Sparse Linear Prediction and Its Applications to Speech Processing

Daniele Giacobello

Aalborg Universitet, Denmark

September, 2009

Motivation

- Linear prediction (LP) is an integral part of many modern speech processing systems.
- Applications ranging from Coding, Synthesis, Spectral Analysis and Recognition.
- The prediction coefficients are usually found through 2-norm minimization of the prediction error.
- Many examples where the 2-norm in LP analysis does not work well.

Why sparse linear prediction?

- Provides interesting modeling properties in many speech applications.
- More synergistic approach to multistage time-domain speech compression.
- Why not! ...new formulations for the LP problem may be of general interest! (e.g. ECG)

Outline

Speech production model

- A sample of speech $x(n)$ is written as a linear combination of K past samples:

$$x(n) = \sum_{k=1}^{K} a_k x(n-k) + e(n), \quad 0 < n \leq N,$$

- The speech production model in matrix form becomes:

$$\mathbf{x} = \mathbf{X}\mathbf{a} + \mathbf{e}$$

where

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

General optimization framework

- Class of problems considered are covered by the optimization problem associated with finding the prediction coefficient vector $\mathbf{a} \in \mathbb{R}^K$ from a set of observed real samples $x(n)$ for $n = 1, \ldots, N$ so that the $p-$norm of the prediction error is minimized:

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

How to choose p, k and γ?

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

- *maximum a posteriori* (MAP) approach for finding **a** under the assumptions that **a** has a Generalized Gaussian Distribution:

$$\begin{aligned} \mathbf{a}_{\mathbf{MAP}} &= \arg\max_{\mathbf{a}} f(\mathbf{x}|\mathbf{a})g(\mathbf{a}) \\ &= \arg\max_{\mathbf{a}} \{\exp(-\|\mathbf{x} - \mathbf{X}\mathbf{a}\|_p^p) \exp(-\gamma\|\mathbf{a}\|_k^k)\}. \end{aligned}$$

- γ is related to the prior knowledge of **a**
- Sparseness is often measured as the cardinality (so-called $\|\cdot\|_0$).
- The $\|\cdot\|_1$ is used as a convex relaxation to a problem of combinatorial nature (NP-hard)

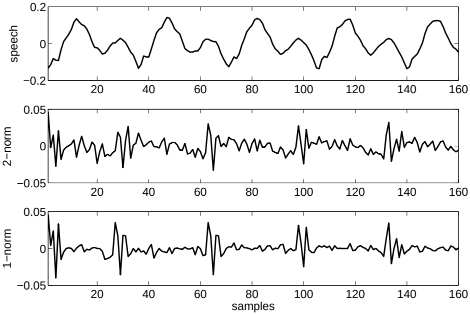
Problem Definition

$$\min_{\mathbf{a}} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_1 .$$

- ML approach when the error sequence is considered to be a set of i.i.d. Laplacian random variables.
- Outperforms the 2-norm in finding a more proper linear predictive representation in voiced speech.
- Better statistical fitting also in unvoiced speech (...and sparser residual!).
- Helpful against over-emphasis on peaks in AR spectral estimation.

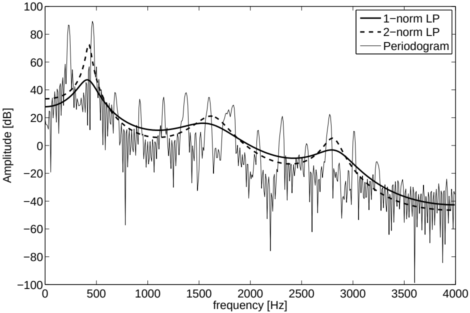
Example

An excitation similar to the impulse response of the long term predictor is found for voiced speech when we look for a sparse residual.



Example

The lower emphasis on peaks in the envelope, when 1-norm minimization is employed, is a direct consequence of the ability to retrieve the spiky pitch excitation.



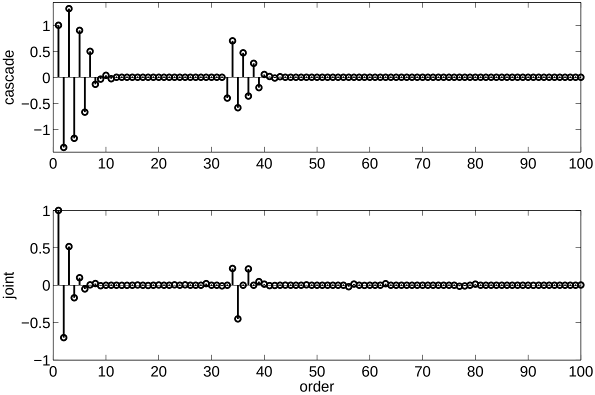
Definition

$$\min_{\mathbf{a}} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_1 + \gamma\|\mathbf{a}\|_1.$$

- With a high prediction order the resulting coefficient vector **a** will be highly sparse.
- An AR filter having a sparse structure is an indication that the polynomial can be factored into several terms.
- The purpose of the high order sparse predictor is to model the *whole* spectrum, i.e., the spectral envelope and the spectral harmonics.
- Strong ability of high order LP to resolve closely spaced sinusoids also helpful in audio processing.

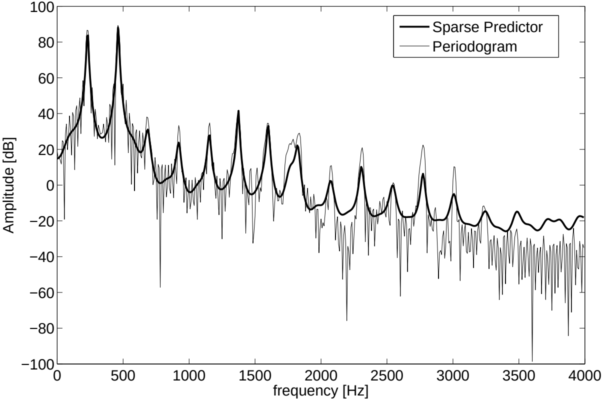
Example

The prediction coefficients vector vector is similar to the multiplication of the short-term prediction filter and long-term prediction filter usually obtained in cascade.

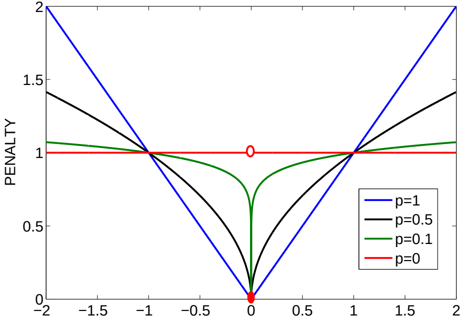


Example

Spectral modeling properties of a high order sparse predictor with only nine nonzero coefficients.



Reducing the 1-norm 0-norm mismatch



- Reweighted 1-norm minimization can help balancing the dependence on the magnitude of the 1-norm.
- Changing the cost function and moving the problem towards the 0-norm minimization with convex tools.

Definition

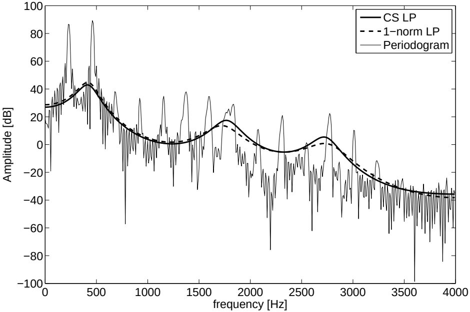
- Exploiting prior knowledge about the sparsity of the signal $\mathbf{x}$, a limited number of random projections are sufficient to recover our predictors and sparse residual with high accuracy. With *known* predictor:

$$\hat{\mathbf{r}} = \arg\min_{\mathbf{r}} \|\mathbf{r}\|_1 \quad \text{s.t.} \quad \Phi\mathbf{x} = \Phi\mathbf{H}\mathbf{r} \tag{1}$$

- To adapt CS principles to the estimation of the predictor as well, we can consider the relation between the synthesis matrix $\mathbf{H}$ and the analysis matrix $\mathbf{A}$ ($\mathbf{A} = \mathbf{H}^{+}$):

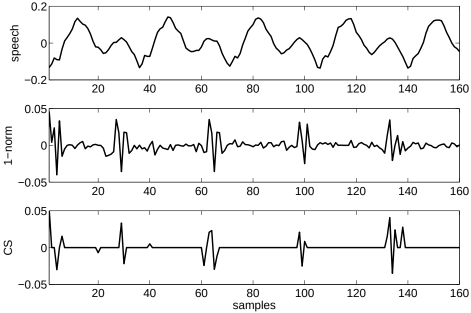
$$\min_{\mathbf{a},\mathbf{r}} \|\mathbf{r}\|_1 \quad \text{s.t.} \quad \Phi\mathbf{r} = \Phi(\mathbf{x} - \mathbf{X}\mathbf{a}). \tag{2}$$

Compressed Sensing in Sparse LP



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

Compressed Sensing in Sparse LP



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

Problem Definition

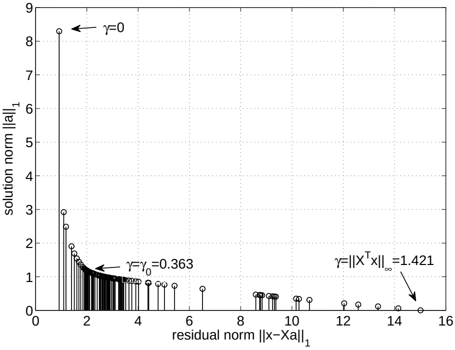
- Better statistical modeling in the context of speech analysis creates an output that offers better coding properties.
- Introducing sparsity constraints in a linear prediction scheme both on the residual and on the high order prediction vector:

$$\min_{\mathbf{a}} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_1 + \gamma \|\mathbf{a}\|_1.$$

- Efficient multipulse residual encoding.
- Robust statistical method for the joint estimation of the short-term and long-term predictors.

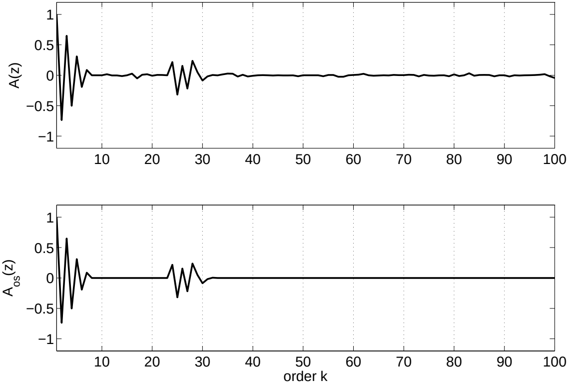
Choosing the regularization parameter γ

- Point of maximum curvature of the modified *L*-curve $(\|\mathbf{x} - \mathbf{X}\mathbf{a}_\gamma\|_1, \|\mathbf{a}_\gamma\|_1)$



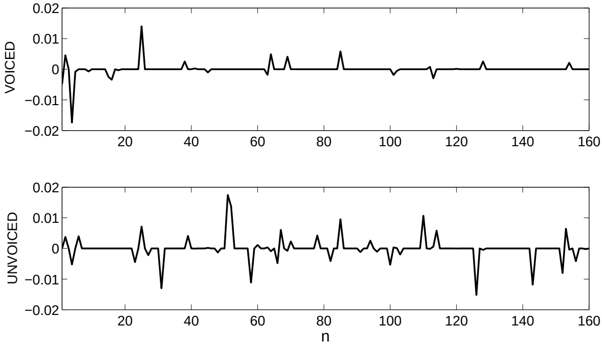
Factorization of the high order predictor

- Removal of spurious quasi-zero components removed through model order selection or reweighted 1-norm



Encoding of the residual

- Use of multipulse encoding (MPE) techniques efficient with the characteristics of the residual.



Discussion

- Possibility Variable rate coding (model order selection and intrinsic V/UV classification).
- Sparse residual allows a more compact representation.
- Joint estimation of short-term and long-term predictors.
- Smoother spectral envelopes robust to quantization.
- Lower order AR models.
- Pitch lag estimation is more accurate.
- Pitch-independence and shift-independence of the estimated predictor.
- NOISE ROBUST!!

Stability

- Stability is not guaranteed.
- Reducing the numerical range of the shift-operator for intrinsic stable solutions.
- Exploiting LSFs interlacing properties.
- Constrained 1-norm based on the alternative Cauchy bound.

Computational costs

- The problem seen are computationally expensive (e.g. 1-norm minimization costs about 20-25 least squares problems).
- Primal-dual interior point methods can help reducing the costs.
- Compressed Sensing reduces the number of constraints.
- Much of the total computational cost in a speech coder is saved by the “one-step” procedure.
- It is a highly structured problem!

Conclusions

- Changing the statistical assumptions in LP brought us to define new formulations of a well-know problem.
- The methods presented are very attractive for the analysis and coding of speech signals outperforming traditional LP.
- Convex optimization algorithms and sparse representation are booming: new powerful estimator can be easily created using these tools.