Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 76 (2016) 143–156

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Fast algorithms for high-order sparse linear prediction with applications to speech processing

Tobias Lindstrøm Jensen [a,*,1], Daniele Giacobello [a], Toon van Waterschoot [b], Mads Græsbøll Christensen [c]

[a] Signal and Information Processing, Department of Electronic Systems, Aalborg University, Denmark
[b] Center for Dynamical Systems Signal Processing and Data Analytics (STADIUS), Department of Electrical Engineering (ESAT), KU Leuven, Belgium
[c] Audio Analysis Lab, AD:MT, Aalborg University, Denmark

Received 9 January 2015; received in revised form 7 September 2015; accepted 21 September 2015
Available online 19 October 2015

Abstract

In speech processing applications, imposing sparsity constraints on high-order linear prediction coefficients and prediction residuals has proven successful in overcoming some of the limitation of conventional linear predictive modeling. However, this modeling scheme, named sparse linear prediction, is generally formulated as a linear programming problem that comes at the expenses of a much higher computational burden compared to the conventional approach. In this paper, we propose to solve the optimization problem by combining splitting methods with two approaches: the Douglas–Rachford method and the alternating direction method of multipliers. These methods allow to obtain solutions with a higher computational efficiency, orders of magnitude faster than with general purpose software based on interior-point methods. Furthermore, computational savings are achieved by solving the sparse linear prediction problem with lower accuracy than in previous work. In the experimental analysis, we clearly show that a solution with lower accuracy can achieve approximately the same performance as a high accuracy solution both objectively, in terms of prediction gain, as well as with perceptually relevant measures, when evaluated in a speech reconstruction application.

Keywords: Sparse linear prediction; Speech and audio processing; Linear programming; Real-time optimization; Speech reconstruction; Packet loss concealment

1. Introduction

Linear prediction (LP) is a well understood technique for the analysis, modeling, and coding of speech signals (Vaidyanathan, 2009). The widespread use of LP of speech can be attributed to its correspondence to the source-filter model of speech production (Makhoul, 1975; Bäckström, 2004). An emitted speech sound can be modeled as a combination of the excitation process (the air flow) and the filtering process (vocal tract effect). The vocal tract can, to a large extent, be modeled as a slow varying low-order all-pole filter, while the air flow can be modeled by a white noise sequence, for unvoiced sounds, or an impulse train generated by periodic vibrations of the vocal chords pulses, for voiced sounds (Hansen et al., 1987).

In speech analysis, the purpose of the all-pole model obtained through LP is to construct a spectral envelope that models the behavior of the vocal tract. For a segment of unvoiced speech, considering the excitation of the

[*] Corresponding author.
E-mail addresses: tlj@es.aau.dk (T.L. Jensen), giacobello@ieee.org (D. Giacobello), toon.vanwaterschoot@esat.kuleuven.be (T. van Waterschoot), mgc@create.aau.dk (M.G. Christensen).

[1] The work of T.L. Jensen is supported by the Danish Council for Independent Research, Technology and Production Sciences. Grant No. 4005-00122.

http://dx.doi.org/10.1016/j.specom.2015.09.013
0167-6393/© 2015 Elsevier B.V. All rights reserved.

all-pole filter as white noise, the envelope is the same as its power spectrum of and the LP model coincides theoretically with the autoregressive (AR) model (Stoica and Moses, 2005). However, for a segment of voiced speech, the connection is more complex. The power spectrum of the voiced speech signal has a clear harmonic structure that can be approximated more effectively as a line spectrum (Christensen and Jakobsson, 2009). The line frequencies are located at the multiples of the pitch frequency and their amplitude are given by the shape of the spectral envelope.

The all-pole coefficients are usually identified by minimizing the mean-squared (2-norm) error of the difference between the observed signal and the predicted signal (Atal and Hanauer, 1971). In the source-filter model, this approach yields the LP all-pole filter, thus the prediction error (the residual signal) represents the source. Unvoiced speech lends itself readily to the principles of the 2-norm error criterion as a means of estimating the model parameters (Makhoul, 1975). Furthermore, the 2-norm is consistent with an i.i.d. Gaussian interpretation of the prediction residual (Saito and Itakura, 1967; Itakura and Saito, 1970). The quality of the 2-norm based LP all-pole model in the context of voiced speech, which is approximately two-thirds of speech, is questionable and, theoretically, not well-founded. In particular, the all-pole spectrum does not provide a good spectral envelope and sampling the spectrum at the line frequencies does not provide a good approximation of their amplitudes (Murthi and Rao, 2000). In general, the shortcomings of LP in spectral envelope modeling can be traced back to the 2-norm minimization.[2] In particular, analyzing the goodness of fit between a given harmonic line spectrum and its LP model, as done in Makhoul (1975), a major flaw can be derived. The LP tries to cancel the input voiced speech harmonics causing the resultant all-pole model to have poles close to the unit circle. Consequently, the LP spectrum tends to overestimate the spectral powers at the formants, providing a sharper contour than the original vocal tract response. A wealth of methods have been proposed to mitigate these effects. Some of the proposed techniques involve a general rethinking of the spectral modeling problem (notably El-Jaroudi and Makhoul, 1991; Murthi and Rao, 2000; Ekman et al., 2008) while some others are based on changing the statistical assumptions made on the prediction error in the minimization process (notably Lee, 1988; Denoel and Solvay, 1985). Many other formulations for finding the parameter of the all-pole model exist, a special mention is for methods that include perceptual knowledge into the estimation process (e.g., Hermansky, 1990; Magi et al., 2009), or account for the non-linearities in the speech production model, e.g., Thyssen et al. (1994).

Despite the wealth of alternative methods introduced to overcome the deficiencies of the 2-norm criterion, traditional usage of LP methods is, however, still confined to modeling only the spectral envelope (the vocal tract transfer function), i.e., the short-term redundancies of speech. Hence, in the case of voiced speech, the predictor does not fully decorrelate the speech signal because of the long-term redundancies of the underlying pitch excitation. This means that the residual will still have pitch pulses present and the spectrum will still show a clear harmonic structure. The usual approach is then to employ a cascaded structure where, after LP is initially applied to determine the short-term prediction coefficients, a long-term predictor is determined to model the harmonic behavior of the spectrum (Hansen et al., 1987). Such a structure is arguably suboptimal since it ignores the interaction between the two different stages (Kameoka et al., 2010; Bensaid and Slock, 2012). This is known in the literature and early contributions have outlined gains in performance in jointly estimating the two filters (the work in Kabal and Ramachandran (1989) is perhaps the most successful attempt). The combination of the two filters determines a high-order linear predictor with a pretty evident sparse characteristics.

In recent work (Giacobello et al., 2008; Giacobello et al., 2012), a more general framework for LP was presented with several benefit by introducing sparsity in the LP minimization framework. This was renamed sparse linear prediction (SpLP). In particular, while reintroducing well-known methods to seek a short-term predictor that produces a residual that is sparse rather than minimum variance (e.g., Denoel and Solvay, 1985; Murthi and Rao, 1998), the idea of employing high-order SpLP (HOSpLP) to model the cascade of short-term and long-term predictors was also introduced (Giacobello et al., 2009,). The application of HOSpLP was originally introduced for speech processing purposes, however its formulation is intimately related to the regularization of ill-conditioned problems and to the precise modeling of long-term redundancies, thus it quickly found applications in diverse fields, such as radar (Erer et al., 2014), geology (Bochud et al., 2013), video packet-loss concealment (Koloda et al., 2013), and general signal representations (Angelosante et al., 2013; Angelosante, 2014).

The SpLP problem can be posed as a linear programming problem, a special case of convex optimization. In order to be deployed in real-time applications, it requires its convex optimization core to be embedded directly in the algorithm that runs online and where strict real-time constraints apply. While convex optimization problems can be efficiently solved, both in theory, with worst-case polynomial complexity (Nesterov and Nemirovskii, 1994), and in practice, such as Andersen et al. (2003), it is rarely

[2] To the authors' knowledge, the "original sin" behind the use of the 2-norm in LP, comes from its first application in speech coding, trying to reduce the entropy of speech for more efficient encoding than simple differential pulse code modulation (Atal, 2006). The fundamental theorem of predictive quantization (Gersho and Gray, 1992) states that the mean-squared reproduction error in predictive encoding is equal to the mean-squared quantization error when the residual signal is presented to the quantizer. Therefore, by minimizing the 2-norm of the residual, these variables have a minimal variance whereby the most efficient coding is achieved.