Sparse Linear Predictors for Speech Processing

Giacobello, Daniele; Christensen, Mads Græsbøll; Dahl, Joachim; Jensen, Søren Holdt; Moonen, Marc

Published in:
Proceedings of the International Conference on Spoken Language Processing

Publication date:
2008

Document Version
Publisher's PDF, also known as Version of record

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
{dg,mgc,joachim,shj}@es.aau.dk, marc.moonen@esat.kuleuven.be

Abstract

This paper presents two new classes of linear prediction schemes. The first one is based on the concept of creating a sparse residual rather than a minimum variance one, which will allow a more efficient quantization; we will show that this works well in presence of voiced speech, where the excitation can be represented by an impulse train, and creates a sparser residual in the case of unvoiced speech. The second class aims at finding sparse prediction coefficients; interesting results can be seen applying it to the joint estimation of long-term and short-term predictors. The proposed estimators are all solutions to convex optimization problems, which can be solved efficiently and reliably using, e.g., interior-point methods.

Index Terms: linear prediction, all-pole modeling, convex optimization

1. Introduction

Linear prediction (LP) is an integral part of many modern speech and audio processing systems ranging from diverse applications such as coding, analysis, synthesis and recognition [1]. Typically, the prediction coefficients are found such that the 2-norm of the residual (the difference between the observed signal and the predicted signal) is minimized [2]. The reason behind this work is that there are many examples where this does not work well, for example when the excitation is not Gaussian, which is the case for voiced speech. In this case the usual approach is to find coefficients for the short-term and long-term signal correlation in two different steps [3]. This obviously leads to inherently suboptimal solutions. In the context of predictive coding, moreover, alternative formulations may be of interest. 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 [4] to quantize the residual. In these techniques, notably the Multi-Pulse and Regular-Pulse Excitation methods (MPE and RPE) [5, 6], 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.

In this paper, we present a framework wherein two kinds of sparse linear predictors are considered corresponding to two different ways of estimating the prediction coefficients. First, we consider the case where the excitation signals are assumed to be sparse, as in the case of voiced speech. Then, we consider the case where, not the residual, but the prediction coefficients are sparse. This latter case allows us to jointly estimate the short-term and long-term predictor coefficients and may be applied in speech coders. Therefore, the novelty introduced is to exploit the statistical characteristics of the algorithms introduced for linear prediction in order to define, in the latter stage, a more efficient quantization scheme.

The paper is organized as follow. A prologue that defines the mathematical formulations of the proposed algorithms will be given. The core will be dedicated to introducing the two algorithms and showing the results obtained with these techniques and some related examples. Then we will discuss and illustrate advantages and drawbacks of them.

2. Fundamentals

The problems considered in this paper are based on the following auto-regressive model, where a sample of speech is written as a linear combination of past samples:

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

where $\{a_k\}$ are the prediction coefficients and $e(n)$ is the excitation. The different predictors considered we will see that apply to different kinds of excitation $e(n)$ and different applications. Mathematically we can state the class of problems considered in this paper 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 $n = 1, \ldots, N$ so that the error is minimized [7]. The vector $\hat{\mathbf{e}} = \mathbf{x} - \mathbf{X}\hat{\mathbf{a}}$ is commonly referred to as the residual which is an estimate of the excitation $\mathbf{e}$, obtained from some estimate $\hat{\mathbf{a}}$ resulting from the following minimization problem:

$$\min_{\mathbf{a}} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_p^p + \gamma \|\mathbf{a}\|_k^k, \tag{2}$$

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 assuming that $x(n) = 0$ for $n < 1$ and $n > N$. For example, considering $p = 2$ and $\gamma = 0$ (maximum likelihood approach for the error being a sequence of i.i.d. Gaussian random variable), setting $N_1 = 1$ and $N_2 = N + K$ will lead us to the autocorrelation method equivalent to solving the Yule-Walker equations; setting $N_1 = K + 1$ and

$N_2 = N$ leads us to the covariance method [8]. We will show that the choice of N_1 and N_2 is not trivial even in the case when $p \neq 2$ where the system in (2) has not a closed-form unique solution.

The question then is how to choose p, k and γ and how to perform the associated minimization, depending on the kind of applications we want to implement. In finding sparse signal representation, there is the somewhat subtle problem of how to measure sparseness. Sparseness is often measured as the cardinality, that would be the so-called 0-norm $\|\cdot\|_0$ [9], therefore, using it in (2) means that we would like to minimize the number of non-zero samples in the error signal. Unfortunately this is a combinatorial problem which generally cannot be solved in polynomial time. Instead of the cardinality measure, we then use the more tractable 1-norm $\|\cdot\|_1$.

The introduction of the regularization term γ in (2) can have two meanings. The first one, where γ is somehow related to the prior knowledge we have of the coefficients vector $\mathbf{a}$, therefore (2) is clearly the *maximum a posteriori* (MAP) approach for finding $\mathbf{a}$ under the assumptions that $\mathbf{a}$ has a Generalized Gaussian Distribution [10]:

$$\begin{aligned}\mathbf{a_{MAP}} &= \arg\max_{\mathbf{a}} f(\mathbf{x}|\mathbf{a})g(\mathbf{a}) \\ &= \arg\max_{\mathbf{a}} \{\exp(-\|\mathbf{x}-\mathbf{X}\mathbf{a}\|_p^p)\exp(-\gamma\|\mathbf{a}\|_k^k)\}.\end{aligned} \tag{3}$$

The second meaning that γ holds can be understood by the following analogy. If in (2) we let $k = 0$ and assume that the number of bits associated with the quantization of the prediction coefficients $\mathbf{a}$ be proportional to the number of non-zero elements in $\mathbf{a}$, then the regularization factor γ plays the role of a Lagrange multiplier in a rate-constrained rate-distortion optimization with p determining the error criterion in question: by adjusting γ, we obtain solutions for $\mathbf{a}$ having different rates.

3. Sparse Linear Predictors

3.1. Finding a Sparse Residual

We now proceed to consider the problem of finding a prediction vector $\mathbf{a}$ such that the residual would be sparse. As we shall see this approach is particularly applicable to analysis and coding of voiced speech. Having defined the 1-norm as an approximation of the cardinality function, the cost function for the problem in question is a special case of (2). By setting $p = 1$ and $\gamma = 0$ we obtain the following optimization problem:

$$\min_{\mathbf{a}} \|\mathbf{x}-\mathbf{X}\mathbf{a}\|_1. \tag{4}$$

The use of a least absolute value error criterion has already been proven to give interesting results in linear prediction of speech signals [11]. Especially 1-norm has been proven to give good results when the error is considered to have long tails, that is due to the fact that when $p = 1$ and $\gamma = 0$, the minimization process corresponds to the maximum likelihood approach when the error sequence is considered to be a set of i.i.d. Laplacian random variables. The excitation in the case of voiced speech is well represented by this statistical approximation, therefore the 1-norm minimization outperforms the 2-norm in finding a more proper linear predictive representation.

It should be noted that standard linear prediction $\|\mathbf{x}-\mathbf{X}\mathbf{a}\|_2$ exhibits spectral matching properties in the frequency domain due to the Parseval's theorem [2]: it is also interesting to note that minimizing the squared error in both time domain and frequency domain leads to the same set of equations, which are the Yule-Walker equations [8]. To our knowledge, the only relations existing between the time and frequency domain error using the 1-norm is the trivial Hausdorff-Young inequality [12]:

$$\sum_{n=-\infty}^{\infty} |e(n)| < \frac{1}{2\pi}\int_{-\pi}^{\pi} |E(e^{j\omega})|d\omega, \tag{5}$$

that explicates the non-correspondence of the frequency domain minimization approach for the 1-norm. It is difficult to say if the 1-norm is always advantageous compared to the 2-norm, since it is not clear the statistical character of the frequency errors. Nevertheless, in our experimental studies, we empirically observed that the use of the 1-norm was helpful against the usual problems that the 2-norm LP analysis has to deal with in the case of voice speech with well-defined harmonics (those would be, for example, over-emphasis on peaks and cancellation of errors [2]).In the case of unvoiced speech, in addition, the residual $e(n)$ has always shown to be sparser than the one obtained with the usual LP analysis.

3.2. Finding Sparse Coefficients

Another intriguing incarnation of the general optimization problem (2) is to minimize the 2-norm of the residual while keeping the coefficient vector $\mathbf{a}$ sparse:

$$\min_{\mathbf{a}} \|\mathbf{x}-\mathbf{X}\mathbf{a}\|_2^2 + \gamma\|\mathbf{a}\|_1. \tag{6}$$

This formulation is relevant because a direct minimization of (2) in the standard LP form ($p = 2, \gamma = 0$) with a high prediction order K, will lead to have a coefficient vector $\mathbf{a}$ containing many non-zero elements even if the true order is less than K. The meaning of looking for a sparse coefficient vector $\mathbf{a}$ can be understood as follows. 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 often used in speech processing is an example. A commonly used long-term predictor is:

$$P(z) = 1 - g_p z^{-T_p}, \tag{7}$$

with T_p being the pitch period (the reciprocal of the fundamental frequency usually found in the range $[50Hz, 500Hz]$) and $g_p > 0$ being the gain. Therefore, the optimization problem in (6) can be interpreted as a joint estimation of the short-term and long-term prediction coefficients, something which is usually achieved in cascade and thus suboptimal way [16, 17]. Also, the proposed approach does not require the pitch period to be known or estimated, unlike some practical long-term predictors. The minimization of the 2-norm in (6) is based on the assumption that aside from the pulse-train, the excitation $e(n)$ also consist of Gaussian noise (as usually represented in the mathematical models of speech production). Regarding the implementation of this algorithm, the optimization problem can be posed as a quadratic programming problem and can also be solved in time equivalent to solving a small number of 2-norm linear prediction problems using an interior-point algorithm [14], as the problem in (4).

4. Numerical Experiments

The results of the approach shown in (4) for a voiced signal exhibit a residual that is surprisingly similar to the impulse response of the long term predictor, an example is presented in Figure 1. It is also easy to see that the 2-norm minimization

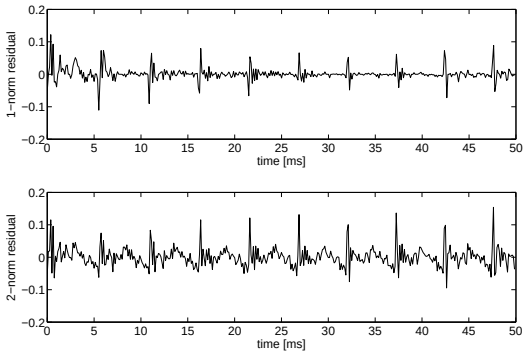

Figure 1: *Residuals for 1-norm and 2-norm minimization.*

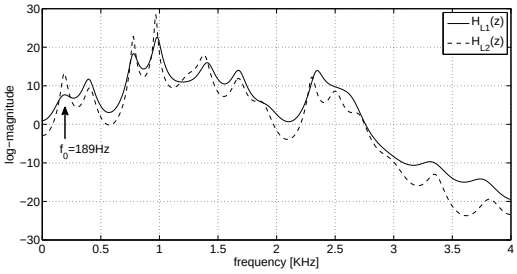

Figure 2: *Frequency response of the filters obtained with 1-norm and 2-norm minimization.*

introduces high emphasis on peaks in its effort to reduce great errors: in this case the outliers due to the pitch excitation, as we can see clearly in Figure 2. Our examples were obtained analyzing the vowel /a/ uttered by a female speaker using $N = 400$, $f_s = 8KHz$ and order $K = 20$. Since the fundamental frequency for the analyzed signal is around $189Hz$, the common LP analysis will try to put a pole very closed to the unit circle around those radians to cancel the harmonic, there explained the peak. The 1-norm approach acknowledges the existence of the pitch harmonic, although it does not try to cancel it because its purpose is not to fit the error into a Gaussian-like probability density function. The result, as clearly shown in Figure 2, is that with the 1-norm minimization we obtain a smoother filter.

In Figure 3 we show an example of the results for our second approach, outlined in section 3.2, on the coefficient vector of the same speech segment analyzed above. The comparison of the prediction coefficients was made between our algorithm for $\gamma = 0.1$ and $\gamma = 1$, with usual LP (order 50) and with the multiplication of the transfer functions of the 10^{th}-order short term predictor (obtained as the mean in the Line Spectral Frequencies domain of four set of LP parameters calculated in the analyzed signal) and the long term predictor obtained by closed loop pitch analysis $P(z) = 1 - 0.22z^{-40}$. In general, we were able to see that using $0.1 \leq \gamma \leq 1$ in (6), the predictive vector $\mathbf{a}$ is similar to the multiplication of the short-term prediction filter $A_{stlp}(z)$ and long-term prediction filter (7) obtained in cascade, in other words in our one step approach we obtained:

$$\frac{1}{A_{sparse}(z)} \simeq \frac{1}{1 - g_p z^{-T_p}} \frac{1}{A_{stlp}(z)}. \qquad (8)$$

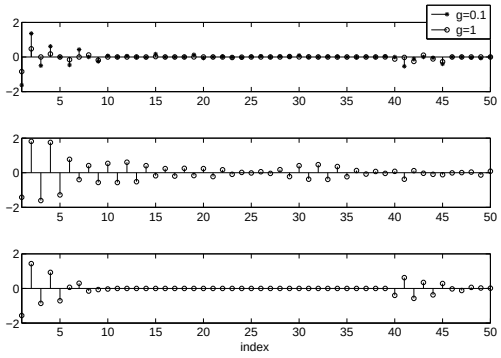

Figure 3: *Comparison of the prediction coefficients (excluding the 0^{th}-order) obtained with our algorithm (top), with usual LP (order 50) and with the convolution of the short-term and long-term coefficients vectors.*

5. Discussion

Denoël and Solvay [11] have pointed out the drawbacks of the absolute error approach that we used in section 3.1. One of them is that the solution (just like the median value of a even number of observations) may not be unique; in this case due to the convexity of the cost function, we can easily state that the all the possible multiple solution would still be the optimal ones [13]; also, seeing the non-uniqueness of the solution as a weakness is arguable: in the set of possible optimal solutions we can probably find a set of coeffients that offer better properties for our purposes.

The stability of this method is not guaranteed, being not intrinsically stable like LP analysis with the autocorrelation method. This drawback was mitigated by choosing $N_1 = 1$ and $N_2 = N + K$ in (2): it also corresponds to the autocorrelation method if the 2-norm was used. This helped us bring the percentage of non-stable filters from 11% (using $N_1 = K + 1$ and $N_2 = N$) to less than 2% in over 10,000 frames analyzed. Although the use of windows to mitigate the spectral peaks or bandwidth expansion method, almost always used in 2-norm minimization problem could have brought the non-stability percentage down to unimportant levels, we decided not to use them as the sparseness properties of the residual were contaminated.

In [11] an interesting method was introduced for both having an intrinsically stable solution as well as keeping the computational cost down using (4): the Burg Method for AR parameters estimation based on the least absolute forward-backward error. In this approach to find a solution, however, the sparseness is not preserved (as shown in Figure 4). This is mostly due to the decoupling of the main K-dimensional minimization problem in K one-dimensional minimization sub-problems, this is in contrast with our algorithm that tries to find a minimum in the $K-$dimensional cost function: therefore this method is suboptimal. The 1-norm Burg algorithm has shown to behave somewhere in between the 1-norm and the 2-norm minimization. Regarding the computational costs, finding the solution of a overdetermined system of equations in the 1-norm using a modern interior point algorithm [14] showed to be comparable to solving around 10-15 least square problem; however the further processes, for example open and closed loop analysis for pitch estimation and algebraic excitation search (in the case of code-excited schemes [15]) and quantization in general, will

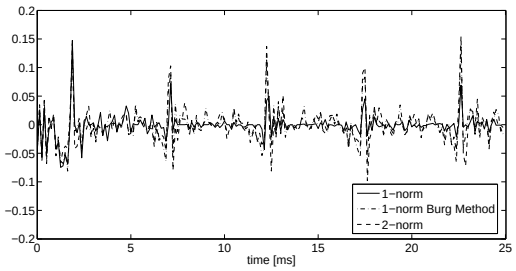

Figure 4: *Comparison of the residuals obtained with the method used in the paper (continuous), the Burg method based on the 1-norm (dash-dotted) and the usual LP (dashed).*

be highly simplified by the characteristics of the output. Furthermore, it's important to notice that the residual signal will be already available at the end of the computation and doesn't have to be calculated.

It is also useful to combine the optimization problems (4) and (6); in this case the following optimization problem arises:

$$\min_{\mathbf{a}} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_1 + \gamma\|\mathbf{a}\|_1. \tag{9}$$

Here, the coefficients of a high-order predictor combining the short and long term predictors are found such that both the coefficient vector and the residual are sparse to better quantize the residual. In our experimental work we were able to efficiently encode a speech signal (with both voiced and unvoiced parts) using a significantly low bit rate by using only 20% of the coefficients of each predictive vector and setting approximately 85% of the residual samples equals to zero with a quantizer that ignores samples below a certain adaptive treshold and a quasi-linear quantization elsewhere. Although more intensive studies are needed to determine the psycho-acoustic level performances of this simple scheme, the time domain distorsion and quality seemed comparable to the common encoding-decoding techniques used in GSM and UMTS based on 2-norm minimization.

6. Conclusions

In this paper, two kinds of sparse linear predictor have been introduced. Specifically, linear predictors that offer a sparse residual or a sparse coefficients vector or the combination of both, as a particular case of the latter one, have been formulated, discussed and evaluated. Although this kind of methods seemed particularly attractive for the analysis and coding of stationary voiced signal, we have seen that the extension of the obtained results to unvoiced signal seemed to be straightforward and it will be subject to further analysis. Furthermore, considering other convex estimators will easily bring to new studies based on different concepts of sparseness. It should be noted that the algorithms introduced are not restricted to speech processing and can be used for several linear prediction problems where either the residual or the coefficient vector is expected to show sparseness properties or where we want these to fit a sparse model.

7. Acknowledgments

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 Mads Græsbøll Christensen is supported by the Parametric Audio Processing project, Danish Research Council for Technology and Production Sciences, grant no. 274060521.

8. References

[1] J. H. L. Hansen, J. G. Proakis, and J. R. Deller, Jr., *Discrete-Time Processing of Speech Signals*, Prentice-Hall, 1987.

[2] J. Makhoul, "Linear Prediction: A Tutorial Review", *Proc. IEEE*, vol. 63(4), pp. 561–580, Apr. 1975.

[3] P. Kroon and W. B. Kleijn, "Linear-prediction based analysis-by-synthesis coding", in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Elsevier Science B.V., 1995, ch. 3, pp. 79–119.

[4] 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*, 2004, pp. 2250–2254.

[5] B. S. Atal and J. R. Remde, "A new model of LPC excitation for producing natural sounding speech at low bit rates", in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 7, 1982, pp. 614 – 617.

[6] P. Kroon, E. D. F. Deprettere, and R. J. Sluyter, "Regular-pulse excitation - a novel approach to effective multipulse coding of speech", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34, pp. 1054–1063, 1986.

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

[8] P. Stoica and R. Moses, *Spectral Analysis of Signals*, Pearson Prentice Hall, 2005.

[9] 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.

[10] J.-R. Ohm, *Multimedia Communication Technology: Representation, Transmission, and Identification of Multimedia Signals*, Springer-Verlag, 2004.

[11] 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.

[12] M. Reed and B. Simon, *Methods of Modern Mathematical Physics II: Fourier Analysis, Self-adjointness*, Academic Press, 1975.

[13] S. C. Narula and J. F. Wellington, "The Minimum Sum of Absolute Errors Regression: A State of the Art Survey", *International Statistical Review*, Vol. 50(3), pp. 317-326, Dec. 1982.

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

[15] W. C. Chu, *Speech Coding Algorithms: Foundation and Evolution of Standardized Coders*, Wiley, 2003

[16] P. Kabal and R. P. Ramachandran, "Joint optimization of linear predictors in speech coders", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37(5), pp. 642–650, May 1989.

[17] H. Zarrinkoub and P. Mermelstein, "Joint optimization of short-term and long-term predictors in CELP speech coders", in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, 2003, pp. 157–160.