High-Order Sparse Linear Predictors for Audio Processing

Daniele Giacobello[1], Toon van Waterschoot[2], Mads Græsbøll Christensen[1], Søren Holdt Jensen[1], Marc Moonen[2]

[1]Aalborg Universitet, Denmark
[2]Katholieke Universiteit Leuven, Belgium

August 24, 2010

Linear Prediction of Speech Signals

- Arguably one of the most successful tools for the analysis and coding of speech signals.
- Analysis: correspondence with modeling the speech production process.
- Coding: interesting attributes like low delay, scalability and, in general, low complexity.

Linear Prediction of Audio Signals

- LP modeling is definitely less popular in audio processing.
- Analysis: the predictor does not necessarily model any physical mechanism that generated the audio signal (ensemble of different sources).
- Coding: general difficulties in the accurate parametrization of audio signals.
- These shortcomings have led the way to a net preference for transform-based audio coders that exploit perceptual models of human hearing.

Motivations

- The all-pole model of the LP filter is generally a quite adequate tool to model the spectral peaks which play a dominant role in perception.
- Properties like low delay, scalability and low complexity make the extension of LP to audio coding also appealing.
- In our recent work, we have shown the benefits of using high-order sparse linear predictors to model the spectrum of voiced speech signals (envelope+harmonics).
- We propose an extension of this work in the case of tonal audio signals.

Outline

Tonal Audio Signal Model

- Spectrum containing a finite number of dominant frequency components:

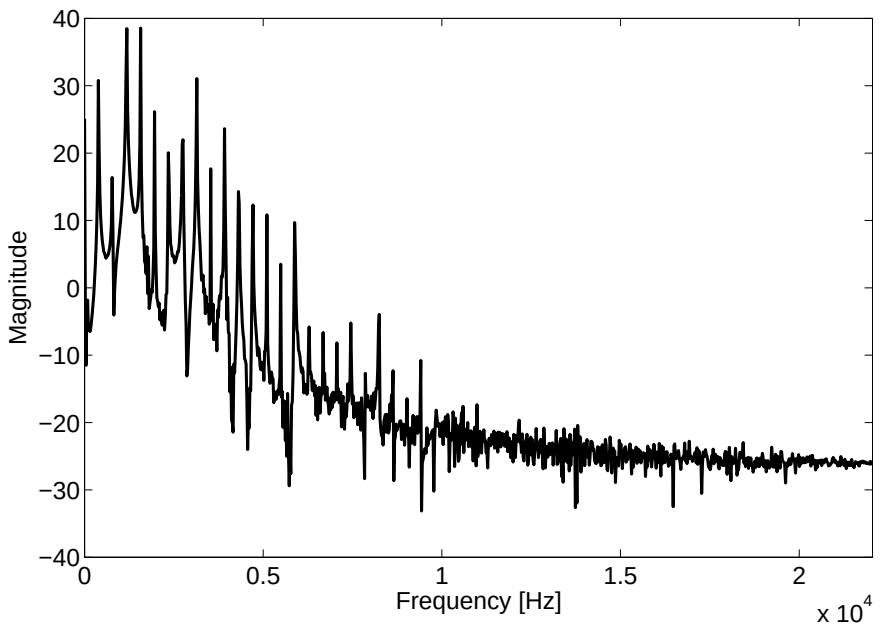
$$x(n) = \sum_{m=1}^{M} \alpha_m \cos(\omega_m n + \phi_m) + r(n), \quad n = 1, \ldots, L,$$

- $r(n)$ contains the nontonal components.
- We will consider $f_s = 44100$ Hz. n normalized with respect to the sampling period $T_s = 1/f_s$.

Monophonic Audio Signals

- It is assumed that all tonal components are harmonically related to a single fundamental frequency f_0:

$$x(n) = \sum_{m=1}^{M} \alpha_m \cos(m\omega_0 n + \phi_m) + r(n), \quad n = 1, \ldots, L.$$



Magnitude spectrum of a monophonic audio signal with $f_0 = 387.6$ Hz.

All-Pole Modeling of Monophonic Audio Signals

- Signal spectrum is made up by two components:
  - a comb-like structure:

$$H_p(z) = \frac{G_p}{P(z)} = \frac{G_p}{1 - pz^{-P}};$$

  - a smooth spectral envelope with low-pass characteristics:

$$H_f(z) = \frac{G_f}{F(z)} = \frac{G_f}{1 - \sum_{k=1}^{N_f} f_k z^{-k}}.$$

- The cascade of the two filters corresponds the multiplication in the z-domain of the their transfer functions:

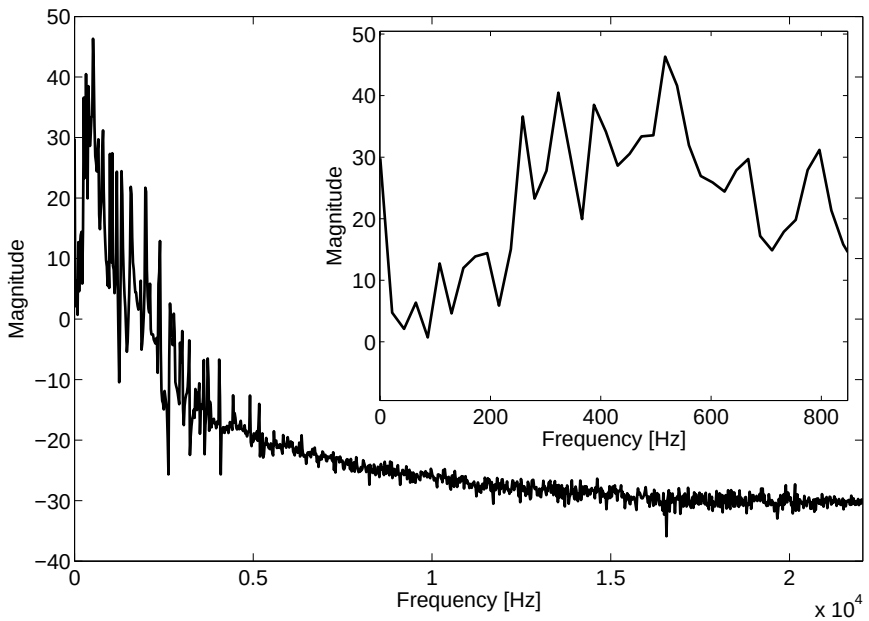
$$H_a(z) = H_p(z)H_f(z) = \frac{G_f G_p}{(1 - \sum_{k=1}^{N_f} f_k z^{-k})(1 - pz^{-P})}.$$

- The signal can therefore be modeled with an order $K \geq P + N_f$ *sparse* all-pole filter $A(z)$ $(1/H_a(z))$.

Polyphonic Audio Signals

- Finite sum of monophonic signals:

$$x(n) = \sum_{m=1}^{M} \left(\sum_{q=1}^{Q_m} \alpha_{m,q} \cos(q\omega_{0,m} n + \phi_{m,q}) \right) + r(n), \quad n = 1, \ldots, L.$$



Magnitude spectrum of a polyphonic audio signal sum of four monophonic signals with ($f_{0,n} = \{258.4, 323.0, 387.6, 516.8\}$).

All-Pole Modeling of Polyphonic Audio Signals

- Presence of a smooth spectral envelope like in the monophonic case.
- The harmonics have now a multi-pitch structure:

$$H_p(z) = \sum_{i=1}^{M} \frac{G_{p_i}}{P_i(z)} = \sum_{i=1}^{M} \frac{G_{p_i}}{1 - p_i z^{-P_i}},$$

that can be approximated by:

$$H_p(z) = \sum_{i=1}^{M} \frac{G_{p_i}}{1 - p_i z^{-P_i}} \approx \frac{G_p}{\prod_{i=1}^{M}(1 - p_i z^{-P_i})}.$$

- The cascade can still be seen as a high-order *sparse* all-pole filter:

$$H_a(z) \approx \frac{G_f G_p}{(1 - \sum_{k=1}^{N_f} f_k z^{-k})(\prod_{i=1}^{M}(1 - p_i z^{-P_i}))}.$$

Fundamentals

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

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

- In matrix form becomes:

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

- Generalized optimization framework:

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

High-Order LP Modeling

- A signal composed of M sinusoids plus (white) noise can be modeled exactly using an ARMA($2M$,$2M$) model[1].
- This model can be arbitrarily closely approximated with an AR model, provided that the model order K is chosen large enough [2].
- When $p = 2$ and $\gamma = 0$:

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_2^2 = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{x},$$

  where $\mathbf{R} = \mathbf{X}^T\mathbf{X}$ is the autocorrelation matrix.
- We will consider $N = 2048$ and $K = 1024$.

---

[1] Y. T. Chan, J. M. M. Lavoie, and J. B. Plant, "A parameter estimation approach to estimation of frequencies in sinusoids," IEEE Trans. ASSP, vol. 29, no. 2, pp. 214-219, 1981.

[2] S. M. Kay, "The effects of noise on the autoregressive spectral estimator," IEEE Trans. ASSP, vol. 27, no. 5, pp.478-485, 1979.

Pitch Prediction

- Only prediction model in which the harmonicity property is exploited.
- A monophonic signal with a pitch period T_0 corresponding to an integer number of sampling periods T_s can be perfectly predicted using the one-tap pitch predictor.
- We will use a 3-tap fractional pitch predictor, efficient in modeling the decreasing comb-like structure of the signals analyzed:

$$P(z) = 1/H_p(z) = 1 + \sum_{i=-1}^{1} p_I z^{-i-P}.$$

- Use of a fractional-delay interpolation filter for non-integer pitch delay P.

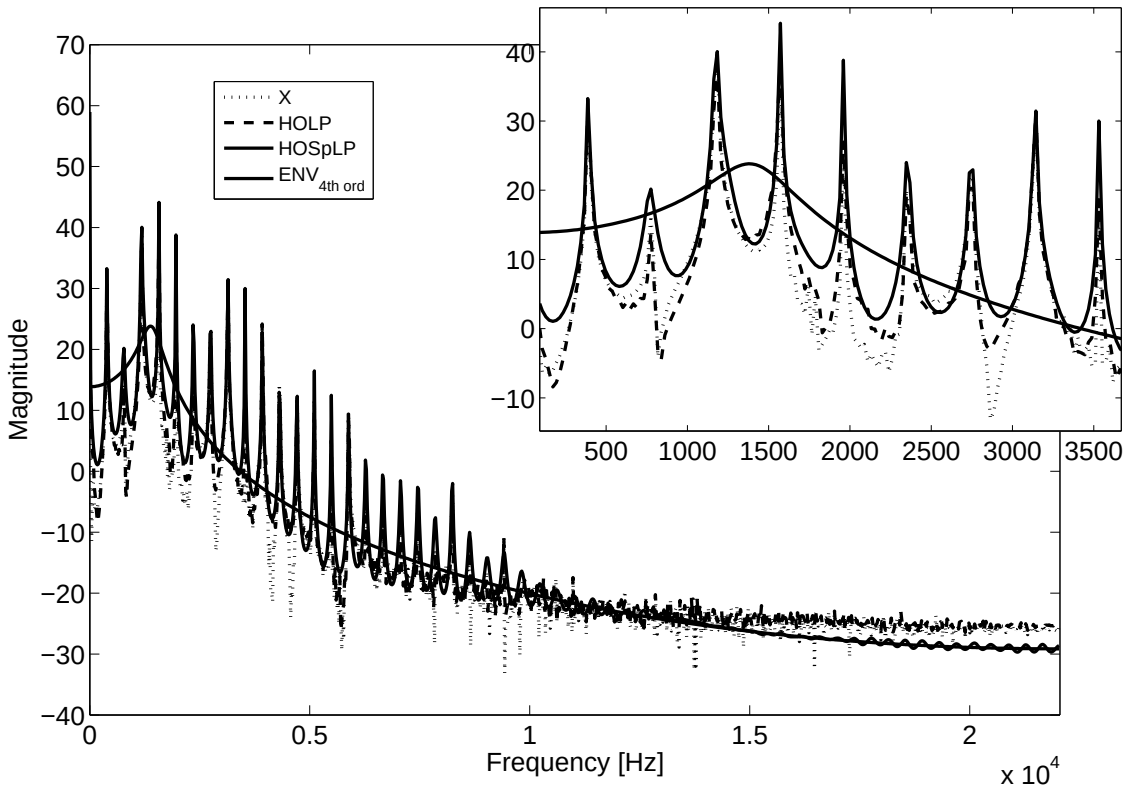
High-Order Sparse LP

- Imposing sparsity constraints on the LP coefficients by relaxing the cardinality constraint (0-norm $\leftarrow$ 1-norm):

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_p^p + \gamma\|\mathbf{a}\|_1.$$

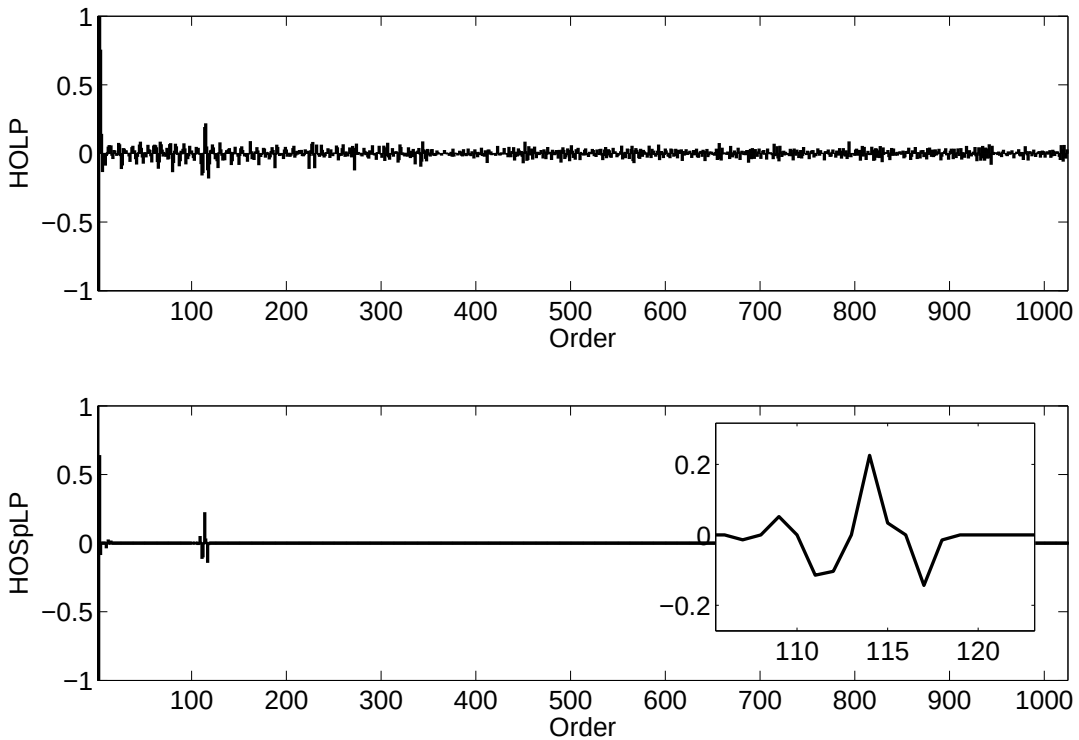
- We will consider $N = 2048$ and $K = 1024$.
- $p = 2$ minimum variance approach, $p = 1$ encourages sparsity also on the residual.
- Meaning of $\gamma\|\mathbf{a}\|_1$:
  - related to the *a priori* knowledge of $\mathbf{a}$ (MAP approach) or *how sparse* the predictor should be;
  - for $K > N/3$ the problem is ill-posed ($\mathbf{X}$ with highly correlated columns), a sparse regularization on the coefficients "trims out" the columns of $\mathbf{a}$ that are redundant for the estimate.

Spectral Modeling



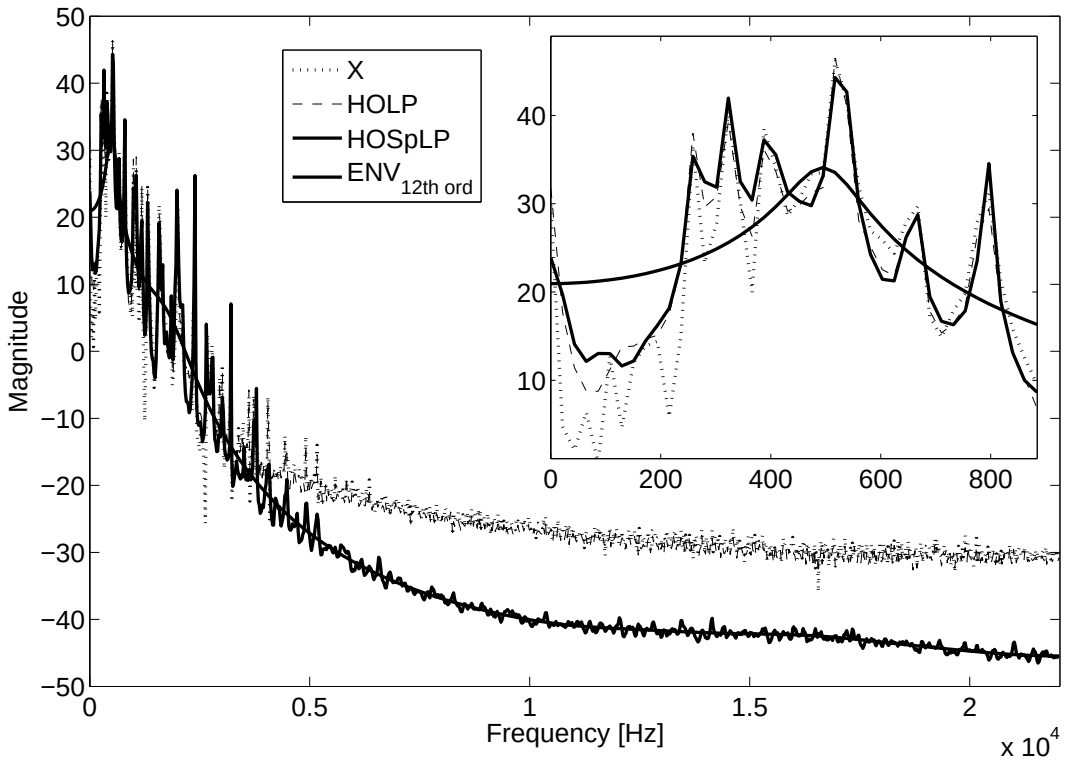
Monophonic audio signal. Frequency response for the all-pole high-order 2-norm LP (HOLP), high-order sparse LP (HOSpLP) and the 4th order smooth spectral envelope (ENV).

High-Order Sparse Predictor



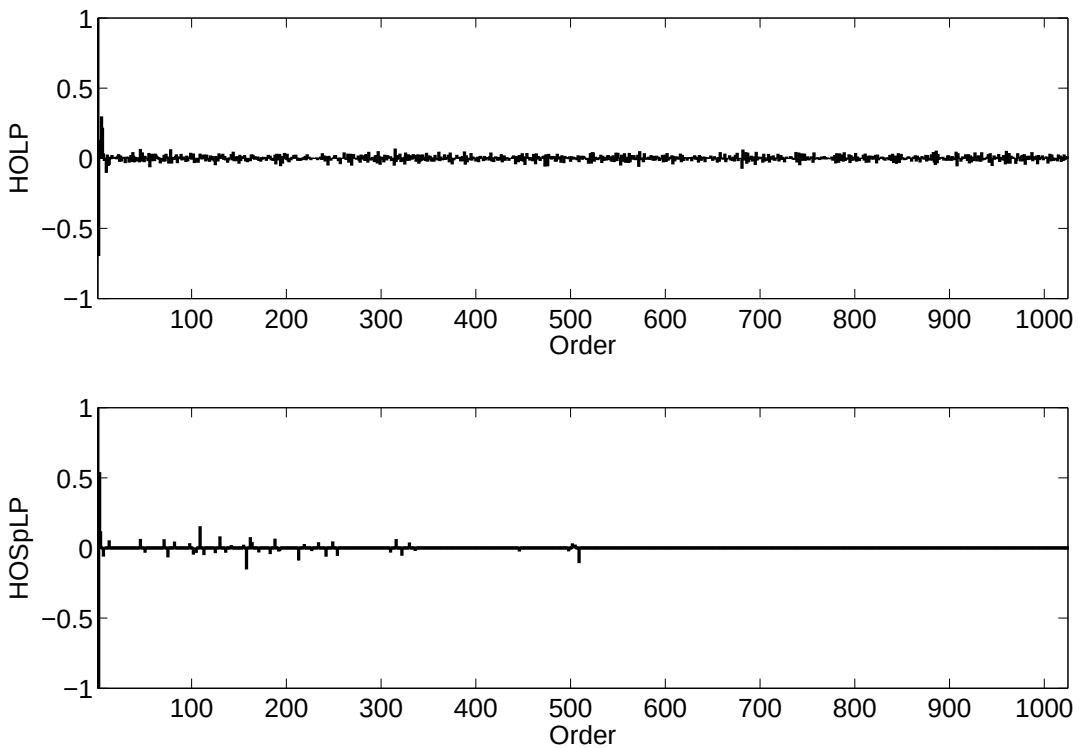
High-order 2-norm LP (HOLP, above) and high-order sparse LP (HOSpLP, below) for a monophonic audio signal. The number of nonzero samples in the sparse predictor is 25. The coefficients with highest magnitude are located around $P = \lceil f_s/f_0 \rfloor = 113$ *(where* $f_s = 387.6$ *Hz). Harmonic components modeled similarly to the fractional delay interpolation filter.*

Spectral Modeling



Polyphonic audio signal. Frequency response for the all-pole high-order 2-norm LP (HOLP), high-order sparse LP (HOSpLP) and the 12th order smooth spectral envelope (ENV). A detail of the first four harmonics (each belonging to a different signal) and the predictors behavior is shown in the smaller frame.

High-Order Sparse Predictor



High-order 2-norm LP (HOLP, above) and high-order sparse LP (HOSpLP, below) for polyphonic audio signal. The number of nonzero samples in the sparse predictor is 53. The predictor is less sparse than in the monophonic case, taking into consideration the different multipitch components.

Spectral Flatness Performances

Difference in spectral flatness between the original audio signals and their LP representations. The high-order sparse LP offers almost the same performance as the high-order 2-norm with significantly less prediction coefficients.

METHOD	ΔSFM_{mono}	ΔSFM_{poly}
HOLP	35.41 dB	37.02 dB
PP	24.37 dB	17.03 dB
HOSpLP	34.59 dB	32.43 dB

Conclusions

- The different components (envelope+harmonics) of the audio signal are modeled efficiently by the high-order predictor.
- While reaching spectral flattening performances comparable with HOLP, the HOSpLP only requires only few nonzero coefficients.
- HOLP does not rely on harmonicity, PP relies only on harmonicity (unsuited for multipitch audio).
- HOSpLP positions itself somewhere in between these two approaches offering a more sophisticated parametric representation.
- Exploiting sparsity, HOSpLP seems to better take into account the different harmonic components of the signal.