Katholieke Universiteit Leuven

Departement Elektrotechniek

ESAT-SISTA/TR 10-66

High-order sparse linear predictors for audio processing[1]

Daniele Giacobello[2 3 4], Toon van Waterschoot[4],
Mads Græsbøll Christensen[5], Søren Holdt Jensen[2], and Marc Moonen[4]

August 2010
Published in *Proc. 18th European Signal Process. Conf. (EUSIPCO '10)*,
Aalborg, Denmark, Aug. 2010, pp. 234–238.

[1]This report is available by anonymous from *ftp.esat.kuleuven.be* in the directory *pub/sista/vanwaterschoot/reports/10-66.pdf*

[2]Dept. of Electronic Systems, Aalborg Universitet, Niels Jernes Vej 12, DK-9220 Aalborg, Denmark, Tel. +45 96 358650, Fax. +45 98 151583, WWW: *http://kom.aau.dk/*. E-mail: *dg@es.aau.dk, shj@kom.aau.dk*.

[3]The work of Daniele Giacobello is supported by the Marie Curie ESTSIGNAL Fellowship (http://est-signal.i3s.unice.fr), contract no. MEST-CT- 2005-021175. The work of Toon van Waterschoot and Marc Moonen is supported by K.U.Leuven Research Council CoE EF/05/006 ("Optimization in Engineering, OPTEC"), the Belgian Programme on Interuniversity Attraction Poles initiated by the Belgian Federal Science Policy Office IUAP P6/04 (DYSCO, "Dynamical systems, control and optimization", 2007-2011), and the Concerted Research Action GOA-MaNet.

[4]K.U.Leuven, Dept. of Electrical Engineering (ESAT), Research group SCD(SISTA), Kasteelpark Arenberg 10, 3001 Leuven, Belgium, Tel. +32 16 321927, Fax +32 16 321970, WWW: *http://homes.esat.kuleuven.be/∼tvanwate*. E-mail: *toon.vanwaterschoot@esat.kuleuven.be, marc.moonen@esat.kuleuven.be*.

[5]Dept. of Architecture, Design & Media Technology, Aalborg Universitet, Niels Jernes Vej 14, DK-9220 Aalborg, Denmark, Tel. +45 99 409793, Fax. +45 99 409788, WWW: *http://imi.aau.dk/ mgc/*. E-mail: *mgc@imi.aau.dk*

HIGH-ORDER SPARSE LINEAR PREDICTORS FOR AUDIO PROCESSING

Daniele Giacobello[1,2], *Toon van Waterschoot*[2], *Mads Græsbøll Christensen*[3], *Søren Holdt Jensen*[1], *Marc Moonen*[2]

[1]Dept. of Electronic Systems, Aalborg Universitet, Denmark
[2]Dept. of Electrical Engineering (ESAT-SCD), Katholieke Universiteit Leuven, Belgium
[3]Dept. of Media Technology, Aalborg Universitet, Denmark
{dg,shj}@es.aau.dk, {tvanwate,moonen}@esat.kuleuven.be, mgc@imi.aau.dk

ABSTRACT

Linear prediction has generally failed to make a breakthrough in audio processing, as it has done in speech processing. This is mostly due to its poor modeling performance, since an audio signal is usually an ensemble of different sources. Nevertheless, linear prediction comes with a whole set of interesting features that make the idea of using it in audio processing not far fetched, e.g., the strong ability of modeling the spectral peaks that play a dominant role in perception. In this paper, we provide some preliminary conjectures and experiments on the use of high-order sparse linear predictors in audio processing. These predictors, successfully implemented in modeling the short-term and long-term redundancies present in speech signals, will be used to model tonal audio signals, both monophonic and polyphonic. We will show how the sparse predictors are able to model efficiently the different components of the spectrum of an audio signal, i.e., its tonal behavior and the spectral envelope characteristic.

1. INTRODUCTION

Linear prediction (LP) is arguably one of the most successful tools for the analysis and coding of speech signals [1]. Its success can be explained by the correspondence between the modeling of the speech production process and the LP analysis. In particular, the all-pole model corresponding to the LP filter can be seen as a good approximation of the vocal tract transfer function [2]. Moreover, the use of LP in speech coding techniques guarantees interesting attributes like low delay, scalability and, in general, low complexity. The predictor in this case is used to decorrelate the speech waveform leaving a prediction residual that is easier to encode.

The LP model is definitely less popular in audio processing. The main reason is that the predictor does not necessarily model any physical mechanism that generated the audio signal. The general difficulties in the accurate parametrization of audio signals [3] have led the way to transform-based audio coders that exploit perceptual models of human hearing [4]. Nevertheless, 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 [5]. This and the properties that made LP successful in speech coding (low delay, scalability and low complexity) make the extension of LP to audio coding also appealing. Several examples can be found in literature (see, e.g., [6, 7, 8, 9]). Furthermore, in audio analysis, LP finds also other interesting applications. For example, the whitening properties of the predictor can be used to obtain fast converging acoustic echo and feedback cancelers (see, e.g., [10, 11]).

The work of Daniele Giacobello is supported by the Marie Curie EST-SIGNAL Fellowship (http://est-signal.i3s.unice.fr), contract no. MEST-CT-2005-021175.

The work of Toon van Waterschoot and Marc Moonen is supported by K.U.Leuven Research Council CoE EF/05/006 (“Optimization in Engineering, OPTEC”), the Belgian Programme on Interuniversity Attraction Poles initiated by the Belgian Federal Science Policy Office IUAP P6/04 (DYSCO, “Dynamical systems, control and optimization”, 2007-2011), and the Concerted Research Action GOA-MaNet.

Since conventional LP, based on the 2-norm minimization of the prediction error, is generally performing poorly in audio processing, several methods have been introduced to improve the LP step in audio processing (see [12] for an overview). High-order autoregressive (AR) models seem to yield some of the highest scores in spectral flatness[1], therefore the predictor retains a great deal of spectral information but it does not provide any useful information for coding purposes.

In our recent work, we have introduced several new predictors for speech processing applications [13]. In particular in [14], we have shown the benefits of using high-order sparse linear predictors to model the cascade of short-term and long-term predictors, providing an efficient decoupling between the two contributions. In general, for a high-order AR filter, a sparse structure is an indication that the polynomial can be factored into several terms. The challenge would now be to extend these early contributions to the case of audio signals. We will test our algorithms and see how the high-order sparse predictors with few nonzero coefficients are capable to model efficiently the tonal behavior of the audio signal as well as the spectral envelope characteristic.

The paper is organized as follows. In Section 2, we introduce the tonal audio signals used in the following sections, providing ideas on how high-order predictors with a sparse structure can model the different components of the audio signal. In Section 3, we illustrate the LP methods used in our experiments and in Section 4 we provide the experimental results. Finally, Section 5 concludes the paper.

2. TONAL AUDIO SIGNAL MODEL

We will only consider tonal audio signals, that is, signals having a spectrum containing a finite number of dominant frequency components at multiples of the fundamental frequency f_0 (usually found in the range 100-1000 Hz). This model covers the majority of audio signals. The performance of the different LP models will be evaluated for three types of audio signals. We will consider true monophonic and true polyphonic audio signals and synthetic audio signals consisting of a sum of harmonic sinusoids.

2.1 Monophonic audio signals

In the monophonic signal model, it is assumed that all tonal components are harmonically related to a single fundamental frequency:

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

where the time index n has been normalized with respect to the sampling period $T_s = 1/f_s$ and $\omega_0 = 2\pi f_0/f_s$. The signal is modeled with M sinusoids (with parameters α_m, $m\omega_0$, ϕ_m) and a noise term $r(n)$ that contains the nontonal components.

[1]The 2-norm minimization of the prediction error is equal, according to the Parseval’s theorem [1], to maximizing the spectral flatness of the residual.

© EURASIP, 2010 ISSN 2076-1465

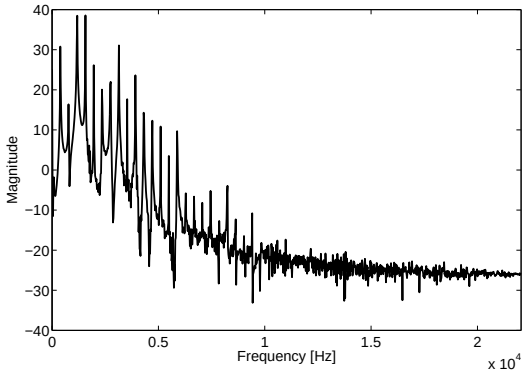

Figure 1: *Magnitude spectrum of the monophonic audio signal of Example 1.*

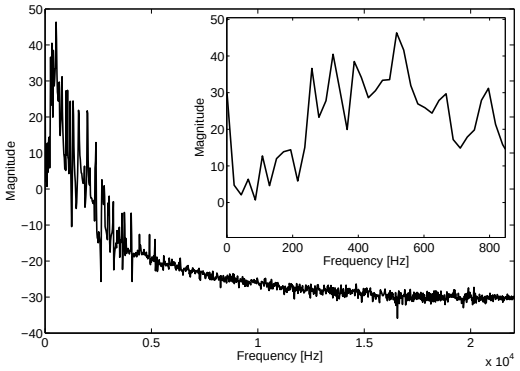

Figure 2: *Magnitude spectrum of the polyphonic audio signal of Example 3. In the smaller frame, we show a detail of the frequency range* $[0,800]$ *Hz where the first harmonics of each of the four monophonic signals are located (*$f_{0,n} = \{258.4, 323.0, 387.6, 516.8\}$*).*

Example 1. The monophonic audio fragment considered was extracted from a Bb clarinet sound recording in the McGill University Master Samples (MUMS) collection ($f_s = 44100$ *Hz). The spectrum of this* $N = 2048$ *samples fragment, which corresponds to the samples 70001 to 72048 of the G4 note recording, is shown in Figure 1. The fundamental frequency corresponds to* $f_0 = 387.6$ *Hz and the signal has* $M = 15$ *relevant harmonics.*

Even though this signal can generally not be considered as output of an AR process, significant considerations can be made. As it is clear from Figure 1, the signal spectrum is made up by two components: a comb-like structure where the peaks are located in the multiples of the fundamental frequency and a smooth spectral envelope that resembles a low-pass filter, since the harmonic structure is more prominent in the lower half of the spectrum. The comb-like structure can be modeled by the filter:

$$H_p(z) = \frac{1}{P(z)} = \frac{G_p}{1 - pz^{-P}}, \tag{2}$$

where $P = T_0/T_s$ ($T_0 = 1/f_0$) and G_p is a scaling factor[2]. The low-pass component can be modeled by an all-pole filter:

$$H_f(z) = \frac{1}{F(z)} = \frac{G_f}{1 - \sum_{k=1}^{N_f} f_k z^{-k}}. \tag{3}$$

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

$$\begin{aligned} H_a(z) &= \frac{1}{A(z)} = \frac{G_f G_p}{F(z)P(z)} = \frac{G_f G_p}{1 - \sum_{k=1}^{K} a_k z^{-k}} \\ &= \frac{G_f G_p}{(1 - \sum_{k=1}^{N_f} f_k z^{-k})(1 - pz^{-P})}. \end{aligned} \tag{4}$$

The signal can therefore be modeled with an order $K \geq P + N_f$ sparse predictor $A(z)$. The resulting predictor coefficient vector $\mathbf{a} = \{a_k\}$ of the high-order polynomial $A(z)$ will therefore be highly sparse. We will see how we can take this into account in the linear prediction model and minimization criterion.

2.2 Synthetic audio signals consisting of a sum of harmonic sinusoids in white noise

Synthetic tonal audio signals are well suited for examining the modeling properties of the high-order sparse LP models presented below, since these provide exact knowledge of the fundamental frequency f_0 and the number of harmonics. The model is similar to (1):

$$x(n) = \sum_{m=1}^{M} \alpha_m \cos(m\omega_0 n + \phi_m) + r(n), \quad n = 1, \ldots, L, \tag{5}$$

except that the noise term $r(n)$ will be white noise, therefore not containing low-power harmonics.

Example 2. We have built a synthetic signal of $N = 2048$ *samples with* $M = 15$ *tonal components and random, uniformly distributed amplitudes (*$\alpha_m \in (0,1]$*) and phases (*$\phi_m \in [0, 2\pi)$*). The radial fundamental frequency was chosen to be* $\omega_0 = 2\pi/64$*, that is, at* $f_s = 44.1$ *kHz,* $f_0 = 689.1$ *Hz. The pitch period* T_0 *being equal to an integer number of sampling periods (*$T_0 = 64T_s$*) will clearly illustrate the effects of the pitch predictor.*

In this case, we can also make considerations similar to those made for the monophonic case. The magnitude spectrum is similar to the one in Figure 1, the main difference being the predominance of the harmonic sinusoids over the rest of the spectrum. While the comb-like behavior can still be modeled by a pitch predictor $P(z)$, the predictor $F(z)$, used to model the smooth spectral envelope of the signal, will now serve to enhance the frequencies where the harmonics are located. In particular, the low-pass filter will exhibit a sharper transition between the lower half of the spectrum and the higher frequencies. This necessarily translates into a higher order N_f for $F(z)$.

2.3 Polyphonic audio signals

The polyphonic audio signals are a finite sum of monophonic signals:

$$\begin{aligned} x(n) = \sum_{m=1}^{M} \Big(\sum_{q=1}^{Q_m} \alpha_{m,q} \cos(q\omega_{0,m} n + \phi_{m,q}) \Big) + r(n), \\ n = 1, \ldots, L \end{aligned} \tag{6}$$

where $\omega_{0,q}$ represents the fundamental frequency of the $q-$th monophonic signal.

Example 3. The polyphonic audio signal considered was generated by adding four monophonic piano sounds from the MUMS concert hall Steinway recordings. The samples 2001 to 4048 of the C4, E4, G4, and C5 note recordings were added to obtain a $N = 2048$ *C major chord, plotted in Figures 2. The four fundamental frequencies are* $f_{0,n} = \{258.4, 323.0, 387.6, 516.8\}$ *Hz, and each of the monophonic components has 7 relevant harmonics.*

Linear prediction of polyphonic audio signals is the most challenging case. It is also the most significant one, since audio signals are usually an ensemble of different sources with different fundamental

[2] If P is non-integer, a fractional-delay filter $P(z)$ can be used [15].

frequencies. The same reasoning we have followed for the case of monophonic audio signals can be used for polyphonic signals with some important differences. The smooth spectral envelope is clearly similar to the monophonic one, therefore requiring a low-order predictor $F(z)$ to model it. The substantial difference comes from the modeling of the sum of the different comb-like components. In particular, the multipitch structure, differently from (2), will have to be modeled by:

$$H_p(z) = \sum_{i=1}^{M} \frac{G_{p_i}}{P_i(z)} = \sum_{i=1}^{M} \frac{G_{p_i}}{1 - p_i z^{-P_i}}, \tag{7}$$

which is a pole-zero filter. Since we are interested in an all-pole filter this may translate into a defect in modeling. Nevertheless, in our experimental analysis, we have noticed that, since $p_i < 1$, we can write:

$$H_p(z) = \sum_{i=1}^{M} \frac{G_{p_i}}{1 - p_i z^{-P_i}} \approx \frac{G_p}{\prod_{i=1}^{M}(1 - p_i z^{-P_i})}. \tag{8}$$

This simplification seems far fetched and obviously requires some further analysis. Nevertheless, we will show it holds quite well in modeling the harmonic behavior. Just as in the monophonic case, also a low-order all-pole model (3) can be used to model the envelope. The high-order sparse predictor resulting from the cascade of the two contributions will still be sparse:

$$\begin{aligned} H_a(z) &\approx \frac{1}{A(z)} = \frac{G_f G_p}{F(z)P(z)} = \frac{G_f G_p}{1 - \sum_{k=1}^{K} a_k z^{-k}} \\ &= \frac{G_f G_p}{(1 - \sum_{k=1}^{N_f} f_k z^{-k})(\prod_{i=1}^{M}(1 - p_i z^{-P_i}))}. \end{aligned} \tag{9}$$

The order of the high-order sparse predictor $A(z)$ will be $K \geq \sum_i P_i + N_f$ in order to accommodate all the cross terms.

3. LINEAR PREDICTION IN AUDIO PROCESSING

The estimation problems considered in this paper are based on the following autoregressive (AR) model, where a signal sample $x(n)$ is written as a linear combination of past samples:

$$x(n) = \sum_{k=1}^{K} a_k x(n-k) + e(n). \tag{10}$$

Here, $\{a_k\}$ are the prediction coefficients and $e(n)$ is the excitation of the corresponding AR filter, also referred to as the prediction error. We consider 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 prediction error is minimized [16]. This corresponds to the following minimization problem:

$$\min_{\mathbf{a}} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_p^p + \gamma \|\mathbf{a}\|_k^k, \tag{11}$$

where

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

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 by assuming $x(n) = 0$ for $n < 1$ and $n > N$. In this paper we will use the most common choice of $N_1 = 1$ and $N_2 = N + K$, which is equivalent, when $p = 2$ and $\gamma = 0$, to the *autocorrelation method*:

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_2^2 = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{x}, \tag{12}$$

where $\mathbf{R} = \mathbf{X}^T\mathbf{X}$ is the autocorrelation matrix (when $N_1 = 1$ and $N_2 = N + K$) [17].

3.1 High-order LP modeling

It is well known that a signal composed of M sinusoids can be modeled exactly using an autoregressive moving average model, i.e., ARMA($2M$,$2M$) model. This model can be arbitrarily closely approximated with an AR model, provided that the model order K is chosen large enough [18]. We will consider for all our audio segments a $K = 1024$ order predictor, solution of the 2-norm minimization problem (12). The general goal of the high-order model is to maximize the spectral flatness of the residual. However, the all-pole model does not provide hints for factorization, as it does not exploits the harmonicity properties of the signal.

3.2 Pitch prediction

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 in Eq. (2). Obviously, the pitch period will generally not be an integer multiple of the sampling period, such that the use of a multi-tap pitch predictor is required for interpolation, or a fractional-delay filter should be used. The drawback with employing only a pitch predictor is that this creates an extremely non-smooth residual signal by also attempting to cancel harmonic frequencies which are not present in the input signal. For these reasons, in this paper we will use a 3-tap pitch predictor [19], efficient in modeling the decreasing comb-like structure of the signals analyzed.

The pitch prediction model is the only prediction model in which the harmonicity property is exploited. The underlying signal model of the monophonic audio signal in (1) is harmonic, while the polyphonic signal model in (6) is not. Therefore, while performing accurately for the monophonic signal, the pitch predictor fails to recover the different pitch components in the polyphonic audio. In particular, we have observed, that its estimation of the fundamental frequency $f_0 = 1/T_0$ is similar to a weighted average of the different fundamental frequencies $f_{0,n}$ of the underlying model.

3.3 High-order sparse LP modeling

Considering the two signal models we have introduced for the monophonic and synthetic audio (4) and for the polyphonic audio (9), we use the minimization problem in (11) to find the LP coefficients imposing $k = 0$. In this way, sparsity of the high-order predictor is taken into consideration directly in the minimization problem. The operator $\|\cdot\|_0$ represents the so-called 0-norm, i.e., the cardinality of the vector. A relaxation of this non-convex problem is obtained by approximating the 0-norm with the more tractable 1-norm or by the iteratively reweighted 1-norm, bringing the solution closer to the 0-norm [13]. In this paper we will limit the analysis to the 1-norm. The regularization term γ is then clearly related to the *a priori* knowledge that we have on the coefficients vector $\{a_k\}$ or, in other terms, to how sparse $\{a_k\}$ is. There are many ways to choose γ. To generate preliminary results, we will consider it fixed ($\gamma = 0.1$). The order of the predictor is $K = 1024$. The choice of p is also non-trivial. For $p = 2$ we will obtain a Gaussian residual, consistent with the equivalent i.i.d. Gaussian maximum likelihood approach to determine the coefficients. The case $p = 1$ is probably more interesting: seeing this as a convex relaxation of the 0-norm, the residual will be also *sparse*, providing interesting coding properties that will be subject to further analysis. The minimization problem considered used is then:

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_1 + \gamma \|\mathbf{a}\|_1. \tag{13}$$

The high-order LP in (12) does not rely on harmonicity, while the pitch predictor relies basically only on harmonicity thus greatly simplifying the calculations. The high-order sparse LP positions itself somewhere in between these two approaches, providing significant modeling properties similar to (12) but parametrizing the signal in a more sophisticated way by taking into account the different components of the signal. Furthermore, when the order K approaches $N/2$ in (12), a number of spurious spectral peaks start

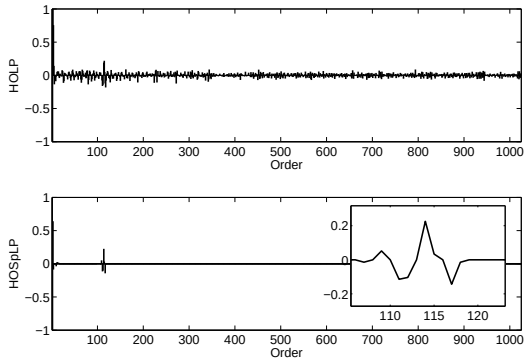

Figure 3: *High-order 2-norm LP (HOLP, above) and high-order sparse LP (HOSpLP, below) for a monophonic audio signal. A detail of the coefficients of order 105-125 is shown in the frame. The number of nonzero samples in the sparse predictor is 25.*

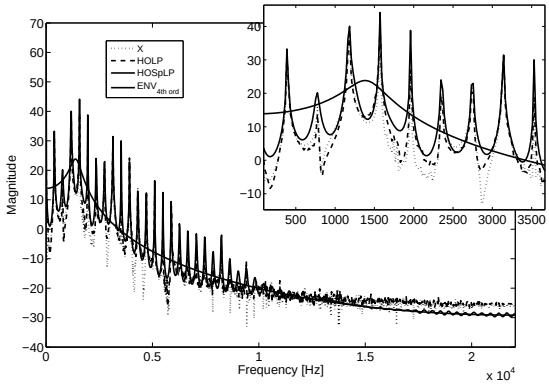

Figure 4: *Monophonic audio signal. Frequency response for the all-pole high-order 2-norm LP (HOLP), high-order sparse LP (HOSpLP) and the 4th order smooth spectral envelope (ENV). A detail of the first nine harmonics and the predictors modeling behavior is shown in the smaller frame.*

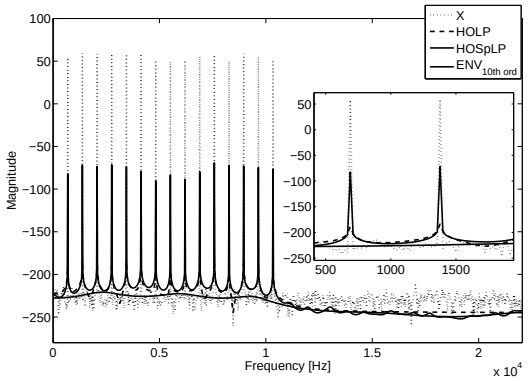

Figure 5: *Synthetic sum of sinusoids. Frequency response for the all-pole high-order 2-norm LP (HOLP), high-order sparse LP (HOSpLP) and the 10th order smooth spectral envelope (ENV). A detail of the first two harmonics and the predictors behavior is shown in the smaller frame.*

to appear. This effects can be traced back to the ill-conditioning of the normal equations $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{x}$ and in particular to the observation matrix $\mathbf{X}$ with highly correlated rows when sinusoids are present in the data [20]. The sparsity of the predictor, helps reducing the ill-conditioning basically applying an "automatic" pruning of the rows of the observation matrix without the necessary *a priori* knowledge used, for example, in [21]. Indeed, the inclusion of the regularization term in (13) can also be seen as a general method for solving ill-posed problems [22].

4. EXPERIMENTAL ANALYSIS

4.1 Spectral modeling

In this section we will compare the use of high-order sparse LP with the conventional high-order 2-norm LP. The comparison is done for the audio signals introduced in Section 1 (Example 1-3). The first N_f coefficients belonging to the low-pass filter are chosen using a model order selection criterion [13].

4.1.1 Monophonic audio signal

The frequency response of the filters is shown in Figure 4 while the two predictors are shown in Figure 3. It is clear that the predictor is an accurate model of the two expected contributions: $P(z)$ and $F(z)$. In particular the convolutive term is clustered around the integer pitch delay corresponding to the inverse of the fundamental frequency and the peak is exactly located in $P = \lceil f_s/f_0 \rfloor = 113$ (where $f_s = 387.6$ Hz). Remarkably, the shape resembles the fractional-delay interpolation filter [23]. The combination of the two contributions models very accurately the comb-like structure and the low-pass behavior (Fig. 4). A 4th order polynomial was enough to model the low-pass behavior, this corresponds to the first four samples of the sparse prediction vector. It is also clear that the order $K = 1024$ is excessive, an order $K \geq P + N_f$ where $N_f \approx 4$ and $P = f_s/f_0$ would have been sufficient. A final word should be spent regarding the sparsity of the vector. The signal, having $M = 15$ relevant harmonics, could be modeled accurately using an ARMA(30,30) model. It is clear that achieving similar performance with just 25 nonzero samples is an important result that can be exploited in coding applications.

4.1.2 Synthetic sum of sinusoids

Similar considerations can be made for the synthetic audio signal. A 10th order polynomial models the envelope enhancing the frequency present in the first half of the spectrum. The pitch predictor models *exactly* the comb like structure since the pitch period T_0 is equal to an integer number of sampling periods ($T_0 = 64T_s$). An example of the modeling behavior of the predictor is shown in Figure 5. For the sake of brevity the predictor structure is not shown.

4.1.3 Polyphonic audio signal

The frequency response of the filters is shown in Figure 6 while the two predictors are shown in Figure 7. The predictor is less sparse than in the monophonic case, taking into consideration the different multipitch components. Furthermore, we notice that the approximation we have performed in (9), holds quite well and the predictor seems to model accurately the whole sum of different harmonics coming from the different signals. The only drawback seems the over-emphasis of the envelope in modeling the low-pass behavior that we have not observed in the other cases. This will be subject to further analysis since at this early point it is difficult to provide an explanation. In this case also the order $K = 1024$ is excessive: recalling that $K \geq \sum_i P_i + N_f$, the order should be a little higher than 500. Moreover, the number of nonzero samples in the sparse predictor is 53, which is considerably less than the number of coefficients of an ARMA(56,56) model (sum of four signal with $M = 7$ relevant harmonics each).

4.2 Spectral Flatness Performance

The spectral flatness measure (SFM) of the LP residual [18] in dB is a negative real number, with SFM= 0 dB corresponding to a flat spectrum. In Table 1 we describe the ΔSFM's, differences in spectral flatness, between the original audio signals (monophonic and polyphonic) and its residual provided by the three methods presented in Section 3. It can clearly be seen that high-order 2-norm minimization certainly provides a higher spectral flatness (as expected) although with a highly dense predictor. The 3-tap pitch-

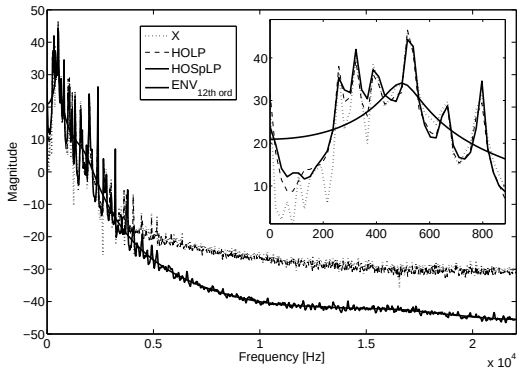

Figure 6: *Polyphonic audio signal. Frequency response for the all-pole high-order 2-norm LP (HOLP), high-order sparse LP (HOSpLP) and the 12th order smooth spectral envelope (ENV). A detail of the first four harmonics (each belonging to a different signal) and the predictors behavior is shown in the smaller frame.*

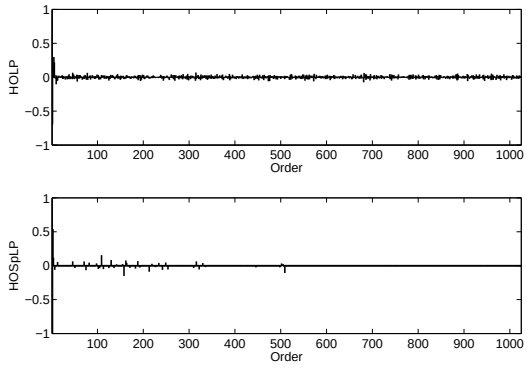

Figure 7: *High-order 2-norm LP (HOLP, above) and high-order sparse LP (HOSpLP, below) for polyphonic audio signal. The number of nonzero samples in the sparse predictor is 53.*

predictor, while performing with a certain degree of accuracy in the monophonic case, fails to model the multipitch behavior of the underlying signal structure in the polyphonic case. The high-order sparse LP offers almost the same performance as the high-order 2-norm with only $1/100$th of the taps necessary. As for the polyphonic case, we notice a more significant difference in performance between sparse LP and 2-norm LP. This is mostly due to the simplification of the pole-zero model structure represented only by the sparse LP and the over-emphasis of the low-pass spectral characteristic in the higher frequency range.

5. CONCLUSIONS

The use of high-order sparse LP in audio processing seems quite promising. In particular, the different components of the audio signal (the spiky harmonics located on the lower half of the spectrum and the low-pass overall behavior of the envelope) are modeled efficiently by the high-order predictor. Furthermore, while reaching spectral flattening performances comparable with high-order 2-norm LP, the high-order sparse LP only requires few nonzero components, offering important hints for coding. In this regard, we should notice that the use of 1-norm residual minimization provides also a *sparse* residual rather than a minimum variance one, arguably related to more efficient coding strategies. Although the frequency behavior corresponding to the 1-norm minimization is unknown, the numerical results obtained clearly show potential advantages of the sparse formulation for spectral modeling. The results presented also make the sparse LP modeling promising for coding applications. This, and other questions left open, such as stability and complexity will be subject of our future work.

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

Table 1: *Difference in spectral flatness between the original audio signals (monophonic and polyphonic) and their residuals for the three methods presented in Section 3: high-order 2-norm LP (HOLP), 3-tap pitch predictor (PP) and high-order sparse LP (HOSpLP).*

REFERENCES

[1] J. Makhoul, "Linear prediction: a tutorial review", *Proc. IEEE*, vol. 63(4), pp. 561–580, 1975.

[2] J. H. L. Hansen, J. G. Proakis, and J. R. Deller, Jr., *Discrete-time processing of speech signals*, Prentice-Hall, 1987.

[3] M. G. Christensen and A. Jakobsson, *Multi-pitch estimation*, Synthesis Lectures on Speech and Audio Processing, Morgan & Claypool.

[4] K. Brandenburg and G. Stoll, "The ISO-MPEG-1 audio: A generic standard for coding of high-quality digital audio," *Journal of the Audio Engineering Society*, vol. 42, no. 10, pp. 780–792, 1994.

[5] M. R. Schroeder, "Linear prediction, extremal entropy and prior information in speech signal analysis and synthesis," *Speech Communication*, vol. 1, no. 1, pp. 9–20, 1982.

[6] G. Schuller and A. Härmä, "Low delay audio compression using predictive coding," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 2 pp. 1853-1856, 2002.

[7] F. Riera-Palou, A. C. den Brinker, and A. J. Gerrits, "A hybrid parametric-waveform approach to bistream scalable audio coding," in *Rec. Asilomar Conf. Signals, Systems, and Computers*, pp. 2250-2254, 2004.

[8] A. A. Biswas, *Advances in perceptual stereo audio coding using linear prediction techniques*, Ph.D. Thesis, Technische Universiteit Eindhoven, 2007.

[9] A. Härmä and U. K. Laine, "A comparison of warped and conventional linear predictive coding," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 579-588, 2001.

[10] T. van Waterschoot, G. Rombouts, P. Verhoeve, and M. Moonen, "Double-talk-robust prediction error identification algorithms for acoustic echo cancellation," *IEEE Trans. Signal Process.*, vol. 55, no. 3, pp. 846-858, 2007.

[11] G. Rombouts, T. van Waterschoot, K. Struyve, and M. Moonen, "Acoustic feedback suppression for long acoustic paths using a nonstationary source model," *IEEE Trans. Signal Process.*, vol. 54, no. 9, pp. 3426-3434, 2006.

[12] T. van Waterschoot and M. Moonen, "Comparison of linear prediction models for audio signals," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2008, Article ID 706935, 24 pages, 2008.

[13] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen, "Sparse Linear Prediction and Its Applications to Speech Processing," submitted to *IEEE Transactions in Audio, Speech and Language Processing*, January 2010. Available at: `http://kom.aau.dk/~dg/publications.html`.

[14] D. Giacobello, M. G. Christensen, J. Dahl, S. H. Jensen, and M. Moonen, "Joint estimation of short-term and long-term predictors in speech coders," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 4109–4112, 2009.

[15] P. Kroon and B. S. Atal, "Pitch predictors with high temporal resolution," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, pp. 661-664, 1990.

[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. M. Kay, "The effects of noise on the autoregressive spectral estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 5, pp. 478-485, 1979.

[19] Y. Qian, G. Chahine, and P. Kabal, "Pseudo-multi-tap pitch filters in a low bit-rate CELP speech coder," *Speech Communication*, vol. 14, no. 4, pp. 339-358, 1994.

[20] D. Tufts and R. Kumaresan, "Singular value decomposition and improved frequency estimation using linear prediction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 30, no. 4, pp. 671–675, 1982.

[21] R. Kumaresan, "Accurate frequency estimation using an all-pole filter with mostly zero coefficients," *Proc. IEEE*, vol. 70, no. 8, pp. 873–875, 1982.

[22] P. C. Hansen and D. P. O'Leary, "The use of the L-curve in the regularization of discrete ill-posed problems," *SIAM Journal on Scientific Computing*, vol. 14, no. 6, pp. 1487–1503, 1993.

[23] T. I. Laakso, V. Välimäki, M. Karjalainen, and U. K. Laine, "Splitting the unit delay [FIR/all pass filters design]," *IEEE Signal Processing Magazine*, vol. 13, no. 1, pp. 30-60, 1996.