Sparse Linear Predictors for Speech Processing

Daniele Giacobello[1,2] Mads Græsbøll Christensen[1] Joachim Dahl[1] Søren Holdt Jensen[1] Marc Moonen[2]

[1]Dept. of Electronic Systems (ES-MISP), Aalborg University, Aalborg, Denmark

[2]Dept. of Electrical Engineering (ESAT-SCD), Katholieke Universiteit Leuven, Leuven, Belgium

1 Motivation

- There are many examples where the 2-norm error minimization in LP analysis does not work well, i.e. when the excitation is not Gaussian.
- In this case the usual approach is to find coefficients for the short-term and long-term signal correlation in two different steps. This obviously leads to inherently suboptimal solutions.
- From an estimation point of view, this suboptimality leads us to define a more appropriate analysis model.
- The 2-norm minimization shapes the residual into variables that exhibit Gaussian-like characteristics; however, so-called sparse coding techniques have been used, for example, in early GSM standards and more recently also in audio coding to quantize the residual. In these techniques the residual is encoded using only few non-zero pulses.
- In this case and quantization-wise in general, we can reasonably assume that the optimal predictor is not the one that minimizes the 2-norm but the one that leaves the fewest non-zero pulses in the residual, i.e. the sparsest one.

2 Fundamentals

- The class of problems considered as those 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 $n = 1, \dots, N$ so that the error is minimized:

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

where $\|\cdot\|_p$ is the p-norm defined as $\|\mathbf{x}\|_p = (\sum_{n=1}^{N} |x(n)|^p)^{\frac{1}{p}}$ for $p \geq 1$.
- The question then is how to choose p, k and γ and how to perform the associated minimization, depending on the kind of application.
- Sparseness is often measured as the cardinality (so-called 0-norm $\|\cdot\|_0$). Unfortunately this is a combinatorial NP-hard problem. Instead of the cardinality measure, we then use the more tractable 1-norm $\|\cdot\|_1$.
- The regularization term γ in our mathematical framework is somehow related to the prior knowledge we have of the coefficients vector $\mathbf{a}$, therefore is clearly the *maximum a posteriori* (MAP) approach for finding $\mathbf{a}$ under the assumptions that $\mathbf{a}$ has a Generalized Gaussian Distribution:

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

3 Sparse Linear Predictors

3.1 Finding a Sparse Residual

- 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.
- Helpful against over-emphasis on peaks and cancellation of errors in voiced speech.
- Invariant to pitch and phase shifts as it goes after the true nature of the formant excitation.

3.2 Finding Sparse Coefficients

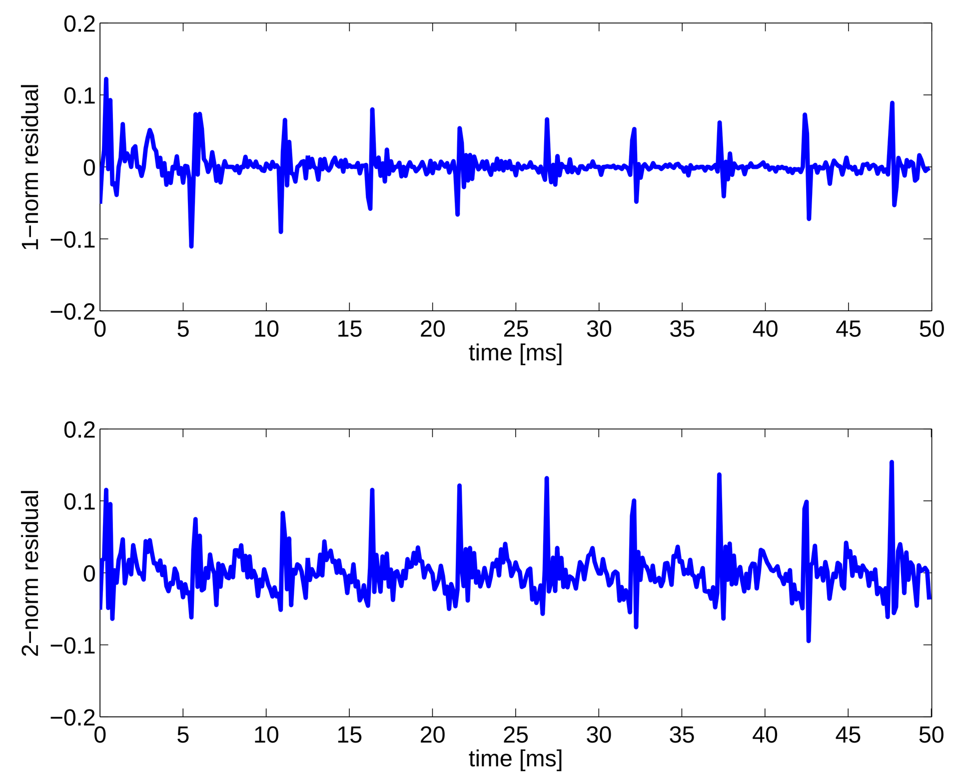
- Problem definition:

$$\min_{\mathbf{a}} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_2^2 + \gamma\|\mathbf{a}\|_1.$$

- With a high prediction order the resulting coefficient vector $\mathbf{a}$ will be highly sparse.
- An AR filter having a sparse structure is an indication that the polynomial can be factored into several terms where one of these exhibits comb-like characteristics: the long term predictor $P(z) = 1 - g_p z^{-T_p}$ often used in speech processing is an example.

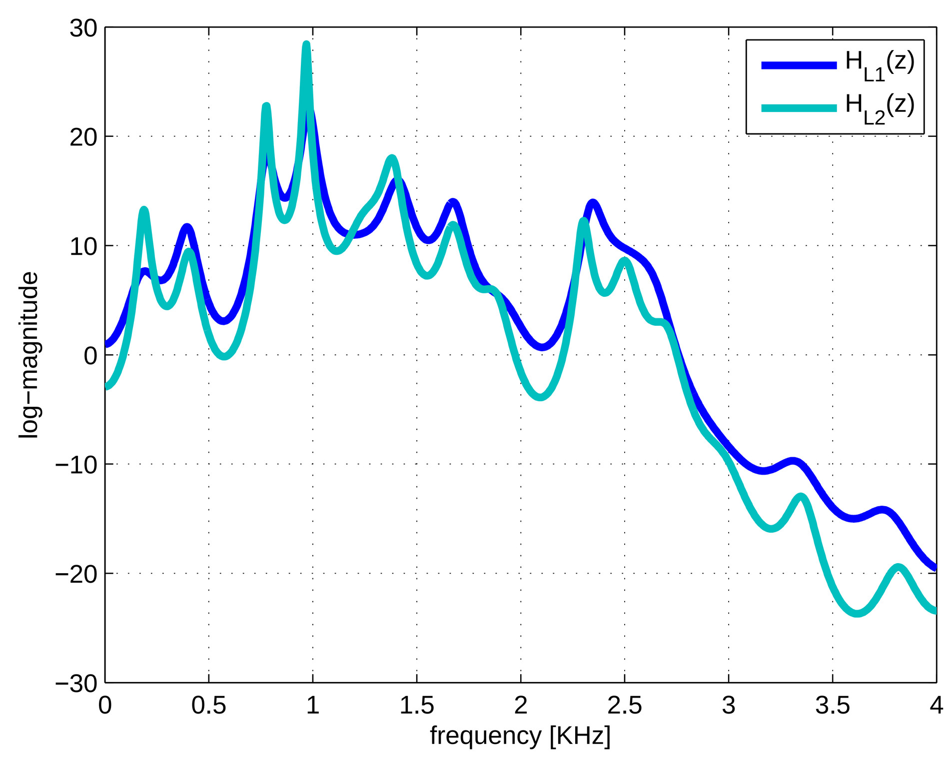
4 Numerical Experiments

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



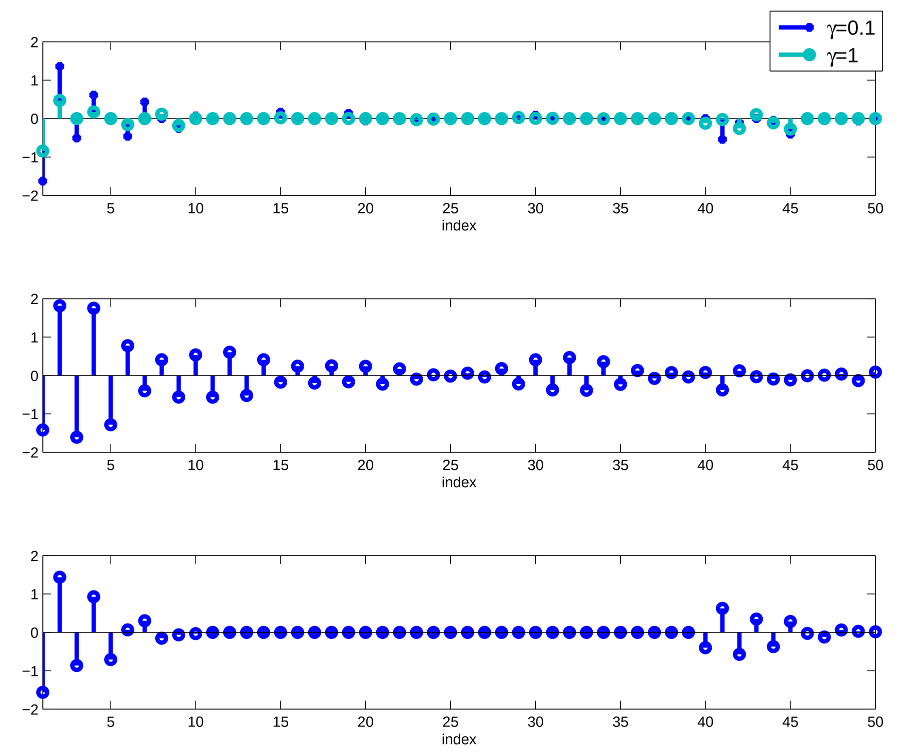
Residuals for 1-norm and 2-norm minimization.

It is also easy to see that the 2-norm minimization introduces high emphasis on peaks in its effort to reduce large errors: in this case the outliers due to the pitch excitation, as we can see clearly in the figure below.



Frequency response of the filters obtained with 1-norm and 2-norm minimization.

In the second approach, we were able to see that using $0.1 \leq \gamma \leq 1$, the predictive vector $\mathbf{a}$ is similar to the multiplication of the short-term prediction filter and long-term prediction filter usually obtained in cascade.



Comparison of the prediction coefficients (excluding the 0^{th}-order) obtained with our algorithm (top), with usual LP (order 50) and with the convolution of the short-term and long-term coefficients vectors.

5 Discussion

The main issues of the absolute error approach have been pointed by Denoël and Solvay:

- Non-uniqueness of the solution.
- Stability not guaranteed.
- Computationally expensive

We can argue that:

- Due to the convexity of the cost function, we can easily state that all the possible multiple solutions will still be optimal.
- By choosing an appropriate windowing of the analyzed signal the percentage of non-stable filters was less then 0.1% in over 10,000 frames analyzed.
- Finding the solution using a modern interior point algorithm showed to be comparable to solving around 10-15 least square problems but further analysis processes can be highly simplified by the characteristics of the output.

6 Conclusion

- These methods are particularly attractive for the analysis and coding of stationary voiced signal but the extension of the obtained results to unvoiced signal seemed to be straightforward and will be subjected to further analysis.
- Considering other convex estimators will easily bring to new studies based on different concepts of sparseness.

References

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

[2] Y. Q. Li, A. Cichocki, S. Amari,"Analysis of sparse representation and blind source separation", *Neural computation*, vol. 16, no.6, pp. 1193-1234, June 2004.

[3] E. Denoël and J.-P. Solvay, "Linear prediction of speech with a least absolute error criterion", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 33(6), pp. 1397–1403, Dec. 1985.

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