method in Algorithm 1. Figure 3 depicts the first four LP coefficients, which change in a sharper way and three long segments are obtained (plus some short spurious changes in the transitions).

A snapshot of the LP model residuals within a stationary segment is depicted in Fig. 4. Figure 4 a) (top) shows the model residual of the method in (8) while Fig. 4 b) (bottom) shows the model residual of (11). From the latter figure it is clear that the spiky excitation can be identified as well as its frequency which might aid speech recognition and analysis.

5. CONCLUSION

A novel method has been developed in this paper for identification of piecewise-constant TV-AR models by exploiting recent advances in robust statistics, variable selection, and compressive sampling. While traditional techniques consist in regularizing a least-squares criterion with the total number of coefficient changes, the novel method relies on a convex regularization function, which resembles the group Lasso and can afford efficient implementation using block-coordinate descent iterations. To cope with modern trends in speech modeling, the method was extended to cope with sparsity in the LP model residual, a case of interest for modeling voiced speech. The latter problem can be solved iteratively via block-coordinate descent and its computational burden scales only linearly in the data size. The methods have been tested using real speech data, and its advantages with respect to classical methods have been highlighted.

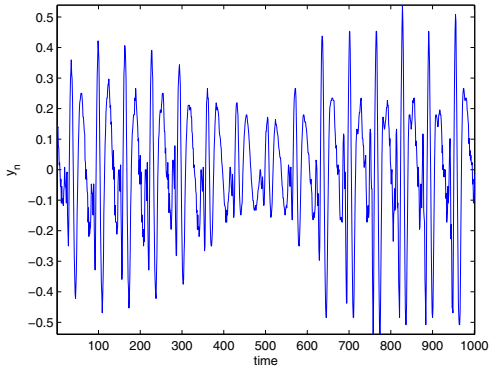

Fig. 1. Voiced speech data.

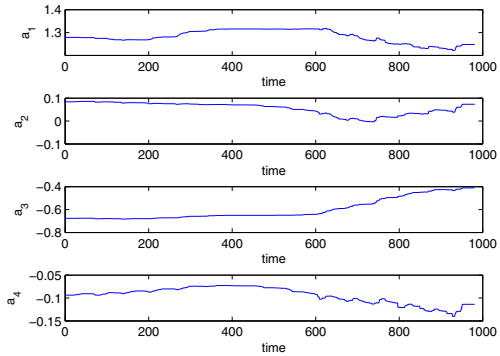

Fig. 2. First four AR coefficients identified via classical ℓ_2-norm cost.

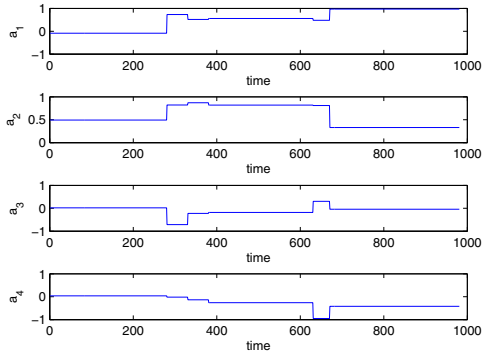

Fig. 3. First four AR coefficients identified via Huber cost.

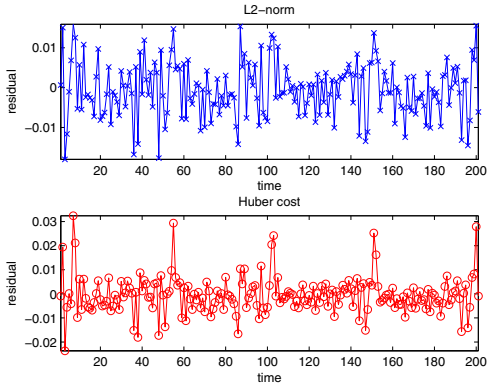

Fig. 4. Residuals of the ℓ_2-norm cost (top) and Huber cost (bottom).

6. REFERENCES

[1] D. Angelosante and G. B. Giannakis, "Group Lassoing Changes in Piecewise-Stationary Autoregressive Processes," *EURASIP Journal on Advances in Signal Processing*, March 2012.

[2] L. Boysen, A. Kempe, V. Liebscher, A. Munk, and O. Wittich, "Consistencies and rates of convergence of jump-penalized least-squares estimators," *Annals of Statistics*, vol. 37, no. 1, pp. 157–183, Feb. 2009.

[3] D. L. Donoho, "Compressed Sensing," *IEEE Trans. on Inf. Th.*, pp. 1289-1306, Vol. 52, 2006.

[4] J.-J. Fuchs, "An inverse problem approach to robust regression," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Phoenix, AZ, 1999, pp. 18091812.

[5] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, M. Moonen, "Sparse Linear Prediction and Its Applications to Speech Processing", *IEEE Transactions in Audio, Speech and Language Processing*, vol. 20, no. 5, pp. 1644-1657, July 2012.

[6] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, M. Moonen, "Stable 1-norm Error Minimization Based Linear Predictors for Speech Modeling", *submitted to IEEE Trans. in Audio, Speech and Language Processing*, 2013.

[7] P. Huber, *Robust Statistics*, Robust Statistics, John Wiley, New York, 1981.

[8] M. Lavielle, "Optimal segmentation of random processes," *IEEE Transactions on Signal Processing*, vol. 46, no. 5, pp. 1365–1373, May 1998.

[9] ——, "Using penalized contrasts for the change-point problem," *Signal Processing*, vol. 85, no. 8, pp. 1501–1510, Aug. 2005.

[10] M. Lavielle and E. Moulines, "Least-squares estimation of an unknown number of shifts in a time series," *Journal of Time Series Analysis*, vol. 21, no. 1, pp. 33–59, Jan. 2000.

[11] E. Lebarbier, "Detecting multiple change-points in the mean of Gaussian process by model selection," *Signal Processing*, vol. 85, no. 4, pp. 717–736, Apr. 2005.

[12] S. Kay, *Fundamentals of Statistical Signal Processing, Volume 2:Detection Theory*, Prentice Hall, 1998.

[13] A. Puig, A. Wiesel, and A. Hero, "A multidimensional shrinkage-thresholding operator," in *Proceedings of the 15th Workshop on Statistical Signal Processing*, Cardiff, UK, Sep. 2009.

[14] P. Stoica and R. L. Moses, *Introduction to Spectral Analysis*. New Jersey: Prentice-Hall, 1997.

[15] P. Tseng, "Convergence of a block coordinate descent method for nondifferentiable minimization," *Journal of Optimization Theory and Applications*, vol. 109, no. 3, pp. 475–494, Jun. 2001.

[16] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society, Series B*, vol. 68, no. 1, pp. 49–67, 2006.