Sparsity in Linear Predictive Coding of Speech

Giacobello, Daniele

Publication date:
2010

Document Version
Early version, also known as pre-print

Sparsity in Linear Predictive Coding of Speech

Ph.D. Thesis

Daniele Giacobello

Multimedia Information and Signal Processing
Department of Electronic Systems
Aalborg University
Niels Jernes Vej 12, 9220 Aalborg Ø, Denmark

Sparsity in Linear Predictive Coding of Speech
Ph.D. Thesis

August 2010

Abstract

This thesis deals with developing improved techniques for speech coding based on the recent developments in sparse signal representation. In particular, this work is motivated by the need to address some of the limitations of the well-known linear prediction (LP) model currently applied in many modern speech coders.

In the first part of the thesis, we provide an overview of *Sparse Linear Prediction*, a set of speech processing tools created by introducing sparsity constraints into the LP framework. This approach defines predictors that look for a sparse residual rather than a minimum variance one with direct applications to coding but also consistent with the speech production model of voiced speech, where the excitation of the all-pole filter can be modeled as an impulse train, i.e., a sparse sequence. Introducing sparsity in the LP framework will also bring to develop the concept of high-order sparse predictors. These predictors, by modeling efficiently the spectral envelope and the harmonics components with very few coefficients, have direct applications in speech processing, engendering a joint estimation of short-term and long-term predictors. We also give preliminary results of the effectiveness of their application in audio processing.

The second part of the thesis deals with introducing sparsity directly in the linear prediction analysis-by-synthesis (LPAS) speech coding paradigm. We first propose a novel near-optimal method to look for a sparse approximate excitation using a compressed sensing formulation. Furthermore, we define a novel re-estimation procedure to adapt the predictor coefficients to the given sparse excitation, balancing the two representations in the context of speech coding. Finally, the advantages of the compact parametric representation of a segment of speech, given by the sparse linear predictors and the use of the re-estimation procedure, are analyzed in the context of frame independent coding for speech communications over packet networks.

List of Papers

The main body of this thesis consists of the following papers:

[A] D. Giacobello, M. G. Christensen, J. Dahl, S. H. Jensen, and M. Moonen, "Sparse Linear Predictors for Speech Processing," in *Proceedings of the 9th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 1353–1356, 2008.

[B] 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," in *Proceedings of the 34th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4109–4112, 2009.

[C] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen, "Speech Coding Based on Sparse Linear Prediction," in *Proceedings of the 17th European Signal Processing Conference (EUSIPCO)*, 2009, pp. 2524–2528.

[D] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen, "Enhancing Sparsity in Linear Prediction of Speech by Iteratively Reweighted 1-norm Minimization," in *Proceedings of the 35th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4650–4653, 2010.

[E] 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 on Audio, Speech, and Language Processing*, 2010.

[F] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen, "Stable Solutions for Linear Prediction of Speech Based on 1-norm Error Criterion," to be submitted to *IEEE Transactions on Audio, Speech, and Language Processing*, 2010.

[G] D. Giacobello, T. van Waterschoot, M. G. Christensen, S. H. Jensen, and M. Moonen, "High-Order Sparse Linear Predictors for Audio Processing," accepted for publication in *Proceedings of the 18th European Signal Processing Conference (EUSIPCO)*, 2010.

[H] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen, "Retrieving Sparse Patterns Using a Compressed Sensing Framework: Applications to Speech Coding Based on Sparse Linear Prediction," in *IEEE Signal Processing Letters*, vol. 17, no. 1, pp. 103–106, 2010.

[I] D. Giacobello, M. N. Murthi, M. G. Christensen, S. H. Jensen, and M. Moonen, "Re-estimation of Linear Predictive Parameters in Sparse Linear Prediction," in *Conference Record of the 43rd Asilomar Conference on Signals, Systems and Computers*, pp. 1770–1773, 2009.

[J] D. Giacobello, M. N. Murthi, M. G. Christensen, S. H. Jensen, and M. Moonen, "Estimation of Frame Independent and Enhancement Components for Speech Communication over Packet Networks," in *Proceedings of the 35th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4682–4685, 2010.

The following papers have also been published by the author of this thesis during the Ph.D. studies:

[1] D. Giacobello, M. Semmoloni, D. Neri, L. Prati and S. Brofferio, "Voice Activity Detection Based on the Adaptive Multi-Rate Speech Codec Parameters," in *Proc. 11th International Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2008.

[2] D. Giacobello, D. Neri, L. Prati and S. Brofferio, "Acoustic Echo Cancellation on the Adaptive Multi-Rate Speech Codec Parameters," in *Proc. 11th International Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2008.

Preface

This thesis is submitted to the International Doctoral School of Technology and Science at Aalborg University in partial fulfillment of the requirements for the degree of Doctor of Philosophy. The main body consists of a number of papers that have been published in or have been submitted to peer-reviewed conferences and journals. The work was carried out during the period from September 2007 through August 2010 at the Multimedia Information and Signal Processing Group of the Department of Electronic Systems at Aalborg University. It was funded by the European Union Marie Curie SIGNAL Fellowship, contract no. MEST-CT-2005-021175.

There are many people I am indebted to and, without their guidance and encouragement, achieving this important goal in my life would have not been possible. First and foremost, my sincere gratitude goes to my supervisor, Prof. Søren Holdt Jensen, for giving me the opportunity of pursuing a Ph.D. degree and for providing me with a perfect working environment to fully develop my potential. He also supported and encouraged me in all my decisions and provided me with very valuable advices. I also thank my co-promoter within the Marie Curie SIGNAL project, Prof. Marc Moonen, for his invaluable comments on all my papers and also for making my stay at the Katholieke Universiteit Leuven a very pleasant experience. I would also like to extend my gratitude to my co-supervisor Prof. Mads Græsbøll Christensen. Since I first started my Ph.D. studies, he has taken me "under his wings" providing me with some of the ideas he had developed by introducing sparsity constraints in the linear predictive framework. As soon as we started working on it, those ideas truly became the "goose that laid the golden eggs," and form the core of this thesis. I owe him a great deal and it has been a privilege to work with him, his mentoring has undoubtedly helped me throughout this great scientific adventure.

This thesis is also, to a large extent, the result of collaboration with other people, and my various co-authors also deserve an honorable mention here. First of all, Prof. Manohar N. Murthi deserves to be thanked for the technical discussions during my stay at the University of Miami and the very fruitful collaboration that sprung out of them. I would also like to thank Dr. Joachim Dahl

for his precious insights on convex optimization, and Dr. Toon van Waterschoot for the highly beneficial talks on how to extend our work to other application scenarios.

The best part of my Ph.D. studies has undoubtedly been getting to meet and work with many amazing people. In this regards, I would like to acknowledge my present and former colleagues at the Multimedia Information and Signal Processing Group at Aalborg University and all the people at Katholieke Universiteit Leuven, Instituto Superior Tecnico Lisbon, and University of Nice who were involved in the SIGNAL project for the countless interesting technical discussions and the fun times we had at our numerous meetings. I would also like to express my gratitude to Charlotte Skindbjerg Pedersen and the whole administrative staff at Aalborg University for taking care of the bureaucratic matters, thus making my working life easier.

In the personal sphere, I would like to thank many people that have been close to me in the past several years and, directly or indirectly, have contributed to this work. In particular (in rigorous alphabetical order): Alessandro, Alessio, Alvaro, Andrea, Behzad, Emilia, Francesca, Gian Paolo, Giulia, Ismael, Kim, Jason, Johan, Lucia, Marco, Mario, Marta, Meg, Pedro, Pierre-Louis, Rocco, Romain, Sabato, Shaminda, Tobias, and Virginia.

My largest debt of gratitude is toward my parents. They have been the pillars on which I could hold on to at any moment in my life. They have guided, inspired, encouraged, and supported me. Above all, they have always believed in me. This thesis is dedicated to them. A special thought goes also to all of my family for their unconditional love and support.

Finally, I would like to thank Shadi for her support, encouragement, patience, and unwavering love, which made these past three years the best of my life.

Daniele Giacobello
Aalborg University, August 2010

Contents

Introduction

In speech coding systems, linear prediction (LP) based all-pole modeling is, arguably, the most used parametric technique for modeling the spectral envelope and capturing the short-term redundancies of a speech signal [1, 2]. These features have led LP to become a fundamental part of many coding architectures since the early works on speech coding [3–5] to the most recent proposals for unified speech and audio coders (e.g., [6–9]). In these cases, LP is used to remove most of the correlations present in a segment of speech, rendering a so-called LP *analysis* filter and a *residual* signal. In order to provide a parsimonious bit representation of this residual signal, a search is usually performed to find the best possible *excitation* of the inverse LP analysis filter, the all-pole *synthesis* filter, given certain constraints on it. This coding paradigm is referred to as Linear Predictive Analysis-by-Synthesis (LPAS) and it has set the standard for speech coding for the past thirty years [10, 11].

The optimization problems encountered in the LPAS speech coding paradigm, namely the LP analysis and the modeling of the excitation, fall in the more general mathematical framework of linear inverse problems, where the model parameters are estimated from a set of observed data [12]. In these problems, the 2-norm minimization criterion has found a widespread use, mostly for its amenability of producing an optimization problem that is attractive both theoretically and computationally. While the 2-norm minimization is consistent with producing a representation with minimal energy, in many signal processing applications it is more beneficial to find solutions with the fewest nonzero coefficients as possible, i.e., a maximal *sparse* solution [13]. Even if examples of the applications of the sparsity measure can be found in early literature for various types of signals and applications (e.g., [14–18]), the use of sparsity in signal processing has grown significantly in the recent years due to the increasing use of transform domain representations (notably, wavelets [19] for images and modified discrete cosine transform, MDCT [20], for audio), for which a concise signal representation in a given domain is required.

This introductory overview is organized as follows. In Section 1 we first elaborate on the speech modeling problem and the popularity of the LP method. In

Section 2 we provide a brief overview of the main stages of the LPAS coding paradigm. In Section 3, we give a summary of the problem formulation and applications of sparsity in signal processing. In Section 4, we address our own contributions where we investigate the properties and applications of sparse signal representation in the LPAS speech coding paradigm. Finally, in Section 5 we sum up the conclusions of this work. As an appendix to this introduction, Section 6 provides some conjectures on the future challenges that await the speech coding community and how some of the topics discussed in this thesis could actually play a role in these challenges.

1 Background

In this section, we first elaborate on the speech modeling problem, and then highlight the limitation of the popular LP method in the context of speech analysis and coding, thus providing a motivation for this research work.

1.1 The Source-Filter Model of Speech Production

The theory behind the widespread use of LP all-pole modeling of speech, arises from the source-filter model of speech production [21]. The general idea is that the emitted speech sound is a combination of the excitation process (the air flow) and the filtering process (vocal tract effect). Historically, the first registered experimental analysis of this theory was done in 1848 by Johannes Müller by blowing air through the larynges excised from a human cadavers [22]. While the experimental evaluation of this theory has evolved since then[1], the fundamentals have not gone through dramatic changes and can be summarized as follows. Speech production is initiated at the lungs by generating air pressure that flows through the trachea, vocal folds, pharynx, oral and nasal cavities[2].

There are, roughly speaking, two different ways in which speech sounds are produced, leading to classify them in two main categories, i.e., *voiced* and *unvoiced* [24]. In the case of voiced speech (e.g., vowels /a/, /o/ and /i/, and nasals /m/ and /n/), the flow of air coming from the lungs excites the vocal folds in an oscillating motion, periodically inhibiting the airflow for a short interval. The periodicity of these intervals determines the fundamental frequency of the source and contributes to the perceived *pitch* of the produced sound and it is then called *pitch period* [24]. Consequently, voiced speech sounds consist of a strong periodic component rich in harmonics. Secondly, for unvoiced speech, airflow is constricted (e.g., fricatives /f/, /s/ and /h/) or completely stopped for a short

[1] Some of the most recent approaches to the analysis of the speech model includes magnetic resonance imaging (MRI) (see, e.g., [23])

[2] In general, cavities above the vocal folds are collectively called the vocal tract.

interval (e.g., stops /t/, /p/ and /k/). Therefore, unvoiced speech is of either noise-like or impulsive-like characteristics, without harmonic structure [21, 25].

The clear relation between the physics of speech production and the theory of sound wave propagation [26], has led to some of the first attempts to provide a mathematical model for speech production in acoustics rather than signal processing [27–31]. In fact, like any acoustic cavity, the vocal tract has resonances that attenuate and amplify different frequency regions. These resonances, in speech science, are called the *formants* and can be modified by movements of the vocal organs, such as tongue, lips and pharynx [32]. While these early works on this topic suffered quite consistently from high requirements on specific a priori knowledge of the voice, Bishnu Atal, in [33], greatly simplified the model by approximating the vocal tract with a lossless tube made by cylindrical sections of equal length but different diameter. In particular, exploiting the relations of the lossless tube model with digital filters, he demonstrated that the formant frequencies and bandwidths are sufficient to uniquely determine the tube model parameters and that this model can always be represented as a transfer function with K poles when the number of sections of the lossless tube is K. This was (and still is) remarkable since it also proved to be also consistent with his early work. Specifically, in [3] Atal first used the concept of *predictive coding* [34] in digital speech processing to decorrelate a speech segment by applying a order K prediction filter. In [33], Atal therefore linked these two theories by showing that the prediction filter is theoretically consistent with the speech production model, since the corresponding order K all-pole model carries the information of the tube model of the vocal tract. In [5], Atal also introduced the discrete speech production model. In this model, the speech signal is analyzed and synthesized as the output of a discrete linear all-pole time-varying filter, which is excited by a periodic pulse train (in the case of voiced speech) or by white noise (in the case of unvoiced speech).

1.2 LP Based Speech Analysis

To understand fully the digital implementations of the source-filter model, it is first useful to distinguish between the power spectrum and the spectral envelope of a speech signal. The goal of the all-pole models is to define a spectral envelope that provides a model of the vocal tract in speech production. For unvoiced speech, considering the excitation of the all-pole filter as white noise, the envelope is the same as the power spectrum. For voiced speech, the connection is more complex. The power spectrum of the voiced speech signal has a clear harmonic structure that can be approximated as a line spectrum. The line frequencies are located at the multiples of the pitch frequency and their amplitude is given by the shape of the spectral envelope.

In the above mentioned pioneering work done by Atal, the all-pole coefficients

are identified by minimizing the mean-squared (2-norm) error of the difference between the observed signal and the predicted signal [5]. This forms a set of equation known in time series literature as the Yule-Walker equations for *autoregressive* (AR) model fitting [35] for which, at that time, already existed a computationally efficient algorithm, the Levinson recursion[3] [36]. 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, which can be modeled as white noise passed through an all-pole filter, lends itself readily to the principles of the 2-norm error criterion as mean of estimating the model parameters [40]. Also, considering the statistical interpretation of the 2-norm minimization with the fitting of the error in a Gaussian i.i.d. distribution [38, 39].

The quality of the LP all-pole model in the context of voiced speech, which is approximately two-thirds of speech[4], is questionable and, theoretically, is 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 approximate of their amplitudes.

In general, the shortcomings of LP in spectral envelope modeling can be traced back to the 2-norm minimization. In particular, analyzing the the goodness of fit between a given harmonic line spectrum and its LP model[5], as done in [40], two major flaws 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 [41–44]) while some others are based on changing the statistical assumptions made on the prediction error in the minimization process (notably [45, 46]). 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., [47–51]). Non-linear prediction methods have also been developed, the most successful attempts are based on the application of neural networks [52] and Volterra filters [53, 54].

1.3 LP Based Speech Coding

The first attempts documented on the application of predictive coding to speech were based on the idea of reducing the first-order entropy [55] of the distribution

[3] In Levinson's own words, a "mathematically trivial procedure."

[4] Of the phonemes in standard English prose, vowels and diphthongs form approximately 38%, voiced consonants 40% and unvoiced consonants 22% [24]

[5] This can be done due to the correspondence of the 2-norm error minimization in time and frequency domain given by Parseval's theorem.

of digital speech so to produce a representation that would require a lower bit rate. According to Atal [56], he was able to reduce the entropy of a 5 ms speech segment sampled at 6.67 kHz from 3.3 b/sample to 1.3 b/sample by applying a $10th$ order predictor. While almost 60 years have passed, the idea is still present today in speech coding, i.e., the 2-norm based LP is used to decorrelate the input leaving a residual that is ideally white, and therefore easier to quantize. This approach is also consistent with the fundamental theorem of predictive quantization. This 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 [57]. Therefore, by minimizing the 2-norm of the residual, these variables have a minimal variance whereby the most efficient coding is achieved.

Nevertheless, the 2-norm based LP shows severe shortcomings also in the speech coding scenario. Firstly, traditional usage of LP is confined to modeling only the spectral envelope, capturing 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. Furthermore, while the 2-norm criterion is consistent with achieving minimal variance of the residual for efficient coding, the excitation is usually estimated with some constrained structure on it. In particular, *sparse* techniques are employed to model the excitation for efficient coding [56]. Examples of this can be seen since early works on speech coding with the introduction of multipulse excitation (MPE [58]) and regular-pulse excitation (RPE [59]) methods and, more recently, in sparse algebraic codes in code-excited linear prediction (ACELP [11]). Early contributions (notably [46, 60, 61]) have followed this line of thought questioning the fundamental validity of the 2-norm criterion with regards to speech coding.

1.4 Why is 2-norm based LP still so popular?

Despite such a rich literature addressing the deficiencies of 2-norm based LP in speech analysis and coding, one might wonder why, to the author's best knowledge, the 2-norm minimization is the only criterion used in commercial speech codecs. There are several explanation that we address below, going around the same concept: simplicity.

- **Mathematical tractability.** The minimization of the 2-norm of the prediction error results in the Yule-Walker equations and can be efficiently solved via the Levinson recursion. The 2-norm cost function is strongly convex allowing for a unique solution [62]. The roots of the corresponding all-pole filter are guaranteed to be inside the unit circle, since stability is intrinsically guaranteed by the construction of the problem [63].

- **Statistical Interpretation.** This method corresponds to the maximum likelihood (ML) approach when the error signal is considered to be a set of i.i.d. Gaussian variables. The Gaussian p.d.f. is arguably the most used and well know distribution for tractable mathematics [64, 65]. In [39], the Yule-Walker equations are derived from the maximum likelihood approach.
- **Frequency-Domain Interpretation.** According to the Parseval's theorem, minimizing the 2-norm of the error in the time-domain is equivalent to minimizing the error ratio between the true and estimated spectra [40]. It is also interesting to notice that minimizing the squared error in the time domain and in the frequency domain leads in both cases to the Yule-Walker equations [66].

2 Linear Prediction Based Analysis-by-Synthesis Coding

In this section, we will give an overview of the the three main stages of the LPAS coding paradigm: LP analysis, pitch analysis, and modeling of the excitation. While several other stages make up the LPAS coding scheme and should not be overlooked for an efficient implementation of a speech coder (i.e., pre-processing, post-processing, quantization, and other implementation issues [67]), in these three stages the three main contributions to the parametrization of a speech signal are estimated.

2.1 Linear Predictive Analysis

The fundamental idea behind LP is that a speech sample $x(n)$ can be approximated as a linear combination of past samples [40]:

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

where $\{a_k\}$ are the prediction coefficients, $e(n)$ is prediction error. Assuming that $x(n) = 0$ for $n < 1$ and $n > N$, the speech production model (1) for a segment of N speech samples in matrix form becomes:

$$\mathbf{x} = \mathbf{X}\mathbf{a} + \mathbf{e}, \tag{2}$$

where:

$$\mathbf{x} = \begin{bmatrix} x(N_1) \\ \vdots \\ x(N_2) \end{bmatrix}, \mathbf{X} = \begin{bmatrix} x(N_1-1) & \cdots & x(N_1-K) \\ \vdots & & \vdots \\ x(N_2-1) & \cdots & x(N_2-K) \end{bmatrix}, \tag{3}$$

the weights used to compute the linear combination are found by minimizing the prediction error:

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_p^p, \tag{4}$$

where

$$\mathbf{x} = \begin{bmatrix} x(N_1) \\ \vdots \\ x(N_2) \end{bmatrix}, \mathbf{X} = \begin{bmatrix} x(N_1 - 1) & \cdots & x(N_1 - K) \\ \vdots & & \vdots \\ x(N_2 - 1) & \cdots & x(N_2 - K) \end{bmatrix}, \tag{5}$$

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 $n > N$ [66]. The most common approach is to choose $N_1 = 1$ and $N_2 = N + K$, equivalent, when $p = 2$, to the *autocorrelation method*:

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_2^2 = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{x}. \tag{6}$$

We can rewrite the system of equation as:

$$\hat{\mathbf{a}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{x} = \mathbf{R}^{-1}\mathbf{r}, \tag{7}$$

where $\mathbf{R} = \mathbf{X}^T\mathbf{X}$ is the autocorrelation matrix and $\mathbf{r} = \mathbf{X}^T\mathbf{x}$ is the cross-correlation vector. In general, the inversion of $\mathbf{R}$ is not necessary, since finding $\hat{\mathbf{a}}$ in (7) corresponds to solving the Yule-Walker equations, and this can be done efficiently with the Levinson recursion (also called Levinson-Durbin algorithm) [40].

2.2 The Excitation Model

In this subsection we describe the most common encoding strategies for the excitation signal. This is the key of the analysis-by-synthesis procedure, in fact, while the previous stage to determine the LP coefficients $\hat{\mathbf{a}}$ is done in an open-loop configuration, the choice of the excitation $\hat{\mathbf{r}}$ is done in a close-loop configuration (so the name analysis-by-synthesis) where the perceptually weighted error between the true speech segment and its synthesized version is minimized. Since $\hat{\mathbf{r}}$ has usually some structural constraints on it, our problem formulation becomes:

$$\hat{\mathbf{r}} = \arg\min_{\mathbf{r}} \|\mathbf{W}(\mathbf{x} - \mathbf{H}\mathbf{r})\|_2^2, \quad \text{s.t.} \quad \mathbf{struct}(\mathbf{r}); \tag{8}$$

where $\mathbf{H}$ is a $N \times N$ lower-triangular convolution matrix, called the *synthesis matrix*, created from the impulse response of the LP synthesis filter and $\mathbf{W}$ is the $N \times N$ perceptual weighting matrix. In speech coding, $\mathbf{W}$ is adaptively chosen according to the prediction filter parameter in order to "concentrate" the error

in the frequency regions perceptually less sensitive, i.e., where the formants are located. Hence the choice of making $\mathbf{W}$ dependent from the prediction filter parameters $\mathbf{a}$ that represent them [68]. It should be noted that, in general, to take into consideration the previous frames of speech, a non-square $\mathbf{H}$ matrix can be used so to include the previous samples of the excitation. The operator $\mathbf{struct}(\cdot)$ we have introduced in (8), represents the structural constraints usually imposed on the excitation, i.e., the modeling strategy used for efficient coding. There are mainly two approaches to model the excitation. The first approach is the multipulse encoding, where only few samples are selected in the excitation, setting to zero most of the other samples. The second approach is to model the excitation from a codebook of predefined possible excitations.

Multipulse Excitation

In multipulse encoding (MPE) coders, the excitation consists of K freely located pulses in each segment of length N. This problem is made impractical by its combinatorial nature and a suboptimal algorithm was proposed in [58] where the sparse residual is constructed one pulse at a time. Starting with a zero residual, pulses are added iteratively adding one pulse in the position that minimizes the error between the original and reconstructed speech. The pulse amplitude is then found minimizing the distortion in the analysis-by-synthesis scheme. The procedure can be stopped either when a maximum fixed number of amplitudes is found or when adding a new pulse does not improve the quality. MPE provides an approximation to the optimal approach, when all possible combinations of K positions in the approximated residual of length N are analyzed, i.e.:

$$\hat{\mathbf{r}} = \arg\min_{\mathbf{r}} \|\mathbf{W}(\mathbf{x} - \mathbf{H}\mathbf{r})\|_2^2 \quad \text{s.t.} \quad \|\mathbf{r}\|_0 = K. \tag{9}$$

The main problem of the MPE procedure is that the K pulses by being freely located, they also require a significant amount of bits to be spent on describing their location on the excitation sequence. The regular-pulse encoding (RPE) [59] addressed exactly this issue, in this case the pulses are constrained on a grid with spacing S. It also allows S possible shifts of the grid and therefore only S possible configuration of the location of the pulses.

Codebook Excitation

The RPE can be considered as the first idea to include a predetermined structure on the excitation [69]. This idea has also been developed, around the same time, in code-excited LP (CELP) [70, 71]. Ideally, the excitation should be a white random sequence and therefore the sequence could be selected by a predetermined codebook populated by "random white noise" sequences. The

problem in (8), would then become:

$$\hat{\mathbf{r}} = \arg\min_{\mathbf{c}} \|\mathbf{W}(\mathbf{x} - \mathbf{H}\mathbf{c})\|_2^2, \quad \text{s.t.} \quad \mathbf{c} \in C; \tag{10}$$

where C is the codebook and $\mathbf{c}$ is a codeword. The general idea, is also to have the sequences pre-quantized, thus truly selecting the optimal sequence to be sent to the encoder. However the basic scheme led to huge computational loads [56]. The introduction of algebraic codebooks, and its corresponding paradigm (algebraic code-excited LP, ACELP), posed a remedy to this. Algebraic codebooks are deterministic codebooks in which the codebook vectors are determined from the transmitted index using simple algebra rather than lookup tables or predefined codebooks. This structure has advantages in terms of storage, search complexity, and robustness [72, 73].

2.3 Modeling the Pitch Periodicity

In speech coding, the LP analysis is usually performed to remove short-term correlation, however, voiced speech segments exhibits strong long-term correlation components due to the presence of a pitch excitation. To account for these correlations, two strategies are usually implemented. The first one is to find a long-term linear predictor, the second one is to model the periodicity directly in the excitation model.

Pitch Prediction

This interpretation is similar to modeling the short-term correlations, and it is the first strategy implemented to account for long-term correlations [4]. The pitch predictor has a small number of taps N_p (usually 1 to 3) and the corresponding delays associated are usually clustered around a value which corresponds to the estimated integer pitch period T_p. The more general form for $N_p = 1$ is:

$$P(z) = 1 - g_p z^{-T_p}. \tag{11}$$

The parameters g_p and T_p are determined by minimizing the residual error signal after the LP predictor, similarly to the minimization problem occurring in estimating the short-term prediction. In order to reduce the computational effort, usually T_p is estimated before the error minimization to find the pitch predictor coefficients [74]. In general, T_p is not integer, thus a noninteger pitch period T_p is usually incorporated in the prediction model in two ways: either by using a multitap pitch prediction model for interpolation (see, e.g., [75]) or by using a fractional delay filter [74], for which numerous design methods exist [76]. The frequency response of $P(z)$ is a comb-like structure, thus resembling a line spectrum, consistent with the harmonic structure of the voiced speech sounds.

Adaptive Codebook

The other interpretation is the one that is currently mostly used in LPAS speech coding. The strategy is to account for the periodicity in the modeled excitation. In particular, the excitation can be seen as a linear combination between a pseudo-random component $\mathbf{c}_f$, and a periodic component given by the pitch excitation $\mathbf{c}_a$ [77]:

$$\hat{\mathbf{r}} = g_f \mathbf{c}_f + g_a \mathbf{c}_a \tag{12}$$

where $\mathbf{c}_f$ is now called the *fixed* codeword ($\mathbf{c}_f \in C_f$) and $\mathbf{c}_a$ is the so-called *adaptive* codeword ($\mathbf{c}_a \in C_a$), g_f and g_a are their respective gains. While including the structure of (12) in (10) is impractical, the common approach is to begin with the search for the adaptive codebook, based on a open-loop estimate of the pitch period T_p, and then determine the fixed codeword [78]. The adaptive codeword is built up based on the pitch period T_p and its gain, similarly to what it is done in (11).

3 Sparsity in Signal Processing

Sparse approximation approaches have enjoyed considerable popularity in recent signal processing applications. The use of sparsity has shown to be particularly efficient in many applications such as signal compression [79], denoising [80], image restoration [81, 82], and, blind source separation [83, 84], etc. Depending on the application, sparsity can be sought on the residual being minimized, or on the solution being computed. In this brief overview, we concentrate on this latter problem, which is also the one mainly covered in the sparse signal processing literature. However, these ideas do have relevance to the problem of computing a sparse residual, as we shall see throughout the contributions of this thesis.

The idea behind sparse approximation is that many natural signals have a concise representation when expressed in the proper basis. In other words, for most signal classes, it is possible to find a basis or a dictionary of elementary building blocks with respect to which most signals in the class may be expanded, so that when the expansion is truncated in a suitable way, high precision approximations are obtained even when very few terms are retained. A large number of signal processing "success stories" may be described in such a way, including image compression and denoising using wavelets [79] (or more sophisticated -lets, such as curvelets [86]) audio coding using MDCT bases [85], and so forth.

It is interesting to notice that some of the first works where sparsity was successfully applied was indeed speech coding. In particular, one of the first ideas for efficient coding was that one could produce speech of any desired quality by providing a sufficient number of pulses at the input of the synthesis filter [58]. Finding the location and amplitudes of the pulses, resulted to solving a linear

inverse problem with the sparsity constraints (9). In this case, the basis is represented by the synthesis matrix and the domain where sparsity is sought after is the excitation domain.

In this section, we introduce the original problem formulation and an overview of the current literature on the several efficient sparse expansion algorithms that have been proposed throughout the years. In particular, we will focus our attention on greedy algorithms [87] and parallel basis selection methods based on the minimization of different diversity measures [88]. While other methods are available in literature to find sparse representation (notably, Bayesian methods [89] and nonconvex optimization [90]), these two approaches are computationally practical and lead to provably correct solutions [91].

3.1 Problem Formulation

The canonical form of the problem of sparse signal representation from a redundant dictionary or basis, is given by:

$$\min_{\mathbf{x}} \|\mathbf{x}\|_0, \quad \text{s.t.} \quad \mathbf{A}\mathbf{x} = \mathbf{b} \tag{13}$$

where $\mathbf{A} \in \mathbb{R}^{N \times M}$ is a matrix whose columns $\mathbf{A}_i$ represent an overcomplete or redundant basis (i.e., $\text{rank}(\mathbf{A}) = N$ and $M > N$) determined from the physics of the problem. The goal is to solve for $\mathbf{x} \in \mathbb{R}^M$ vector, from the measurements vector (or given signal) $\mathbf{b} \in \mathbb{R}^N$. The cost function being minimized $\|\cdot\|_0$ is the 0-norm of $\mathbf{x}$, i.e. the cardinality of $\mathbf{x}$. The general idea is that $\mathbf{x}$ is K-sparse ($K \ll M$), i.e., only K entries in $\mathbf{x}$ are sufficient to reconstruct $\mathbf{b}$ without distortion. An alternative formulation to (13), popular when accounting for modeling errors or measurement noise is:

$$\min_{\mathbf{x}} \|\mathbf{x}\|_0, \quad \text{s.t.} \quad \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 \leq \epsilon \tag{14}$$

Unfortunately, both (13) and (14) are combinatorial problem, and the search for the optimal K-sparse representation would require solving up to $\binom{M}{K}$ linear systems, making it impractical for even modest values of M and K. Consequently, in practical situations, there is a need for approximate methods that efficiently solve (13) or (14).

3.2 Algorithms

As mentioned above, winnowing through the all $\binom{M}{K}$ possibilities to determine the optimal K-sparse solution is impractical. In this subsection, we will describe the general concepts behind the most used methods for determining a sparse solution. The methods can be divided in two classes. Greedy methods that "break" the optimization problem in a sequence of smaller problems in which

a optimal solution can be easily found. Convex optimization relaxations that replace the combinatorial problem with a related convex program.

Greedy Algorithms

The first approaches to solve (13) and (14), are the one based on greedy algorithms, iteratively solving the sparse approximation problem applying a sequence of locally optimal choices in an effort to determine a globally optimal solution. In this category, notably falls the matching pursuit (MP) algorithm [92], a technique which involves finding the "best matching" projections of multidimensional data onto an overcomplete dictionary ($\mathbf{A}$ in our formulation). This is a recursive strategy that involves choosing, at a given iteration, the column $\mathbf{A}_i$ that is most aligned with the current residual vector. The procedure usually terminates when the given sparsity level K is achieved.

The main deficiency of MP type algorithms is related to the general limits of greedy algorithms, i.e., if the algorithms picks a wrong column at a given iteration, there is no possibility of correcting this error in the following iterations [93]. To cope with this problem, an alternative method to MP, but based on the same concept, was developed. This is the orthogonal matching pursuit (OMP) [94–96]. The main idea behind OMP is to add a least-square minimization in the selection of the basis so to obtain a better approximation over the columns of $\mathbf{A}$ that have already been chosen. Following this line of thought, the cyclic matching pursuit (CMP) was also developed [97].

Minimizing Diversity Measures

Backed up by significant improvements in convex optimization algorithms [100, 101], this category is certainly the one that has received the most interest lately.

The first ideas was introduced in [98] with the development of the basis pursuit (BP) principle. Differing substantially from MP and OMP, BP was based on the idea that the number of terms in a representation (i.e., the cardinality), can be approximated by the absolute sum of coefficients. Thus, the idea is to perform a convex relaxation of the 0-norm, replacing the combinatorial sparse approximation with a problem solvable with convex tools that also lead to sparse solutions. Differently from greedy algorithms, it is based on global optimization, thus, in general, finds improved sparse solutions [91]. The 1-norm is arguably chosen for this purpose as the closest convex approximation to the non-convex 0-norm [99]. The problems (13) will then become:

$$\min_{\mathbf{x}} \|\mathbf{x}\|_1, \quad \text{s.t.} \quad \mathbf{A}\mathbf{x} = \mathbf{b} \tag{15}$$

and (14) equivalently:

$$\min_{\mathbf{x}} \|\mathbf{x}\|_1, \quad \text{s.t.} \quad \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 \leq \epsilon \tag{16}$$

Furthermore, many recent algorithms have exploited the sparsity inducing properties of the 1-norm to find more focal solutions to the original problems by iteratively reweighting the minimization process [102–104]. The choice of the weights, as the inverse of the magnitude of the coefficients, is made to penalize every nonzero coefficient equally, as done by the 0-norm. In [102] and [104], it is also shown that the reweighted 1-norm algorithm, at convergence, is equivalent to the minimization of the log-sum penalty function. This is relevant to the original problem formulation in (13) and (14): the log-sum cost function has a sharper slope near zero compared to the 1-norm, providing more effective sparsity inducing properties. Furthermore, since the log-sum is not convex, the iterative algorithm corresponds to minimizing a sequence of linearizations of the log-sum around the previous solution estimate, providing at each step a sparser solution (until convergence). In the class of methods to compute sparse solutions through reweighting, thus by emphasizing and de-emphasizing the different contributions of the columns of $\mathbf{A}$ in the solution $\mathbf{x}$, a distinctive mention is for the FOcal Underdetermined System Solver (FOCUSS) algorithm [105] based on the reweighted 2-norm algorithm.

4 Summary of Contributions

The main contributions of the work that is documented in this thesis is to propose new approaches to the optimization problems encountered in LPAS coding by introducing the sparsity constraint. Papers A through F deal with sparse speech modeling, obtained introducing sparsity constraints directly in the LP based all-pole modeling of speech. Paper G extends the use of sparsity in the LP framework to the analysis of monophonic and polyphonic audio signals. In Paper H sparsity is introduced in the stage of selection the approximated excitation in the analysis-by-synthesis equations that follow the all-pole modeling stage. Paper I defines a new approach to LPAS coding, taking into account the approximated excitation in deriving a new set of LP parameters; in paper J, we apply this method to define a two-layered speech coder for packet networks. We will now go through the contributions of the individual papers that constitute the main body of this thesis.

Paper A

This paper introduces a generalized LP framework and provides some preliminary numerical experiments and conjectures on the use of sparsity constraints in it. Two classes of LP schemes are presented for voiced speech analysis. The first class aims at finding a predictor that outputs a sparse residual rather than a minimum variance one. Its use produces a residual with a clearer spiky behavior compared to traditional LP. The second class aims at finding a high-order

sparse predictor. The estimated sparse high-order predictor exhibits a clear resemblance to the high-order predictor obtained by convolving the short-term and long-term predictors obtained in two different stages.

Paper B

The objective of this paper is to investigate the use of the high-order sparse predictor for the joint estimation of short-term and long-term predictor. In particular, the high-order sparse predictor can be factorized into a short-term predictor and long-term predictor that offer a better estimate compared to the traditional multistage approach. The high-order predictor is also more effective in finding a prediction error that is also spectrally whiter and therefore easier to model and quantize through pseudo-random codewords. This method is implemented into an ACELP scheme and offer improvements in coding efficiency, also compared to other joint estimation methods.

Paper C

This paper describes a novel speech coding concept created by introducing sparsity constraints in the linear prediction scheme both on the residual and on the high-order prediction vector. The sparse residual obtained allows a more compact representation, while the sparse high-order predictor engenders a robust joint estimation of short-term and long-term predictors. Thus, the main purpose of this work is showing that better statistical modeling in the context of speech analysis creates an output that offers better coding properties. We compare the implemented coder with the RPE-LTP coder, showing that just a change in the LP estimation approach achieves a more parsimonious description of a speech segment with interesting direct applications to low bit-rate speech coding.

Paper D

While in Papers A-C, the 1-norm has been reasonably chosen as a convex approximation of the so-called 0-norm, the true sparsity measure, in this paper we apply the reweighted 1-norm algorithm in order to produce a more focused solution to the original combinatorial problem that we are originally trying to solve. The purpose of the reweighted scheme is then to overcome the mismatch between 0-norm minimization and 1-norm minimization while keeping the problem solvable with convex estimation tools. The experimental analysis shows improvements over the previously used 1-norm based estimators, producing sparser solutions.

Paper E

The objective of this paper is twofold. Firstly, we put our earlier contributions (Papers A-D) in a common framework giving an introductory overview of Sparse Linear Prediction and we also introduce its compressed sensing formulation. Secondly, we provide a detailed experimental analysis of its usefulness in modeling and coding applications transcending the well known limitations related to traditional LP. In particular, we provide a thorough analysis of the effectiveness of the sparse predictors in modeling the speech production process. Furthermore, we give several results as proof of the usefulness of introducing sparsity in the LP framework for speech coding applications. This provides, not only a more synergistic new approach to encode a speech segment, but also several interesting properties such as shift independence, pitch independence and a slower decaying quality for decreasing SNR. The compressed sensing formulation for sparse LP introduced is also very helpful in reducing the size of the minimization problem, and hence to keep the computational costs reasonable.

Paper F

Compared to traditional LP based on the 2-norm minimization, the minimization of the 1-norm process will offer a residual that is sparser, providing tighter coupling between the multiple stages of time-domain speech coders, and thereby enabling more efficient coding. Nevertheless, unlike those obtained through 2-norm minimization, the predictors obtained through 1-norm minimization are not intrinsically stable and, in coding application, unstable filters may create problems, generating saturations in the synthesized speech. In this paper, we introduce several alternative methods to 1-norm linear prediction comparing the spectral modeling and coding performances of the alternative predictors.

Paper G

The main purpose of this paper is to extend the use of high-order sparse predictors to the audio processing scenario. In particular several experiments will be provided to show how these 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. The main strength of the high-order sparse predictors, as evinced from this paper, is that they can achieve spectral flatness properties comparable to traditional high-order LP with very few coefficients compared to the order of the predictor. This shows possible applications for a more efficient use of LP in several audio related problems.

Paper H

In this paper, we devise a compressed sensing formulation to compute a sparse approximation of speech in the residual domain in the Analysis-by-Synthesis equations. In particular, in our previous work defined a sparse predictive framework that aims for a sparse prediction residual rather than the traditional minimum variance residual. We have also shown that MPE techniques are better suited in this framework for finding a sparse approximation of the residual rather than pseudo-random sequences (e.g., algebraic codes). Considering that MPE is itself a suboptimal approach to modeling prediction residuals, in this paper we aim at improving the performance of MPE by moving toward a better approach of capturing the approximated excitation without increasing complexity. We compare the method of computing a sparse prediction residual with the optimal technique based on an exhaustive search of the possible nonzero locations and the well known MPE. Experimental results demonstrate the potential of compressed sensing in speech coding techniques finding with high probability the true sparse solution.

Paper I

The usual approach in Analysis-by-Synthesis coding is to first find the linear prediction parameters in a open-loop configuration then searching for the best excitation given certain constraints on it. This is done in a closed-loop configuration where the perceptually weighted distortion between the original and synthesized speech waveform is minimized. The conceptual difference between a quasi-white true residual and its approximated version, where usually sparsity is taken into consideration (e.g. ACELP, RPE, MPE coding schemes), creates a mismatch that can raise the distortion significantly. In this paper, we estimate the optimal truncated impulse response that creates the given sparse coded residual without distortion. An all-pole approximation of this impulse response is then found using a least square approximation. The all-pole approximation is a stable linear predictor that allows a more efficient reconstruction of the segment of speech. In this case, the autoregressive modeling is no more employed as a method to remove the redundancies of the speech segment but as IIR approximation of the optimal FIR filter, adapted to the quantized approximated residual, which is used in the synthesis of the speech segment.

Paper J

In this paper, we exploit the compact speech segment representation given by sparse linear prediction and the re-estimation procedure introduced in Paper I to create two representations within a segment of coded speech. One representation that allows for decoding a speech frame independently, and one that acts as an

enhancement layer and it is frame dependent. This introduces a new approach to speech coding over packet networks, creating a coder that has speech frames with a core that is independently decodable and an enhancement layer that is based on the previously received frames. In particular, we create a coder that can select between two decoding procedures, if the previous frames are received correctly, then it decodes using all the information, otherwise, it uses only the frame independent information. By doing so, we offer the flexibility of a frame independent codec if the loss probability is significant but, if the probability is low (or ideally null), then it will exploit inter-frame dependencies to perform similarly to a frame dependent coder.

5 Conclusions

In this work, we have introduced several new approaches to the LPAS problem for speech analysis and coding obtained by introducing sparsity into the LPAS coding framework.

When sparsity is applied in the generalized LP minimization framework, the sparse linear predictors have been shown to provide a more efficient decoupling between the pitch harmonics and the spectral envelope. This translates into predictors that are not corrupted by the fine structure of the pitch excitation and offer interesting properties such as shift invariance and pitch invariance. In the context of speech coding, the sparsity of residual and of the high-order predictor provides a more synergistic new approach to encode a speech segment by reducing the burden on the excitation sequence, offering significant benefits for low bit-rate applications. In particular, the sparse residual obtained allows a more compact representation, while the sparse high-order predictor engenders joint estimation of short-term and long-term predictors. A compressed sensing formulation is used to reduce the size of the minimization problem, and hence to keep the computational costs reasonable. The sparse linear prediction based robust encoding technique provided a competitive approach to speech coding with a synergistic multistage approach and a slower decaying quality for decreasing SNR. Some preliminary results on the possible applications of the sparse linear predictive framework in audio processing, has also shown to be effective transcending some of the limitation of traditional linear prediction.

In the second part of this work, we have concentrated our attention on the complete structure of the encoder, introducing new strategies to code the excitation sequence based on the compressed sensing formulation, creating a computationally efficient near-optimum multipulse approach. We have also proposed a new method for the re-estimation of the prediction parameters in speech coding. In particular, the autoregressive modeling is no more employed as a method to remove the redundancies of the speech segment but as IIR approximation of the

optimal FIR filter, adapted to the quantized approximated excitation, that is used in the synthesis of the speech segment. The method has shown improvements in the general performances of the sparse linear prediction framework, providing tradeoffs between the complexity, and thus the bit-rate, of the two descriptions hitherto not possible. An interesting incarnation of the proposed framework is the possibility of estimating predictors and residuals that create a independently decodable frame of speech. This has been successfully applied in a novel way to code speech in packet networks, creating a frame independent description and an frame dependent description that acts as enhancement layer, exploiting inter-frame redundancies.

6 Outlook

In the author's opinion, the increasing demand for Voice-over-IP (VoIP) telephony, that can carry also music and mixed audio contents, will arguably offer some of the most important challenges in the speech coding community in the following years. The current trend is to merge well deployed existing codecs optimized for speech and audio and used them jointly to offer the best possible quality [6–9]. In particular, embedded coding, proposes a multi-layer approach to sound coding mixing transform based (e.g., MDCT) codecs for audio with traditional LP based codecs for speech. This approach has two main weaknesses. Firstly, it does not provide a common coding strategy to speech and audio and its flexibility is to simply switch between different codecs depending on the input signal (this also pointed out in [110]). Secondly, these codecs achieve high quality with low bit-rate mostly thanks to the exploitation of inter-frame dependencies showing severe shortcomings in the presence of packet loss. Therefore, it is interesting to focus future research to find a common coding framework for speech and audio that achieves superior robustness to packet loss by providing frame independent coding.

We here give a brief overview to the future role that some of the topics discussed in this thesis could play in these above mentioned issues.

6.1 Provide a Common Coding Framework for Speech and Audio Coding

It is well known that transform based coders are not suitable for speech coding, mostly due to their inadequate modeling of the speech signal that cannot achieve a low bit rate. Other reasons are, the computational demands of the transforms used [106], and the algorithmic delay that necessarily arise, especially at high sampling rate. On the other hand, LP has been fundamentally abandoned as a possible candidate for audio coding since low-order LP seems to

be appropriate in modeling only when the harmonic components are distributed uniformly on the spectrum [107]. Nevertheless, the LP filter is generally a quite adequate tool to model the spectral peaks which play a dominant role in perception [108]. 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. In our work, we have proposed to use high-order sparse linear predictors for audio and speech processing. These tools have shown to be quite attractive in modeling the harmonic behavior of audio and speech signals, achieving a concise parametric representation by exploiting harmonicity and achieving accurate spectral modeling consistent with high-order LP [109]. Their use could provide a possible common coding framework for both speech and audio signals.

Furthermore, the complexity of the encoding strategy in audio and wideband speech (and recently super-wideband) coding is strongly dependent on the sampling frequency of the initial acquisition procedure. The encoding structure often relies on mirror filterbanks in order to proceed with a less computationally demanding subband approach. In our approach, we come across the same complexity issue dealing with high-order predictors and long residual vectors. Nevertheless, since we have defined that both audio and speech are *sparse* in the prediction and residual domain, we can effectively reduce the number of measurement applying a compressed sensing formulation. This formulation, which we have efficiently applied in finding a sparse residual, can be also easily extended to the estimation of the high-order sparse predictor. Also, we are not using any predefined basis, thus providing a truly adaptive sparse representation for our processed speech and audio signals.

6.2 Redefine the LPAS Coding Scheme

In simple terms, the LPAS approach is to first find the linear prediction parameters in a open-loop configuration then searching for the best excitation given certain constraints on it. This second step is done in a closed-loop configuration where the perceptually weighted distortion between the original and synthesized speech waveform is minimized. Since the predictor is quantized *transparently*, all the responsibility for the distortion falls on the choice of the excitation. A consequence of this approach can be seen, for example, in the AMR-WB coder, where, in its 23.85 kbit/s configuration, 80% of the bits are allocated for the excitation and only 10% for the predictor [111].

In our work, we have proposed several ways to generally improve performances of the LPAS scheme, reducing the burden on the excitation signal. For example, the general idea of our proposed re-estimation procedure for the predictor was to find a tradeoff between the complexity of the excitation and the complexity of the predictor. This idea can be easily extended to performing a

tradeoff of the sparse representation of the excitation and the sparse representation of the high-order sparse predictor also considering that there is, arguably, a clear relation between sparsity and rate. Early approaches have also outlined gains by including the LP parameters in the closed-loop configuration [112, 113].

Furthermore, the LP model computation (ignoring quantization) accounts for a minimal part of the total computational effort in the LPAS encoders, significantly less than the search for the excitation [111]. Thus, the time might be right to revisit the current approaches in LPAS coding, balancing both the bit allocation and computational effort.

6.3 Provide Frame Independent Coding

As mentioned above, the codecs used for embedded coding present strong dependencies from both present and future frames. The exploitation of the redundant information present in neighboring frames helps considerably in reducing the bit rate. Nevertheless, while this approach is consistent in the case of telephony with dedicated circuits, in packet networks these dependencies create well known problems. While Packet Loss Concealment (PLC) strategies have achieved a certain degree of maturity [114–121], it is still important to reduce, if not eliminate, these dependencies making each frame independently decodable, as done, for example in [122]. The coding algorithm we have presented is representative of a more general rate-distortion problem. In our case, the distortion will be dependent on how the representation of the speech segment is divided between a frame independent core and a frame dependent enhancement layer. In particular, the distortion term can be made dependent on the loss rate and therefore adjusting the bit allocation on the frame dependent and frame independent parts. While future studies are obviously necessary, the preliminary studies and results presented in this thesis have shown this to be a viable road.

References

[1] J. D. Markel and A. H. Gray, *Linear prediction of speech*, Springer-Verlag, New York, 1980.

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

[3] B. S. Atal and M. R. Schroeder, "Predictive coding of speech," in *Proc. Conf. Communications*, pp. 360–361, 1967.

[4] B. S. Atal and M. R. Schroeder, "Adaptive predictive coding of speech," *Bell Syst. Tech. J.*, vol. 49, no. 8, pp. 1973–1986, 1970.

[5] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Amer.*, vol. 50, pp. 637–655, 1971.

[6] B. Geiser, P. Jax, P. Vary, H. Taddei, M. Gartner, and S. Schandl, "A Qualified ITU-T G.729EV Codec Candidate for Hierarchical Speech and Audio Coding," *Proc. IEEE Workshop on Multimedia Signal Processing*, pp. 114–118, 2006.

[7] S. Ragot, B. Kövesi, R. Trilling, D. Virette, N. Duc, D. Massaloux, S. Proust, B. Geiser, M. Gartner, S. Schandl, H. Taddei, Y. Gao, E. Shlomot, H. Ehara, K. Yoshida, T. Vaillancourt, R. Salami, M. S. Lee, and D. Y. Kim, "ITU-T G.729.1: An 8-32 kbit/s scalable coder interoperable with G.729 for wideband telephony and Voice over IP," *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 4, pp. 529–532, 2007.

[8] B. Geiser, H. Kruger, H. W. Lollmann, P. Vary, D. Zhang, H. Wan, H. T. Li, and L. B. Zhang,, "Candidate proposal for ITU-T super-wideband speech and audio coding," *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 4121–4124, 2009.

[9] M. Neuendorf, P. Gournay, M. Multrus, J. Lecomte, B. Bessette, R. Geiger, S. Bayer, G. Fuchs, J. Hilpert, N. Rettelbach, R. Salami, G. Schuller, R. Lefebvre, and B. Grill, "Unified speech and audio coding scheme for high quality at low bitrates," *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 1–4, 2009.

[10] 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., ch. 3, pp. 79–119, 1995.

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

[12] R. C. Aster, C. H. Thurber, and B. Borchers, *Parameter Estimation and Inverse Problems*, Elsevier 2004.

[13] B. D. Rao, "Signal Processing with the Sparseness Constraint," *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 1, pp. 369–372, 1998.

[14] I. F. Gorodnitsky, J. S. George, and B. D. Rao, "Neuromagnetic source imaging with FOCUSS: a recursive weighted minimum norm algorithm," *J. of Electroenceph. and Clinical Neuroph.*, vol. 95, no. 4, pp. 231–251, 1995.

[15] R. M. Leahy and B. D. Jeffs, "On the design of maximally sparse beamforming arrays," *IEEE Transactions on antennas and propagation*, vol. 29, no. 8, pp. 1178–1187, 1991.

[16] S. D. Cabrera and T. W. Parks, "Extrapolation and spectral estimation with iterative weighted norm modification," *IEEE Trans Acoust., Speech, Signal Processing*, vol. 39, no. 4, pp. 842–851, 1991.

[17] S. Singhal and B. S. Atal, "Amplitude Optimization and Pitch Prediction in Multipulse Coders," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 3, pp. 317–327 , 1989.

[18] M. O' Brien, A. N. Sinclair, S. M. Kramer, Recovery of a sparse spike time series by L_1 Norm Deconvolution, *IEEE Trans. Signal Processing*, vol. 43, no. 12, pp. 3353–3365, 1994.

[19] S. G. Mallat, *A wavelet tour of signal processing*, Academic Press, 1999.

[20] A. W. Johnson and A. B. Bradley, "Adaptive transform coding incorporating time domain aliasing cancellation," *Speech Comm.*, vol. 6, no. 4, pp. 299–308, 1987.

[21] G. Fant, *Acoustic Theory of Speech Production*, Mouton, The Hague, 1960.

[22] P. Lieberman, *The Biology and Evolution of Language*, Harvard University Press, Cambridge, 1984.

[23] T. Baer, J. C. Gore, S. Boyce, P. W. Nye, "Application of MRI to the analysis of speech production," *Magn. Reson. Imaging*, vol. 5, no. 1, pp. 1–7, 1987.

[24] J. L. Flanagan, *Speech Analysis Synthesis and Perception*, Springer Verlag, Berlin, 1972.

[25] D. O'Shaughnessy, *Speech Communication - Human and Machine*, Addison-Wesley, New York, 1987.

[26] L. L. Beranek, *Acoustics*, McGraw-Hill, New York, 1954.

[27] M. R. Schroeder, "Determination of the Geometry of the Human Vocal Tract," *J. Acoust. Soc. Amer.*, no. 41, pp. 1002–1010, 1967.

[28] A. P. Mermelstein, "Determination of the Vocal Tract Shape from Measured Formant Frequencies," *J. Acoust. Soc. Amer.*, no. 41, pp. 1283–1294, 1967.

[29] J. Heinz, "Perturbation functions for the determination of Vocal Tract Area Functions from Vocal Tract Eigenvalues," *Quart. Progr. Status Rep.*, Speech Transmission Lab., Roy. Inst. Tech. Stockholm, Sweden, pp. 1–14, 1967.

[30] B. Gopinatha and M. M. Sondhi, "Determination of the Shape of the Human Vocal Tract from Acoustical Measurements," *Bell Sys. Tech. J.*, vol. 49, pp. 1195–1214, 1970.

[31] M. M. Sondhi and B. Gopinatha, "Determination of Vocal-Tract Shape from Impulse Response at the Lips," *J. Acoust. Soc. Amer.*, vol. 49, pp. 1867–1873, 1971.

[32] T. D. Rossing, *The Science of Sound*, Addison-Wesley, Reading, 1990.

[33] B. S. Atal, "Determination of the vocal tract shape directly from the speech wave," *J. Acoust. Soc. Amer.*, vol. 47, p. 65, 1970.

[34] P. Elias, "Predictive Coding I," *IRE Transactions on Information Theory*, vol. 1, no. 1, pp. 16–24.

[35] P. J. Brockwell and R. A. Davis, *Time series: theory and methods*, Springer-Verlag, 1987.

[36] N. Levinson, "The Wiener RMS error criterion in filter design and prediction," *J. Math. Phys.*, vol. 25, pp. 261–278, 1947.

[37] J. Durbin, "The fitting of time series models," *Rev. Inst. Int. Stat.*, vol. 28, pp. 233–243, 1960.

[38] S. Saito and F. Itakura, "Theoretical consideration of the statistical optimum recognition of the spectral density of speech," *J. Acoust. Soc. Japan*, 1967.

[39] F. Itakura and S. Saito, "A statistical method for estimation of speech spectral density and formant frequencies," *Electron. Commun. Japan*, vol. 53A, pp. 36–43, 1970.

[40] J. Makhoul, "Linear Prediction: A Tutorial Review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[41] A. El-Jaroudi and J. Makhoul, "Discrete All-Pole Modeling," *IEEE Trans. Signal Processing*, vol. 39, pp. 411–423, 1991.

[42] M. N. Murthi and B. D. Rao, "All-pole modeling of speech based on the minimum variance distortionless response spectrum," *IEEE Trans. Speech and Audio Processing*, vol. 8, pp. 221–239, 2000.

[43] L. A. Ekman, W. B. Kleijn and M. N. Murthi, "Regularized linear prediction of speech," *IEEE Trans. Audio, Speech, Language Processing*, vol. 16, no. 1, pp. 65–73, 2008.

[44] H. Hermansky, H. Fujisaki, Y. Sato, "Spectral envelope sampling and interpolation in linear predictive analysis of speech," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 9, pp. 53–56, 1984.

[45] C.-H. Lee, "On Robust linear prediction of speech," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. 36, No. 5, pp. 642–650, 1988.

[46] E. Denoël and J.-P. Solvay, "Linear prediction of speech with a least absolute error criterion," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 33, no. 6, pp. 1397–1403, 1985.

[47] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Amer.*, vol. 87, no. 4, pp. 1738–1751, 1990.

[48] P. L. Chu and D. G. Messerschmitt, "Frequency weighted linear prediction," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 1318–1321, 1982.

[49] D. Petrinovic, "Discrete weighted mean square all-pole modeling," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, pp. 828–831, 2003.

[50] S. Varho and P. Alku, "Separated linear prediction - a new all-pole modeling technique for speech analysis," *Speech Comm.*, vol. 24, pp. 111–121, 1998.

[51] C. Magi, J. Pohjalainen, T. Backstrom, P. Alku, "Stabilised weighted linear prediction," *Speech Communication*, vol. 51, no. 5, pp. 401–411, 2009.

[52] B. Townshend, "Nonlinear prediction of speech," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, pp. 425–428, 1991.

[53] J. Thyssen, H. Nielsen, and S. D. Hansen, "Non-linear short-term prediction in speech coding," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, pp. 185–188, 1994.

[54] T. Shimamura and H. Hayakawa, "Adaptive nonlinear prediction based on order statistics for speech signals," *Proc. Eurospeech 1999*, pp. 347–350, 1999.

[55] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, 1948.

[56] B. S. Atal, "The History of Linear Prediction," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 154–161, 2006.

[57] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, Kluwer Academic Publishers, 1993.

[58] 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, pp. 614–617, 1982.

[59] P. Kroon, E. F. Deprettere, and R. J. Sluyter, "Regular-pulse excitation - a novel approach to effective and efficient multipulse coding of speech", *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 34, no. 5, pp. 1054–1063, 1986.

[60] J. Lansford and R. Yarlagadda, "Adaptive L_p approach to speech coding," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, pp. 335–338, 1988.

[61] M. N. Murthi and B. D. Rao, "Towards a synergistic multistage speech coder," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, pp. 369–372, 1998.

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

[63] L. Knockaert, "Stability of linear predictors and numerical range of shift operators in normed spaces," *IEEE Trans. on Information Theory*, vol. 38, no. 5, pp. 1483–1486, 1992.

[64] S. Kotz, N. Balakrishnan, N. L. Johnson, *Continuous Multivariate Distributions, Volume 1, Models and Applications*, 2nd edition, Wiley, 2000.

[65] S. Nagarajah, "A generalized normal distribution," *Journal of Applied Statistics*, vol. 32, no. 7, pp. 685–694, 2005.

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

[67] W. B. Kleijn and K. K. Paliwal Eds., *Speech Coding and Synthesis*, Elsevier Science B.V., 1995.

[68] B. S. Atal and M. Schroeder, "Predictive coding of speech signals and subjective error criteria," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 27, pp. 247–254, 1979.

[69] E. F. Deprettere and P. Kroon, "Regular excitation reduction for effective and efficient LP-coding of speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 25.8.1–25.8.4, 1985.

[70] B. S. Atal and M. R. Schroeder, "Stochastic coding of speech signals at very low bit rates," *Proc. Int. Conf. Commun.*, pp. 1610–1613, 1984.

[71] M. R. Schroeder and B. S. Atal, "Code-excited linear prediction (CELP): High-quality speech at very low bit rates," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 937–940, 1985.

[72] J. Adoul, P. Mabilleau, M. Delprat, and S. Morisette, "Fast CELP coding based on algebraic codes," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 1957–1960, 1987.

[73] J. Adoul and C. Lamblin, "Comparison of some algebraic structures for CELP coding of speech," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 1953–1956, 1987.

[74] 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.

[75] R. P. Ramachandran and P. Kabal, "Pitch prediction filters in speech coding," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 37, no. 4, pp. 467–478, 1989.

[76] T. I. Laakso, V. Valimaki, M. Karjalainen, and U. K. Laine, "Splitting the unit delay," *IEEE Signal Processing Magazine*, vol. 13, no. 1, pp. 30–60, 1996.

[77] W. B. Kleijn, "On the periodicity of speech coded with linear-prediction based analysis by synthesis Coders," *IEEE Trans. on Speech, and Audio Processing*, vol. 2, no. 4, pp. 539–542, 1994.

[78] R. C. Rose, "Design and performance of an analysis-by-synthesis class of predictive speech coders," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 38, no. 9, pp. 1489–1503, 1990.

[79] S. G. Mallat, *A wavelet tour of signal processing: the Sparse way*, Academic press, 2009.

[80] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Transactions on Inf. Theory*, vol. 41, no. 3, pp. 613–627, 1995.

[81] M. A. Figueiredo and R. D. Novak, "An EM algorithm for wavelet-based image restoration," *IEEE Trans. Image Process.*, vol. 12, no. 8, pp. 906–916, 2003.

[82] J. Kalifa, S. Mallat, and B. Rouge, "Deconvolution by thresholding in mirror wavelet bases," *IEEE Trans. Image Process.*, vol. 12, no. 6, pp. 446–457, 2003.

[83] R. Gribonval and S. Lesage, "A survey of sparse component analysis for blind source separation: principles, perspectives, and new challenges," *Proceedings of ESANN*, vol. 6, pp. 323–330, 2006.

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

[85] D. Pan, "A tutorial on MPEG/audio compression," *IEEE Trans. on Multimedia*, vol. 2, no. 2, pp. 60–74, 1995.

[86] J.-L. Starck, E. J. Candes, and D. L. Donoho, "The curvelet transform for image denoising," *IEEE Trans. Image Process.*, vol. 11, no. 6, pp. 670–684, 2002.

[87] J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Trans. Information Theory*, vol. 50, no. 10, pp. 2231–2242, 2004.

[88] J. A. Tropp, "Just relax: Convex programming methods for identifying sparse signals," *IEEE Trans. Information Theory*, vol. 52, no. 3, pp. 1030–1051, 2006.

[89] D. Wipf and B. D. Rao, "Sparse Bayesian learning for basis selection," *IEEE Trans. Signal Processing*, vol. 52, no. 8, pp. 2153–2164, 2004.

[90] R. Chartrand, "Exact reconstruction of sparse signals via nonconvex minimization," *IEEE Signal Processing Lett.*, vol. 14, no. 10, pp. 707–710, 2007.

[91] J. A. Tropp and S. J. Wright, "Computational methods for sparse solution of linear inverse problems," Accepted to *Proc. IEEE, special issue on applications of sparse representation and compressive sensing*, 2010.

[92] S. G. Mallat and Z. Zhang, "Matching Pursuits with Time-Frequency Dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, 1993.

[93] R. DeVore and V. N. Temlyakov, "Some remarks on greedy algorithms," *Adv. Comput. Math.*, vol. 5, pp. 173–187, 1996.

[94] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Rec. Asilomar Conf. Signals, Systems and Computers*, pp. 40–44, 1993.

[95] G. Davis, S. Mallat, and Z. Zhang, "Adaptive time-frequency decompositions with matching pursuit," *Opt. Eng.*, vol. 33, no. 7, pp. 2183–2191, 1994.

[96] G. Davis, S. Mallat, and M. Avellaneda, "Adaptive greedy approximation," *J. Constr. Approx.*, vol. 13, pp. 57–98, 1997.

[97] M. G. Christensen and S. H. Jensen, "The Cyclic Matching Pursuit and Its Application to Audio Modeling and Coding," in *Rec. Asilomar Conf. Signals, Systems, and Computers*, pp. 550–554, 2007.

[98] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61, 1999.

[99] D. L. Donoho and M. Elad, "Optimally sparse representation from overcomplete dictionaries via ℓ^1-norm minimization," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 100, no. 5, pp. 2197–2202, 2002.

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

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

[102] E. J. Candés, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5, pp. 877–905, 2008.

[103] B. D. Rao, K. Engan, S. F. Cotter, J. Palmer, and K. Kreutz-Delgado, "Subset selection in noise based on diversity measure minimization," *IEEE Trans. Signal Process.*, vol. 51, no. 3, pp. 760–770, 2003.

[104] D. Wipf and S. Nagarajan, "Iterative reweighted ℓ_1 and ℓ_2 methods for finding sparse solutions," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 2, pp. 317–329, 2010.

[105] I. Gorodnitsky and B. D. Rao, "Sparse signal reconstruction from limited data using FOCUSS: A re-weighted minimum norm algorithm," *IEEE Trans. Signal Processing*, vol. 45, no. 3, pp. 600–616, 1997.

[106] V. Britanak and K. Rao, "An efficient implementation of the forward and inverse MDCT in MPEG audio coding," *IEEE Signal Processing Letters*, vol. 8, no. 2, pp. 48–51, 2001.

[107] 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.

[108] 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.

[109] P. Stoica and T. Soderström, "High-order Yule-Walker equations for estimating sinusoidal frequencies: The complete set of solutions," *Signal Processing*, Vol. 20, No. 3, pp. 257–263, 1990.

[110] N. H. van Schijndel, J. Bensa, M. G. Christensen, C. Colomes, B. Edler, R. Heusdens, J. Jensen, S. H. Jensen, W. B. Kleijn, V. Kot, B. Kovesi, J. Lindblom, D. Massaloux, O. A. Niamut, F. Norden, J. H. Plasberg, R. Vafin, S. van de Par, D. Virette, and O. Wubbolt, "Adaptive RD optimized hybrid sound coding," *J. Audio Eng. Society*, vol. 56, no. 10, pp. 787–809, 2008.

[111] 3GPP TS 26.190 - Speech codec speech processing functions; Adaptive Multi-Rate - Wideband (AMR-WB) speech codec; transcoding functions.

[112] J. P. Woodard and L. Hanzo, "Improvements to the analysis-by-synthesis loop in CELP codecs," *IEE Conference on Radio Receivers and Associated Systems*, pp. 114–118, 1995.

[113] F. F. Tzeng, "Near-optimum linear predictive coding,", *Global Telecommunications Conference*, pp. 962–966, 1990.

[114] H. Sanneck, A. Stenger, K. B. Younes, and B. Girod, "A new technique for audio packet loss concealment," in *Proc. Global Telecommunications Conf.*, pp. 48–52, 1996.

[115] K. Clüver and P. Noll, "Reconstruction of missing speech frames using sub-band excitation," in *Proc. IEEE-SP Int. Symp. Time-Frequency and Time-Scale Analysis*, pp. 277–280, 1996.

[116] E. Gündüzhan and K. Momtahan, "A linear prediction based packet loss concealment algorithm for PCM coded speech," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 6, pp. 778–785, 2001.

[117] M. Elsabrouty, M. Bouchard, and T. Aboulnasr, "A new hybrid longterm and short-term prediction algorithm for packet loss erasure over IP-networks," in *Proc. 7th Int. Symp. Signal Processing and Its Applications*, vol. 1, pp. 361–364, 2003.

[118] J. C. D. Martin, T. Unno, and V. Viswanathan, "Improved frame erasure concealment for CELP-based coders," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 1483–1486, 2000.

[119] J. Wang and J. D. Gibson, "Parameter interpolation to enhance the frame erasure robustness of CELP coders in packet networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, pp. 745–748, 2001.

[120] C. A. Rodbro, M. G. Christensen, S. V. Andersen, and S. H. Jensen, "Compressed domain packet loss concealment of sinusoidally coded speech," in*Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, pp. 104–107, 2003.

[121] C. A. Rodbro, M. G. Christensen, S. V. Andersen, and S. H. Jensen, "Hidden Markov model-based packet loss concealment for voice over IP," *IEEE Trans. on Speech, Audio and Language Processing*, vol. 14, no. 5, pp. 1609–1623, 2006.

[122] S. V. Andersen, et al., "iLBC - A linear predictive coder with robustness to packet loss", in *Proc. IEEE Workshop on Speech Coding*, pp. 23–25, 2002.

Paper A

Sparse Linear Predictors for Speech Processing

D. Giacobello, M. G. Christensen, J. Dahl,
S. H. Jensen, and M. Moonen

This paper has been published in
Proceedings of the 9th Annual Conference of the International Speech Communication Association (INTERSPEECH),
pp. 1353–1356, 2008.

The layout has been revised.

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.

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 intro-

duced 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 disadvantages of these.

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:

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

where $\{a_k\}$ are the prediction coefficients and $e(n)$ is the excitation. We will see that the different predictors considered 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:

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

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 variables), 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 (A.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 (A.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 (A.2) can have two meanings. The first one, it is somehow related to the prior knowledge we have of the coefficients vector $\mathbf{a}$, therefore (A.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]:

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

The second meaning that γ holds can be understood by the following analogy. If in (A.2) we let $k = 0$ and assume that the number of bits associated with the quantization of the prediction coefficients $\mathbf{a}$ is 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 (A.2). By setting $p = 1$ and $\gamma = 0$ we obtain the following optimization problem:

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

The use of a least absolute value error criterion has already been proven to give interesting results in linear prediction of speech signals [13]. 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 [14]:

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

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 the statistical character of the frequency errors is not clear. 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 voiced 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 (A.2) is to minimize the 2-norm of the residual while keeping the coefficient vector $\mathbf{a}$ sparse:

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

This formulation is relevant because a direct minimization of (A.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:

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

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 (A.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 [11, 12]. 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 (A.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). As to 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 [16], as the problem in (A.4).

4 Numerical Experiments

The results of the approach shown in (A.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 A.1. 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 Figure A.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 A.2, is that with the 1-norm minimization we obtain a smoother filter.

In Figure A.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

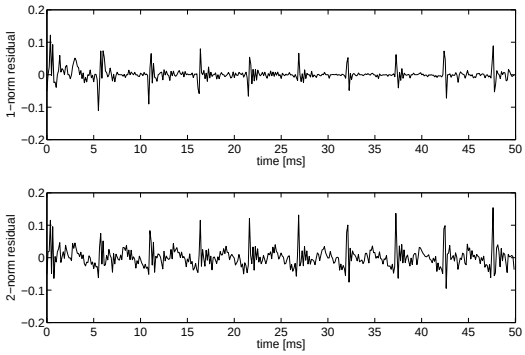

Fig. A.1: *Residuals for 1-norm and 2-norm minimization.*

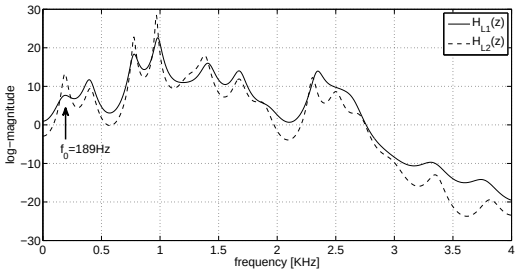

Fig. A.2: *Frequency response of the filters obtained with 1-norm and 2-norm minimization.*

were able to see that using $0.1 \leq \gamma \leq 1$ in (A.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 (A.7) obtained in cascade, in other words in our one step approach we obtained:

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

5 Discussion

Denoël and Solvay [13] 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 an 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

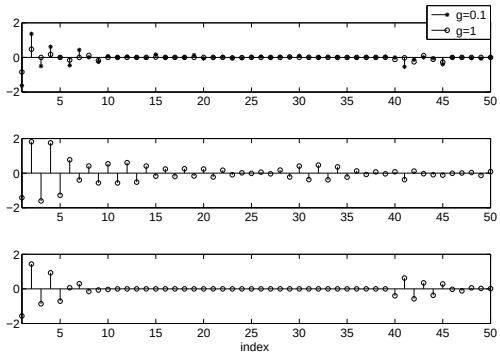

Fig. A.3: *Comparison of the prediction coefficients (excluding the 0^{th}-order) obtained with our algorithm (top), with usual LP (order 50) and with the convolution of the short-term and long-term coefficients vectors.*

all the possible multiple solutions will still be optimal [15]; 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, not being intrinsically stable like LP analysis with the autocorrelation method. This drawback was mitigated by choosing $N_1 = 1$ and $N_2 = N + K$ in (A.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 [13] an interesting method was introduced for both having an intrinsically stable solution as well as keeping the computational cost down using (A.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 A.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 an overdetermined system of equations in the 1-norm us-

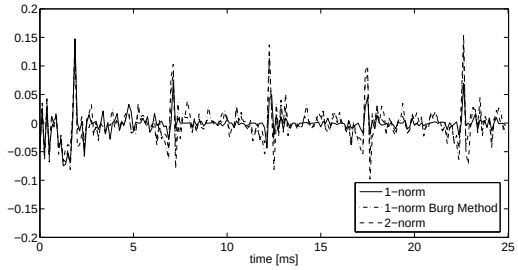

Fig. A.4: *Comparison of the residuals obtained with the method used in the paper (continuous), the Burg method based on the 1-norm (dash-dotted) and the usual LP (dashed).*

ing a modern interior point algorithm [16] showed to be comparable to solving around 10-15 least square problems; 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 [17]) and quantization in general, will be highly simplified by the characteristics of the output. It is also important to notice that the residual signal will already be available at the end of the computation and doesn't have to be calculated.

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

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

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 equal 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 these kinds 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 will be subjected 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.

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] 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.

[12] 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.

[13] 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.

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

[15] 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.

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

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

Paper B

Joint Estimation of Short-Term and Long-Term Predictors in Speech Coders

D. Giacobello, M. G. Christensen, J. Dahl,
S. H. Jensen, and M. Moonen

This paper has been published in
Proceedings of the 34th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),
pp. 4109–4112, 2009.

The layout has been revised.

Abstract

In low bit-rate coders, the near-sample and far-sample redundancies of the speech signal are usually removed by a cascade of a short-term and a long-term linear predictor. These two predictors are usually found in a sequential and therefore suboptimal approach. In this paper we propose an analysis model that jointly finds the two predictors by adding a regularization term in the minimization process to impose sparsity constraints on a high order predictor. The result is a linear predictor that can be easily factorized into the short-term and long-term predictors. This estimation method is then incorporated into an Algebraic Code Excited Linear Prediction scheme and shows to have a better performance than traditional cascade methods and other joint optimization methods, offering lower distortion and higher perceptual speech quality.

1 Introduction

Traditionally, low bit-rate speech coders involve short-term linear prediction (LP) in order to reduce the highly redundant speech signal into a sequence of i.i.d. samples that is easier to quantize. The prediction coefficients are found by minimizing the 2-norm of the prediction error signal (difference between original and predicted signal) [1]; this corresponds to finding the prediction coefficients in a maximum likelihood sense by fitting the error signal into a white Gaussian model. Although this approach is used in almost all commercial speech coder, the theoretical basis is fundamentally wrong as this analysis is optimal only if the input to the AR synthesis model is indeed spectrally white and Gaussian [1]: this is hardly the case for voiced speech and a large set of unvoiced speech sounds. In order to counter this model mismatch, the general approach is to add a long-term predictor in the whitening process: the short-term predictor will first remove the redundancies due to the formants while the long-term predictor will subsequently remove the redundancies due to the presence of a pitch excitation. This scheme is inherently suboptimal for the short-term analysis that will necessarily be biased by the presence of the pitch excitation. The suboptimality of the first short-term prediction step will subsequently corrupt the long-term analysis: the minimum variance residual will not retain the structure of the original excitation but reflect something that has been attenuated and distorted making the analysis more difficult. The most significant works that have pointed out the sub-optimality of the sequential approach were [2] and, more recently [3]. In [2], information about the intermediate short-term residual is included in a new minimization framework that determines jointly the formants and pitch predictors. In [3] a correction factor based on a previous pitch excitation is included in the short-term error minimization. Our main objection to these two methods is that they

do not take into consideration the statistical properties of the analyzed signal as well as how the cascade of the two predictors influences their own coefficients.

The objective of this paper is to define a new one-step minimization framework corresponding to a new way of determining a prediction vector that can then be used to find jointly a non-biased short-term predictor and a more accurate pitch predictor, this also results in a residual error that is spectrally whiter and therefore easier to quantize. This is done by increasing the prediction order and by imposing in the 2-norm minimization of the prediction error signal a penalty term in order to keep the predictor sparse. This sparse predictor can then easily be factorized into the short-term and long-term predictor. The former will not be biased by the presence of a pitch excitation because this is already taken into account by the predictor while the latter will have a higher accuracy than those found through traditional methods. The residual is highly uncorrelated and with very few outliers. Thus, the novelty introduced in this paper is a minimization framework that better matches the statistical characteristics of the speech in order to define, in a latter stage, a more efficient quantization scheme.

The paper is organized as follow. A prologue will be given in Section 2 that illustrates the general formulation for linear predictors employed in speech coders. Section 3 and Section 4 will be dedicated to introducing the mathematical framework in which the joint estimator is developed and how this is formulated. In Section 5 we will show and discuss the performances of our estimator in an Algebraic Code Excited Linear Prediction (ACELP) scheme.

2 General Formulation for Linear Predictors

The general approach in low bit-rate predictive coding is to employ a cascade of a short-term linear predictor $F(z)$ and a long-term linear predictor $P(z)$ in order to remove respectively near-sample redundancies, due to the presence of formants, and distant-sample redundancies, due to the presence of a pitch excitation in voiced speech. The general form of the short-term linear predictor is:

$$F(z) = 1 - \sum_{k=1}^{N_f} f_k z^{-k}. \tag{B.1}$$

The coefficient vector $\mathbf{f} = \{f_k\}$ is determined by minimizing the norm of the prediction error signal:

$$\min_{\mathbf{f}} \|\mathbf{e}\|_p^p = \min_{\mathbf{f}} \|\mathbf{x} - \mathbf{X}\mathbf{f}\|_p^p \tag{B.2}$$

where

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

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, for $p = 2$, setting $N_1 = 1$ and $N_2 = N + N_f$ will lead to the autocorrelation method equivalent to solving the Yule-Walker equations; setting $N_1 = N_f + 1$ and $N_2 = N$ leads to the covariance method [4]. The order of the short-term predictor N_f is usually chosen to be between 8 and 16 and the frame length N between 5 to 20 ms (40 to 160 samples at 8 kHz).

The long-term predictor works in a similar way on the residual of the short-term analysis but using a larger number of data samples ($2N$ to $4N$) in order to find values of the pitch lags that are higher than the length of the short-term window and to better spot long-term redundancies. The pitch predictor has a small number of taps N_p (usually 1 to 3) and the corresponding delays associated are usually clustered around a value which corresponds to the estimated pitch period T_p, the general form is:

$$P(z) = 1 - \sum_{k=1}^{N_p} g_k z^{-(T_p + k)}. \tag{B.3}$$

The parameters $\{g_k\}$ and T_p are determined by minimizing the norm of the residual error signal after the two predictors, just like in the short-term prediction. $P(z)$ often has only one tap and the analysis is done by finding a first *open-loop* estimation of the long-term parameters and successively a *closed-loop* estimation where this is refined and finalized.

The final step is to encode the residual error signal after the two predictors that is hoped to be white and Gaussian. The encoding of the residual signal uses very few bits: in ACELP coders usually the residual is encoded with only 20-30% of non-zeros samples with constrained values of ± 1 and a gain $g_{ac}(n)$ [5].

3 Formulation of the Joint Estimator

The cascade of the predictors in (B.1) and (B.3) corresponds the multiplication in the z-domain of the two transfer functions:

$$\begin{aligned} A(z) = F(z)P(z) &= 1 - \sum_{k=1}^{K} a_k z^{-k} \\ &= (1 - \sum_{k=1}^{N_f} f_k z^{-k})(1 - \sum_{k=1}^{N_p} g_k z^{-(T_p+k)}). \end{aligned} \tag{B.4}$$

The resulting coefficients vector $\mathbf{a} = \{a_k\}$ of the high order polynomial $A(z)$ will therefore be highly sparse. We will then take this sparsity into account in a minimization process similar to (B.2) by adding a regularization term that imposes sparsity on the coefficient vector:

$$\min_{\mathbf{a}} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_2^2 + \gamma \|\mathbf{a}\|_0, \tag{B.5}$$

where $\|\cdot\|_0$ represents the so-called 0-norm, i.e. the cardinality of the vector. A relaxation of this non-convex problem is done by approximating the 0-norm with the more tractable 1-norm [6]:

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

Note that $\mathbf{X}$ has now been redefined as:

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

where $K \geq N_f + N_p$.

The optimization problem in (B.6) can be posed as a quadratic programming problem and can also be solved in time equivalent to solving a small number of 2-norm problems (like the one in (B.2)) using an interior-point algorithm [7]. The left term is strongly convex, sufficient condition for the uniqueness of the solution [7] and also the corresponding polynomial $A(z)$ is minimum phase when the choice of windowing is done as the autocorrelation method (see Section 2).

If we consider the problem in (B.6) from a Bayesian point of view, we notice that this may be interpreted as the *maximum a posteriori* (MAP) approach for finding $\{a_k\}$ under the assumption that the coefficients vector is an i.i.d. Laplacian set of variables and the error is an i.i.d. Gaussian set of variables:

$$\begin{aligned} \mathbf{a}_{MAP} &= \arg\max_{\mathbf{a}} f(\mathbf{x}|\mathbf{a}) g(\mathbf{a}) \\ &= \arg\max_{\mathbf{a}} \{\exp(-\|\mathbf{x} - \mathbf{X}\mathbf{a}\|_2^2) \exp(-\gamma \|\mathbf{a}\|_1)\}, \end{aligned} \tag{B.7}$$

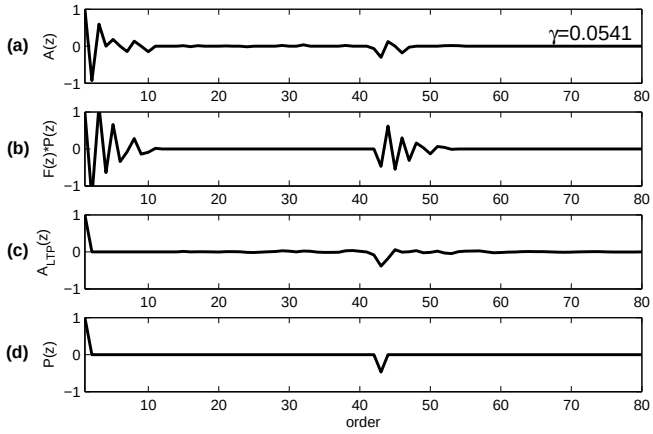

Fig. B.1: (a) *and* **(b)** *show a comparison between the polynomial obtained with regularized minimization $A(z)$ and multiplication of the two predictors $F(z)P(z)$ obtained in cascade;* **(c)** *and* **(d)** *a comparison of the two long-term predictors $A_{LTP}(z)$ and $P(z)$.*

which can be considered to be true observing the coefficients of the polynomial in (B.4). The regularization term γ is then intimately 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, considering (B.6) as an approximation of (B.5). The problem of finding γ that offers the best fitting of the model in (B.6) will be addressed in the next section.

Once the solution of (B.6) has been found, corresponding to the estimated version of the coefficients of $A(z)$ in (B.4), the first N_{stp} coefficients are used as the estimated coefficients of the short-term predictor $A_{stp}(z)$. Then the polynomial $A_{LTP}(z)$ is created by taking the quotient of the division between $A(z)$ by $A_{stp}(z)$. In other words:

$$A(z) = A_{LTP}(z)A_{stp}(z) + R(z); \tag{B.8}$$

where the deconvolution residual $R(z)$ can be considered negligible. Once we have $A_{LTP}(z)$ we can find the pitch gain and delay by taking the minimum value and its position in the corresponding coefficients vector:

$$\begin{aligned} g_p &= \min\{a_{LTP}\}, \\ T_p &= \arg\min\{a_{LTP}\}. \end{aligned} \tag{B.9}$$

where $\{a_{LTP}\}$ are the coefficients of $A_{LTP}(z)$. An example is shown in Figure B.1.

One of the main drawbacks is that even though the polynomial corresponding to the solution of (B.6) is intrinsically stable, by selecting the first N_{stp} coefficients we can risk having the roots of the corresponding short-term prediction polynomial outside the unit circle. This problem is not easy to solve and a deeper analysis has to be done. However, we have observed that if the choice of γ is accurate, the coefficients of the short-term polynomial $A_{stp}(z)$ will usually occupy the first 8 to 16 positions of the high order polynomial $A(z)$ and their absolute value usually decays rapidly. We can reasonably assume that taking the first $N_{stp} \geq 10$ coefficients and ignoring the rest $A_{stp}(z)$ will still be a stable filter. Our intuitive analysis is corroborated by the results obtained: less than 0.01% of short-term filters where unstable in a large set of frames analyzed. As for the long-term predictor, if we choose a one tap filter, having $g_p < 1$ guarantees stabilty; an event in which $g_p \geq 1$ has not been observed in our analysis. It is important to notice that even if a pitch periodicity is not present, the algorithm will still find a pitch gain and delay. The delay values are usually in the same range as the estimates in case of pitch presence, while the pitch gain usually is small ($g_p < 0.01$) not creating any artifacts in the reconstructed signal.

An interesting aspect of this algorithm is that the number of taps is highly customizable. For example, we can choose fixed orders for both predictors or we can adjust them iterating over several values in an analysis-by-synthesis scheme without adding too much complexity to the architecture of the coder, considering that the order of the system of equations in (B.6) is fixed and we are just manipulating the resulting prediction coefficients vector $\{a_k\}$.

4 Selection of the Regularization Term

In previous works on Tikhonov regularized minimization, notably [8], the L-curve has been used in order to examine which value of the regularization parameter γ offers the best trade-off between the variance of the residual and the variance of the solution vector. In our case, we will just substitute the variance of the solution vector with the sum of absolute values. This is done by means of plotting $\|\mathbf{x} - \mathbf{X}\mathbf{a}_\gamma\|_2$ versus $\|\mathbf{a}_\gamma\|_1$ for several values of γ, more precisely for $0 < \gamma < \|\mathbf{X}^T\mathbf{x}\|_\infty$ (where $\|\cdot\|_\infty = \|\cdot\|_1^*$ denotes the dual norm) the solution of (B.6) is a piecewise linear function of γ. It is clear that for values of γ that are too close to the bounds the optimal solution will be useless. In particular, for $\gamma = 0$ we will find a high order polynomial that cannot be easily factorized and for $\gamma \geq \|\mathbf{X}^T\mathbf{x}\|_\infty$ the coefficients $\{a_k\}$ will be all zeros. The L-curve is monotonically decreasing and we can easily find the “corner" that characterizes the L-curve [8] in which the best trade-off can be found. Analyzing about 100.000 frames of speech coming from speakers with different characteristics (gender, age, pitch, regional accent), we have found that the interval of values of γ in which (B.6)

offers the best performances in terms of mere optimization is $0.02 \leq \gamma \leq 0.2$. We will concentrate further analysis, based on the magnitude of the difference between the encoded-decoded signal and the original signal, in this range.

We investigate three approaches, one with γ chosen to be constant, one with γ adaptively chosen based on the statistics of the signal and one with γ found in an optimal sense:

- **constant** γ
  The regularization parameter value that on average gave the best results was $\gamma = 0.0631$. This is the mean of the set of optimal γ's found for each frame.

- **adaptive** γ
  The probability density function of γ shows to have a high variance due to the change in statistics of the analyzed frames of speech. Studying the behavior of the optimal γ we have seen that this is strictly related to how "voiced" the speech is in the analyzed frame, therefore it is intimately related to the pitch gain g_p. By observing the data of the values of the optimal γ over g_p at the n^{th} frame, we have found this approximate relation:
  $$\gamma(n) = -0.18 g_p^2(n) + 0.2. \tag{B.10}$$
  Considering the slow change in value of the pitch gain from a frame to another, starting with $\gamma(n = 0) = 0.0631$, we can update the value of γ using (B.10). A similar relation was used in another regularized linear prediction scheme [9].

- **optimal** γ
  An alternative approach is also investigated where γ is tuned for every frame analyzed in order to obtain the best result. This part of the process is based on the magnitude of the difference between the encoded-decoded signal and the original signal.

5 Validation

5.1 Experimental Setup

In order to obtain comparable results, the regularized method are also implemented in an ACELP scheme, the order of the optimization scheme in (B.6) is $K = 110$ and the frame length is $N = 160$ (20 ms). The order of the short-term and long-term predictors are respectively $N_{stp} = 12$ and $N_{LTP} = 1$, obtained with the procedure of Section 3. The choice of $K = 110$ means that we can cover accurately pitch delays in the interval $[N_{stp} + 1, K - N_{stp} - 1]$ or equivalently

Table B.1: Improvements over conventional ACELP $\mathbf{A_c}$ in the decoded speech signal in terms of reduction of log magnitude distortion ($\mathbf{\Delta}$DIST) and Mean Opinion Score ($\mathbf{\Delta}$MOS). A 95% confidence intervals is given for each value.

METHOD	$\mathbf{\Delta}$DIST	$\mathbf{\Delta}$MOS
$\mathbf{R_o}$	2.05±0.06 dB	0.11±0.00
$\mathbf{R_a}$	1.65±0.11 dB	0.07±0.00
$\mathbf{R_c}$	1.04±0.27 dB	0.03±0.03
$\mathbf{A_j}$	0.32±0.13 dB	0.00±0.02

pitch frequency in the interval $[82Hz, 571Hz]$. The prediction residual vector is encoded according to [5] using 40 non-zero samples constrained with ± 1 values and a gain. In the classical and optimized ACELP scheme, the order of the short-term and long-term analysis are the same ($N_f = 12$ and $N_p = 1$). The coefficients of the short-term filter are found using the autocorrelation method on a subframe basis of 80 samples. The pitch delay and gain are found on the residual error signal according to traditional ACELP encoding [5]. The final residual error signal is also encoded according to [5] but on the subsamples frame basis with 20 non-zero samples and a gain that is averaged with the next one. In order to obtain the same number of parameters for both regularized and traditional ACELP, the values obtained with regularized ACELP are being interpolated (the short-term filter interpolation is done in the LSF domain [5]), so that for each n-th subframe of each method, the transfer function is:

$$H_n(z) = \frac{g_{ac}(n)}{\left(1 - g_p(n) z^{-T_p(n)}\right)\left(1 - \sum_{k=1}^{12} a_k(n) z^{-k}\right)}, \tag{B.11}$$

and the excitation is a 80 samples vector with 20 non-zero as seen above. It should be noted that the interpolation can be performed in the decoder with an important decrease in the number of parameters that have to be transmitted.

5.2 Results

For each method, the signals coming out of the encoding-decoding scheme are compared to the original speech. The results have shown that the regularized methods offer a higher accuracy compared to traditional ACELP as shown in table B.1, both in reducing objective as well as subjective distortion using PESQ evaluation [10]. The performances have shown what could have been reasonably assumed in the preliminary studies. $\mathbf{R_o}$ clearly shows the highest performances having the minimization process tuned to the optimal value of γ. $\mathbf{R_a}$, by taking

into consideration the statistics of the signal, performs at a comparable level to the optimal procedure confirming the good adaptive criterion used in (B.10). $\mathbf{R_c}$ has the drawback of performing poorly when the statistics of the analyzed frame fail to fit into the fixed minimization framework. The jointly optimized method $\mathbf{A_j}$ gives in general higher performances compared to $\mathbf{A_c}$ but the method does not perform well in the unvoiced case where the correction term used in the autocorrelation method has been observed to perturb the minimization process.

There are two main reasons for the increase in accuracy in our methods. First, the spectrally white residual coming out of the optimization process in (B.6) that shows fewer outliers and therefore does not bias the search of an algebraic codeword as much as the traditional ACELP does. Also, the search of the pitch parameters done with the open-loop estimation on the autocorrelation can fail due to the presence of multiples of the pitch delay, this does not happen in our scheme that outperforms the traditional open-loop and closed-loop procedure for pitch estimation. Furthermore, we have observed that the sensitivity of the short-term prediction vectors in our method is generally lower than with traditional LP. This is due to the lower emphasis on peaks that this kind of analysis makes by intrinsically taking into consideration that the signal has outliers due to the pitch excitation. In the traditional short-term linear predictive analysis (B.1) this is not taken into consideration and the minimum-variance approach in finding the residual causes the polynomial to have zeros very close to the unit circle in order to try to cancel the pitch excitation: the result is a transfer function that suffers greatly from this bias and presents a spikier frequency response. This does not happen in our approach. Thus, we have found another meaning for the regularization term γ as related to the bandwidth expansion that is usually operated on the LP filter [9].

6 Conclusion

The analysis method presented in this paper has shown to have attractive performances for the coding of speech signals offering both higher accuracy and lower number of parameters needed. This was done by presenting a new formulation for the minimization process involved in the linear prediction that offers a better statistical fitting for the model of speech making coding more straightforward and accurate.

References

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

[2] 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.

[3] H. Zarrinkoub and P. Mermelstein, "Joint Optimization of Short-Term and Long-Term Predictors in CELP Speech Coders", *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, pp. 157–160, 2003.

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

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

[6] D. Giacobello, M. G. Christensen, J. Dahl, S. H. Jensen and M. Moonen, "Sparse Linear Predictors for Speech Processing", *Proc. INTERSPEECH*, 2008.

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

[8] P. C. Hansen, "Analysis of Discrete Ill-Posed Problems by Means of the L-Curve", *SIAM Review*, vol. 34, no. 4, pp. 561–580, December 1992.

[9] L. A. Ekman, W. B. Kleijn and M. N. Murthi, "Regularized Linear Prediction of Speech", *IEEE Trans. Audio, Speech, Language Processing*, vol. 16, no. 1, pp. 65–73, 2008.

[10] ITU-T Recommendation P.862, "Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for End-To-End Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs", February 2001.

[11] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett and N. Dahlgren, "DARPA-TIMIT acoustic-phonetic continuous speech corpus", *Technical Report NISTIR*, no. 4930, 1993.

Paper C

Speech Coding Based on Sparse Linear Prediction

D. Giacobello, M. G. Christensen, M. N. Murthi,
S. H. Jensen, and M. Moonen

This paper has been published in
Proceedings of the 18th European Signal Processing Conference (EUSIPCO),
pp. 2524–2528, 2007.

The layout has been revised.

Abstract

This paper describes a novel speech coding concept created by introducing sparsity constraints in a linear prediction scheme both on the residual and on the prediction vector. The residual is efficiently encoded using well known multi-pulse excitation procedures due to its sparsity. A robust statistical method for the joint estimation of the short-term and long-term predictors is also provided by exploiting the sparse characteristics of the predictor. Thus, the main purpose of this work is showing that better statistical modeling in the context of speech analysis creates an output that offers better coding properties. The proposed estimation method leads to a convex optimization problem, which can be solved efficiently using interior-point methods. Its simplicity makes it an attractive alternative to common speech coders based on minimum variance linear prediction.

1 Introduction

Linear prediction (LP) is an integral part of many modern speech coding systems and is commonly used to estimate the autoregressive (AR) filter parameters describing the spectral envelope of a segment of speech. Typically, the prediction coefficients are found such that the 2-norm of the difference between the observed signal and the predicted signal is minimized [1]. However, the minimization criterion has been shown to be not optimal in many cases. For example, in voiced speech, when the excitation is not Gaussian, the estimation of the short-term spectrum is contaminated by the spectral fine structure due to the presence of a pitch excitation. In this case, the usual approach is to find coefficients for the short-term and long-term signal correlation in two different steps leading to inherently suboptimal solutions. Furthermore, the 2-norm minimization shapes the residual into variables that exhibit Gaussian-like characteristics; however, in order to encode the residual efficiently, usually only few non-zero pulses are used. We can then reasonably assume that the ideal 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. generates the sparsest residual.

In this paper, we present a method for estimating jointly the short-term and long-term predictors that results in a sparse residual. With this, we transcend the well known problems related to traditional LP based coding discussed above. The novelty introduced is then to exploit the sparse characteristics imposed by the new linear predictive scheme on the predictor and on the residual in order to define, in the latter stage, a more efficient quantization. The strength of our method is seen when these characteristics are used to realize a low bit rate coder that keeps the perceptual quality at high levels.

The paper is organized as follow. We first outline the mathematical for-

mulations of the proposed algorithms. The core of the paper is dedicated to introducing the speech coding procedure and showing the performance results obtained with this technique. Then we will discuss and illustrate advantages and disadvantages of this method before concluding on our work.

2 Sparse Linear Prediction

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

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

Where $\{a_k\}$ are the prediction coefficients and $e(n)$ is the excitation of the corresponding AR filter. 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 [2]. The prediction error 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:

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

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 start and end 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 when the excitation is a sequence of i.i.d. Gaussian random variables), setting $N_1 = 1$ and $N_2 = N + K$ will lead to the autocorrelation method equivalent to solving the Yule-Walker equations, while setting $N_1 = K + 1$ and $N_2 = N$ leads us to the covariance method [3].

The question then is how to choose p, k and γ and how to solve the corresponding minimization problem, depending on the kind of applications we want to implement. In finding a sparse signal representation, there is the somewhat subtle problem of how to measure sparseness. Sparseness is often measured as the cardinality, corresponding to the so-called 0-norm $\|\cdot\|_0$. Therefore, using $p = 0$ in (C.2) means that we would like to minimize the number of non-zero samples

in the error vector. 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 $\|\cdot\|_1$ widely used as a linear programming relaxation of this problem [4]. When $p = 1$ and $k = 1$, our optimization problem then becomes:

$$\min_{\mathbf{a}} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_1 + \gamma\|\mathbf{a}\|_1. \tag{C.3}$$

This optimization problem can be posed as a linear programming problem and can be solved using an interior-point algorithm [2]. The introduction of the regularization parameter γ in (C.2) is intimately related to the *a priori* knowledge that we have on the coefficient vector $\{a_k\}$ or, in other words, to how sparse $\{a_k\}$ is, considering the 1-norm as an approximation of the 0-norm. Furthermore, from a Bayesian point of view, this may be interpreted as the *maximum a posteriori* (MAP) approach for finding $\{a_k\}$ under the assumption that the coefficient vector and the error vector are both i.i.d. Laplacian sets of variables:

$$\begin{aligned} \mathbf{a_{MAP}} &= \arg\max_{\mathbf{a}} f(\mathbf{x}|\mathbf{a})g(\mathbf{a}) \\ &= \arg\max_{\mathbf{a}} \{\exp(-\|\mathbf{x} - \mathbf{X}\mathbf{a}\|_1)\exp(-\gamma\|\mathbf{a}\|_1)\}. \end{aligned} \tag{C.4}$$

3 Basic Coding Structure

The core of the speech coder is based on the optimization problem (C.3) seen in the previous section. In order to obtain appropriate solutions, we have to choose a proper regularization parameter γ in order to obtain the best statistical model for the analyzed segment of speech. For each segment, once we have chosen γ, we can solve the minimization problem in (C.3). At this point we obtain a high order solution vector (the prediction polynomial) and a residual vector that clearly exhibits sparsity. We will then look at efficient ways to encode these.

3.1 Selection of the Regularization Parameter

The regularization parameter γ plays a fundamental role in finding an appropriate statistical model for the segment of speech that is being analyzed. Previous works based on the regularized minimization problem in (C.2), with $p = 2$ and $k = 2$, suggest that the choice should be done based on an algorithm that locates the "corner" of the L-curve [5], defined as the point of maximum curvature of the L shaped curve obtained by plotting $(\|\mathbf{x} - \mathbf{X}\mathbf{a}_\gamma\|_2, \|\mathbf{a}_\gamma\|_2)$ for several values of γ. This value of γ then offers the best trade-off in the minimization problem (C.3).

In our case we modify this principle by replacing the 2-norm with the 1-norm: the new L-curve $(\|\mathbf{x} - \mathbf{X}\mathbf{a}_\gamma\|_1, \|\mathbf{a}_\gamma\|_1)$ will still be a monotonically decreasing curve and the solution $\mathbf{a}_\gamma$ is a piecewise linear function of γ. We can use the

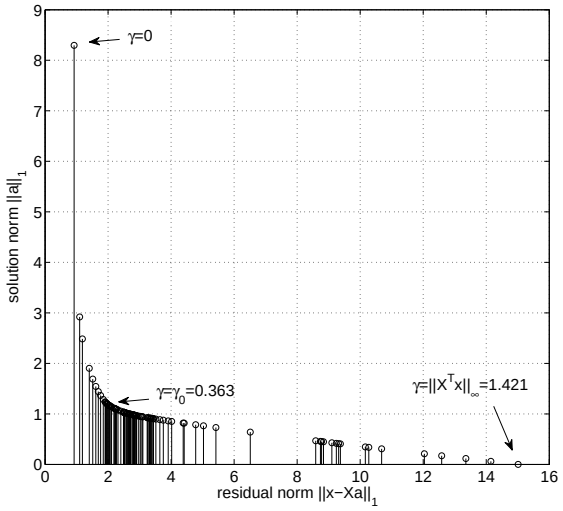

Fig. C.1: *An example of the L-curve ($\|\mathbf{x} - \mathbf{X}\mathbf{a}_\gamma\|_1$,$\|\mathbf{a}_\gamma\|_1$) obtained for a segment of 160 samples of speech (20 ms at 8 kHz); the order is $K = 110$. The lower and upper bounds of γ and their respective solution norm and residual norm are also shown. γ_0 represents the optimal value of the regularization parameter for the current segment found with the algorithm shown in [5].*

same algorithm used in [5] in order to find the point of maximum curvature, that will correspond to the value γ_0. An example of the L-curve so obtained is shown in Figure C.1. Considering the 1-norm as an approximation of the 0-norm, this process may be seen as a trade-off between the sparsity of the residual and the sparsity of the predictor. In particular for $\gamma \geq \|\mathbf{X}^T\mathbf{x}\|_\infty$ (where $\|\cdot\|_\infty = \|\cdot\|_1^*$ denotes the dual norm) the entries of $\mathbf{a}_\gamma$ will all be zeros while for $\gamma = 0$ the predictor sparsity is not controlled and so the number of zeros in the residual will be proportional to the order of the predictor K.

3.2 Factorization of the High Order Predictor

For each segment of speech, the high order predictor $A(z)$, obtained by solving (C.3) using γ_0 as regularization parameter, has mostly zeros as entries due to the sparsity that we have imposed on it. However, the quantization of this predictor may not be trivial due to spurious near-zero components. In this section we will present a robust method to remove these spurious components by creating a new polynomial $A_{os}(z)$ that will then be efficiently factorized into a short-term predictor $A_{stp}(z)$and a long-term predictor $P(z)$.

The removal of the spurious near-zero components in $A(z)$ can be done by applying a model order selection criterion that identifies the useful coefficients in the predictor. Most model order selection criteria for autoregressive (AR) spectral estimation are based on the assumption that the minimization term is the prediction error power of the AR filter. A criterion first introduced by Jenkins and Watts [7] can be generalized to the minimization of the sum of absolute values. The model order selection criterion will then be based on the function:

$$\alpha_k = \frac{1}{N-2k} \sum_{n=k}^{N-1} \left| x(n) + \sum_{i=1}^{k} a_k(n) x(n-i) \right|, \tag{C.5}$$

where the prediction vector $\mathbf{a}$ is obtained by solving the minimization problem in (C.3) for different orders k, using the regularization parameter γ_0 found in the previous step. It has been shown [6] that when solving (C.3) for a segment of voiced speech, the high order polynomial $A(z)$ will be very similar to the convolution of a short-term linear predictor and a long-term linear predictor. According to this, α_k will have a shape that helps us to identify the locations in $A(z)$ of both the short-term predictor and the locations of the coefficients obtained from the convolution between the short-term and long-term predictors. In particular, in traditional AR model selection, α_k will be rapidly decreasing toward a global minimum k_{GMIN} and then monotonically increasing; the order of the AR model is then chosen as k_{GMIN}. This would still be case for segments of signal where long-term redundancies are not present (unvoiced speech). However, in the case when these redundancies are present (voiced speech), the function α_k assumes a very interesting behavior: it will still initially decrease toward a global minimum k_{GMIN} and start increasing again; but then, when the polynomial of order k in (C.5) will start including the positions where the convolution between the short-term and long-term predictors includes important coefficients, α_k will then decrease, increase and decrease again exhibiting also two local minima (k_{LMIN1}, k_{LMIN2}) and two local maxima (k_{LMAX1}, k_{LMAX2}). By extending the polynomial in (C.5), past the positions where the important long-term contribution are, α_k will then increase monotonically toward the global maximum. The first local maximum k_{LMAX1} and the second local minimum k_{LMIN2} then define the location of the convolution of the short-term and the long-term predictor as they are acknowledged by the model order selection curve α_k by making it descend (or, in other words, being useful in the minimization process). Thus, the coefficients with indexes $[k_{LMAX1}+1, \ldots, k_{LMIN2}]$ and the first k_{GMIN} coefficients (corresponding to the location of the short-term predictor) are the only useful non-zero elements in $A(z)$ that we need. An example of the function α_k for voiced speech is shown in Figure C.2 and an example of the two high order polynomials before and after removing the spurious components through the model order selection information ($A(z)$ and $A_{os}(z)$) are shown in Figure C.3.

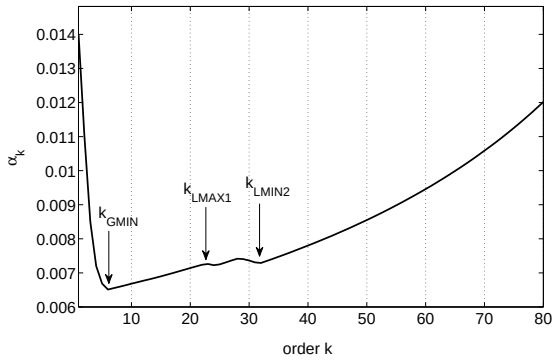

Fig. C.2: *An example of the cost function α_k for a segment of voiced speech. The values used for the order selection $k_{GMIN} = 6$, $k_{LMAX1} = 23$ and $k_{LMIN2} = 32$ are shown.*

The prediction vector $A_{os}(z)$ may now be relatively easy to quantize, having usually few non-zero coefficients. However we can make a further simplification that makes our solution more meaningful by proceeding with the deconvolution of the high-order polynomial. Knowing that the short-term predictor $A_{stp}(z)$ is located in the first k_{GMIN} positions of $A_{os}(z)$:

$$A_{stp}(z) = 1 - \sum_{k=1}^{N_{stp}} a_{os,k} z^{-k}, \tag{C.6}$$

where $N_{stp} = k_{GMIN}$, we can separate $A_{os}(z)$ into its two contributions, short-term $A_{stp}(z)$ and long-term $A_{LTP}(z)$:

$$A_{os}(z) = A_{LTP}(z)A_{stp}(z) + R(z) \approx A_{LTP}(z)A_{stp}(z), \tag{C.7}$$

where we can reasonably assume that the deconvolution residual $R(z)$ is negligible. The resulting polynomial $A_{LTP}(z)$ can then be further reduced into the classical form for a long-term predictor:

$$P(z) = 1 - \sum_{k=0}^{N_p-1} g_k z^{-(T_p+k)}, \tag{C.8}$$

where $T_p = k_{LMAX1} + 1$. The number of taps N_p, i.e., the order of $P(z)$, is chosen by looking at the difference between the magnitude of the frequency response between the true long-term contribution polynomial $A_{LTP}(z)$ and its approximation $P(z)$.

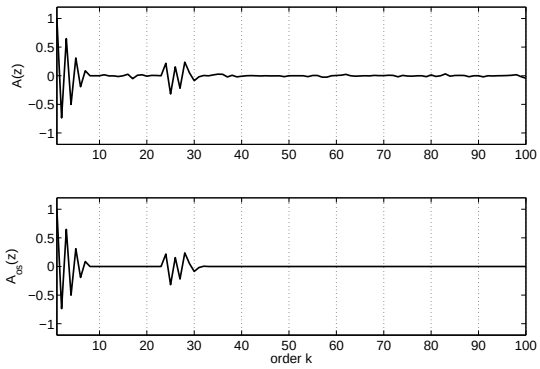

Fig. C.3: *An example of the high order predictor coming out of the minimization process $A(z)$ and its "clean" version $A_{os}(z)$*

3.3 Encoding of the Residual

In traditional LPC coding schemes, the 2-norm shapes the residual such that it exhibits Gaussian-like characteristics. This is not the case in our scheme, where the residual exhibits sparse characteristics (Figure C.4). In early GSM standards, notably in the Multi-Pulse and Regular-Pulse Excitation methods (MPE and RPE) [8], the residual is encoded using only few non-zero pulses. We will then go back to these previous methods as encoding procedures, as they are reasonable approaches to encoding the residual.

In the MPE scheme, an efficient solution is found by determining in an analysis-by-synthesis scheme the locations and amplitudes of the pulses composing the synthetic excitation, one at the time. Finding the location in our case will be much simplified by the sparsity of the residual (Figure C.4). The RPE scheme is based on a similar concept, except that the location of the non-zero samples in the residual is now constrained. In particular, the excitation sequence will be an upsampled version of an optimal vector found using an analysis-by-synthesis criterion. This encoding procedure also allows for a shift of the upsampled sequence [8]. In our work, we will consider this second formulation which will result in a more efficient bit allocation. In the analysis-by-synthesis procedure we will use the polynomial obtained as the multiplication of $A_{stp}(z)$ and $P(z)$.

4 Validation

To validate our method, we will compare it with the GSM 6.10 RPE-LTP Coder [8] and the low rate CELP coder presented in [10]. The comparison

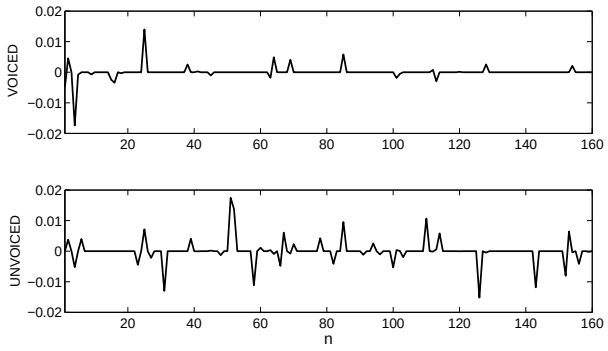

Fig. C.4: *An example of the sparse residual vector for a segment of voiced (above) and unvoiced speech (below).*

with the former method will show that the different ways of estimating the parameters and the residual will lead to a significant decrease in the bit rate with similar perceptual quality. The comparison with the latter method will show the higher perceptual quality obtained with similar bit rate. We have analyzed about one hour of clean speech coming from several different speakers with different characteristics (gender, age, pitch, regional accent) taken from the TIMIT database, re-sampled at 8 kHz. In order to obtain comparable results, the frame length is $N = 160$ (20 ms). The order of the optimization problem in (C.3) is $K = 110$ and the order of the short-term and long-term predictors are chosen according to the method presented in 3.2. For voiced speech we have noted that the order of the short-term predictor is usually between $N_{stp} = 6$ and $N_{stp} = 8$ and the corresponding long-term predictor order is between $N_p = 1$ (usual single lag implementation) and $N_p = 3$, while for unvoiced speech the order is usually between $N_{stp} = 8$ and $N_{stp} = 11$, without long-term information. The choice of $K = 110$ means that we can cover accurately pitch delays in the interval $[N_{stp} + 1, K - N_{stp} - 1]$, including the usual range for the pitch frequency $[70\text{Hz}, 500\text{Hz}]$.

In our method, as well as for the other two coding schemes, the coefficients of the short-term predictor are encoded using their Line Spectral Frequencies (LSFs) representation. The number of bits needed for each LSFs vector it is fixed to 20 bits for a 10 coefficients predictive vector in the RPE and ACELP coders. In our scheme, it will depend on the predictor length from 12 ($N_{stp} = 6$) to 22 ($N_{stp} = 11$) bits per frame. In all three schemes, the method presented in [9] is used; the number of bits chosen is consistent with the transparent coding properties (spectral distortion between quantized and unquantized spectrum less than 1 dB).

Each long-term prediction coefficient is encoded directly with 6, 5, and 4 bits (depending on the position) and the pitch period is encoded with 7 bits. The number of pulses to be used in the regular-pulse encoding of the residual is based on the intrinsic classification between voiced an unvoiced speech performed in the factorization procedure of the high-order polynomial. For voiced speech, the residual will have only very few significant non-zero values, while for unvoiced speech the residual will have a less clear sparse structure (Figure C.4). Therefore we will represent the excitation with 20 samples (pulse spacing $Q = 8$) in the case of unvoiced speech and only 10 samples (pulse spacing $Q = 16$) in the case of voiced speech. A 8-level uniform quantizer is used in both cases. The quantizer normalization factor (the peak magnitude) is encoded with 6 bits per frame; the initial shift is encoded with 3 or 4 bits depending on the number of pulses used in the residual.

The maximum bit rate for voiced speech segments is 87 bits/frame (4300 bits/s) obtained when $N_{stp} = 8$, $N_p = 3$ and we use 10 pulses to code the excitation. The maximum bit rate for unvoiced speech segments is 110 bits/frame (4800 bits/s) obtained when $N_{stp} = 11$ and we use 20 pulses to code the excitation. The choice of the maximum possible number of coefficients is given by the analysis phase. For voiced speech the largest observed value of N_{stp} was 8 and to model the long-term predictor no more than 3 taps have been needed. Similarly, for unvoiced speech the largest observed value of N_{stp} was 11. The average bit rate is around 4600 bits/s. It should be noted that our scheme requires for each frame 1 bit to indicate the voiced/unvoiced decision, 2 bits to indicate the order of the short-term predictor and 2 bits to indicate the order of the long-term predictor.

A perceptual evaluation using PESQ (ITU-T P.862) has been done and the coding scheme has been compared by means of the Mean Opinion Score (MOS) with the other two schemes. The results are shown in Table C.1. The evaluation clearly shows that the large reduction in the bit rate, compared to the RPE, is paid by just a slight decrease in accuracy, demonstrating the robustness of our method. The CELP scheme, that works with a similar bit rate, has a significantly worse perceptual quality.

5 Discussion

In this section we will discuss some of the drawbacks and advantages of the LPC method presented in the paper.

Stability

Stability is important in common linear predictive coding for various reasons, the most important one being its employment in the analysis-by-synthesis schemes, to choose the best approximate excitation, and in the synthesis of the recon-

Table C.1: Comparison in terms of bit rate and Mean Opinion Score (MOS) between our coder based on Sparse LP, the RPE-LTP and the CELP scheme according to [10]. A 95% confidence intervals is given for each value.

Coder	Bit Rate	MOS
Sparse LP	4.6 Kb/s	3.49±0.03
RPE-LTP	12.4 Kb/s	3.59±0.06
CELP	4.7 Kb/s	3.21±0.01

structed speech signal. Our scheme presents a low rate (around 2%) of unstable combined filters $A_{stp}(z)P(z)$ and an important aspect is that the instability in this polynomial is given, except very few exceptions, only by the long-term predictor $P(z)$. This is consistent with traditional coding procedures in which the pitch gain is allowed to be greater than 1 (one tap implementation). It should be noted that $A(z)$, $A_{os}(z)$ and the combined polynomial $A_{stp}(z)P(z)$ exhibit the same instability rate, a further proof of the good criterion employed to factorize the polynomial. Although stability has been considered a fundamental property to be kept in speech coding frameworks, we have noted in our scheme that instability does not affect the performances of our coder (i.e., the output of the system does not "explode"). We have found as a main reason for this is that the roots outside the unit circle are usually only given by the long-term predictor and they are still very close to the unit circle. A proof is that performing a bandwidth expansion, using a fixed value found in the analysis process as low as 0.9965 (about 20 Hz of expansion), would force the number of non-minimum phase combination filters $A_{stp}(z)P(z)$ to zero. The unstable filters are also isolated events that do not create problems in the reconstruction phase. In practice, using a minimum phase $A_{stp}(z)P(z)$ results in slightly higher time-domain distortion than the original composite filter.

Uniqueness

The minimization problem in (C.3) allows for the solution not to be unique. In these rare cases of multiple solutions, due to the convexity of the cost function, we can easily state that the all the possible multiple solutions will still be optimal [2].

Computational costs

Regarding the computational costs, finding the solution of the overdetermined system of equations in (C.3) using a modern interior point algorithm [2] can be shown to be comparable to solving around 20-30 least square problems. However, our advantage is that we have found a one step way to calculate both the short-term and the long-term predictors while the encoding of the residual is facilitated by its sparse characteristics. The process of selecting the regularization parameter γ_0 can also be highly simplified by choosing it in a fixed or adaptive

way based on the properties of the signal as done in other regularized prediction methods [6, 11]. The factorization process can also be done by choosing a fixed set of possible values of N_{stp} and N_p and selecting the ones that creates the best fitting of $A(z)$, skipping the model order selection procedure [6].

Sensitivity of the short-term predictor coefficients

In the experimental analysis, the coefficients of the short-term prediction polynomial $A_{stp}(z)$ obtained with our LP method have shown to have lower sensitivity than the one obtained with usual LPC procedures. This allows one to also have reflection coefficients, Log-Area-Ratio coefficients or Line Spectral Frequencies, with a lower sensitivity as well, therefore allowing more efficient quantization. In particular, we have observed a lower log spectral distortion (LSD) between the estimated short-term AR model obtained with our method $S_1(\omega, \mathbf{a})$ and its corresponding quantized version $\hat{S}_1(\omega, \mathbf{a})$, compared to the one obtained with the 2-norm autocorrelation method $S_2(\omega, \mathbf{a})$ (applying a 60 Hz bandwidth expansion) and its quantized version $\hat{S}_2(\omega, \mathbf{a})$. Another comparison, between a reference spectrum $S_{ref}(\omega)$ and the quantized versions of the two AR models has also demonstrated that our method is generally more efficient in quantization purposes by achieving a lower distortion at lower bit rates. The reference used was found through a cubic spline interpolation between the harmonic peaks of the logarithmic periodogram and used as an approximation of the true vocal tract transfer function [11]. An example of the LSD values obtained for different rates is shown in Figure C.5.

Pitch-independence and shift-independence

Two properties of the method presented in this paper that have stunned us, and will be subject to further investigations, are the pitch-independence of the short-term predictor $A_{stp}(z)$ and the shift-independence of the solution predictor $A(z)$. Our analysis has shown that shifting the frame boundaries by few samples does not change significantly the statistics of the predictor as much as with the traditional linear predictive coding. The pitch-independence has been observed by re-synthesizing segments of speech changing only the pitch value. Analyzing again the new synthetic signal and comparing the new short-term envelopes with the original ones, the new short-term envelopes have not exhibited any significant changes when our method is employed, while dramatic differences have been observed when traditional 2-norm LP analysis is used. Both properties are most likely due to the robustness of the estimation based on the 1-norm to outliers. The shift-independence may be mainly due to the reduced dependence of the solution to all of the values taken into consideration in the minimization process (just like when calculating the median value of an even number of observations). The pitch-independence may be due to the reduced emphasis put on the envelope peaks by the 1-norm LP estimation than the traditional 2-norm LP estimation in the minimization process to reduce the outliers of the pitch excitation. The common LP analysis tries to cancel the pitch harmonics by putting some of

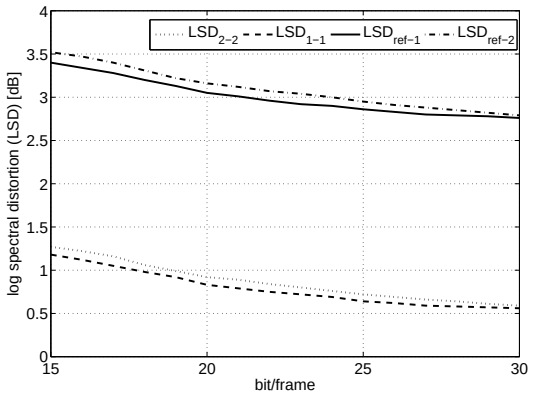

Fig. C.5: *Values of average log spectral distortion (LSD) for voiced speech at different bits per LSFs frame. The figure shows the LSD values between the two AR models (obtained with our scheme ($N_{stp} = 8$) and with 2-norm minimization ($N_{stp} = 10$)) and their quantized version (LSD_{1-1q} vs. LSD_{1-2q}). The total LSD is also shown comparing the quantized AR models with a ground truth reference spectrum (LSD_{ref-1q} vs. LSD_{ref-2q}). In our method the bit rate includes the 2 bits necessary to indicate the model order at the receiver.*

the poles very closed to the unit circle. The 1-norm approach acknowledges the existence of the pitch harmonics, although it does not try to cancel them because its purpose is not to fit the error into a Gaussian-like probability density function and consequently it will let through the pitch excitation outliers. This results in smoother short-term filters that are independent from the underlying pitch excitation in voiced speech. This makes the pitch detection much easier in the case of a conventional analysis based on the short-term residual. In our case, we go even beyond this sequential approach having jointly estimated short-term and long-term predictors. The pitch-tracking properties have been shown to outperform the traditional closed-loop pitch estimation done on the short-term prediction residual. We compared the results of both with a robust reference based on subspace pitch estimation [12]; an example is shown in Figure C.6.

6 Conclusions

In this paper we have introduced a new formulation in the context of speech coding where the concept of sparsity is used in the linear predictive scheme.The sparse residual obtained allows a more compact representation, while the sparse high order predictor engenders joint estimation of short-term and long-term predictors that achieve better spectral matching properties than conventional methods. The short-term predictors obtained are not corrupted by the fine structure

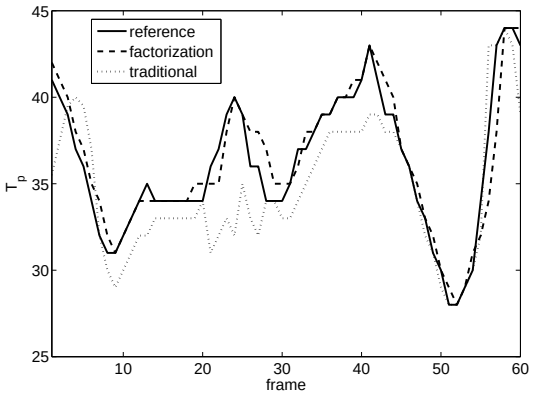

Fig. C.6: *Integer pitch lag (T_p) tracking performances for our method based on the factorization of short-term and long-term predictor compared with traditional close-loop method based on autocorrelation and a reference value based on subspace pitch estimation [12].*

belonging to the pitch excitation and their smoother spectral envelopes are robust to quantization. These envelopes are also represented using lower order AR models compared to traditional LP based coders, thus requiring fewer bits. The long-term predictors and, in particular, the pitch lag estimation are also more accurate. These and other interesting properties, like pitch-independence of the short-term spectral envelopes and shift-independence of the combined envelopes, lead to attractive performance in speech coding.

References

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

[2] S. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge University Press, 2004.

[3] P. Stoica and R. Moses, *Spectral analysis of signals*, Pearson Prentice Hall, 2005.

[4] D. Giacobello, M. G. Christensen, J. Dahl, S. H. Jensen, and M. Moonen, "Sparse linear predictors for speech processing", *Proc. INTERSPEECH*, pp. 1353–1356, 2008.

[5] 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.

[6] 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", to appear in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 2009.

[7] G. M. Jenkins and D. G. Watts, *Spectral analysis and its applications*, Holden-Day, 1968.

[8] P. Kroon, E. F. Deprettere, and R. J. Sluyter, "Regular-pulse excitation - a novel approach to effective and efficient multipulse coding of speech", *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 34, no. 5, pp. 1054–1063, 1986.

[9] A. D. Subramaniam, B. D. Rao, "PDF optimized parametric vector quantization of speech line spectral frequencies", *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 2, 2003.

[10] M. Schroeder and B. Atal, "Code-excited linear prediction(CELP): high-quality speech at very low bit rates", in *Proc. IEEE International Conference Acoustics, Speech, and Signal Processing*, vol. 10, pp. 937–940, 1985.

[11] L. A. Ekman, W. B. Kleijn, and M. N. Murthi, "Regularized linear prediction of speech", *IEEE Trans. Audio, Speech, Language Processing*, vol. 16, no. 1, pp. 65–73, 2008.

[12] M. G. Christensen, A. Jakobsson, and S. H. Jensen, "Joint High-Resolution Fundamental Frequency and Order Estimation", *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 15, no. 5, pp. 1635–1644, 2007.

Paper D

Enhancing Sparsity in Linear Prediction of Speech by Iteratively Reweighted 1-norm Minimization

D. Giacobello, M. G. Christensen, M. N. Murthi,
S. H. Jensen, and M. Moonen

This paper has been published in
Proceedings of the 35th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),
pp. 4650–4653, 2010.

The layout has been revised.

Abstract

Linear prediction of speech based on 1-norm minimization has already proved to be an interesting alternative to 2-norm minimization. In particular, choosing the 1-norm as a convex relaxation of the 0-norm, the corresponding linear prediction model offers a sparser residual better suited for coding applications. In this paper, we propose a new speech modeling technique based on reweighted 1-norm minimization. The purpose of the reweighted scheme is to overcome the mismatch between 0-norm minimization and 1-norm minimization while keeping the problem solvable with convex estimation tools. Experimental results prove the effectiveness of the reweighted 1-norm minimization, offering better coding properties compared to 1-norm minimization.

1 Introduction

In Linear Predictive Coding of speech signals (LPC), the prediction coefficients are typically obtained by minimizing the 2-norm of the residual (the difference between the observed signal and the predicted signal) [1]. The 2-norm minimization shapes the residual into variables that exhibit Gaussian-like characteristics. However, in order to reduce the information content of the residual and to allow for a low bit rate encoding, a sparse approximation of the residual is often used. This conceptual difference between a quasi-white minimum variance residual and its approximated version creates a mismatch that can raise the distortion significantly. In our recent work, we have defined a new predictive framework that provides a tighter coupling between the linear predictive analysis and the residual encoding by looking for a sparse residual rather than a minimum variance one [2, 3]. Early encoding techniques such as Multi-Pulse Excitation (MPE) [4] or Regular-Pulse Excitation (RPE) [5], have shown to be more consistent with this kind of predictive framework unlike, e.g., Code Excited LP (CELP) [6] that uses pseudo-random sequences to encode the residual.

In our previous work we have used the 1-norm as a convex relaxation of the so-called 0-norm, the cardinality of a vector. The 0-norm, and more generally the p-norm with $0 \leq p < 1$, is not a proper norm and its minimization yields a combinatorial problem (NP-hard). We therefore aim to "adjust" the error weighting difference between the 1-norm and the 0-norm keeping the feasibility of the problem in polynomial time. To do so, in this paper we propose a new method for the estimation of the prediction filter based on iteratively reweighted 1-norm minimization [7]. We will see how this method, by enhancing the sparsity of the residual, yields a better and simpler formulation of the coding problem, hence allowing for a general improvement in performance.

The paper is organized as follows. In Section 2 we give the general problem

formulation of sparse linear prediction. In Section 3 we introduce the algorithms used to enhance sparsity in linear predictive coding and in Section 4 we provide a statistical interpretation. In Section 5 and Section 6 we illustrate the effects of the algorithm for analysis and coding of speech. Section 7 concludes the paper.

2 Sparse Linear Prediction

The problem considered in this paper is based on the following Auto-Regressive (AR) speech production model, where a sample of speech $x(n)$ is written as a linear combination of K past samples:

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

where $\{a_k\}$ are the prediction coefficients and $r(n)$ is the driving noise process (commonly referred to as the prediction residual). The speech production model (D.1) in matrix form becomes:

$$\mathbf{x} = \mathbf{X}\mathbf{a} + \mathbf{r} \tag{D.2}$$

where:

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

The prediction coefficient vector $\mathbf{a} \in \mathbb{R}^K$ is found by minimizing the $p-$norm of the residual $\mathbf{r}$ [8]:

$$\hat{\mathbf{a}}, \hat{\mathbf{r}} = \arg\min_{\mathbf{a}} \|\mathbf{r}\|_p^p, \quad \text{s.t.} \quad \mathbf{r} = \mathbf{x} - \mathbf{X}\mathbf{a}; \tag{D.3}$$

where $\|\cdot\|_p$ is the p-norm. The starting and ending points $N_1 = 1$ and $N_2 = N + K$ are chosen assuming that $x(n) = 0$ for $n < 1$ and $n > N$ [9]. Sparsity is often measured as the cardinality, i.e., the so-called 0-norm. Therefore, setting $p = 0$ in (D.3) means that we aim to minimize the number of non-zero samples in the error signal. Unfortunately this corresponds to a combinatorial problem which generally cannot be solved in polynomial time. Instead of the 0-norm, we then use the more tractable 1-norm [2]:

$$\hat{\mathbf{a}}, \hat{\mathbf{r}} = \arg\min_{\mathbf{a}} \|\mathbf{r}\|_1 \quad \text{s.t.} \quad \mathbf{r} = \mathbf{x} - \mathbf{X}\mathbf{a}; \tag{D.4}$$

An interesting alternative problem formulation is obtained when sparsity is also imposed on the predictor:

$$\hat{\mathbf{a}}, \hat{\mathbf{r}} = \arg\min_{\mathbf{a}} \|\mathbf{r}\|_1 + \gamma\|\mathbf{a}\|_1, \quad \text{s.t.} \quad \mathbf{r} = \mathbf{x} - \mathbf{X}\mathbf{a}; \tag{D.5}$$

Algorithm 1 Iteratively Reweighted 1-norm Minimization of the Residual

Inputs: speech segment $\mathbf{x}$
Outputs: predictor $\hat{\mathbf{a}}^i$, residual $\hat{\mathbf{r}}^i$
$i = 0$, initial weights $\mathbf{W}^{i=0} = \mathbf{I}$
while halting criterion false **do**
1. $\hat{\mathbf{a}}^i, \hat{\mathbf{r}}^i \leftarrow \arg\min_{\mathbf{a}} \|\mathbf{W}^i\mathbf{r}\|_1 \quad \text{s.t.} \quad \mathbf{r} = \mathbf{x} - \mathbf{X}\mathbf{a}$
2. $\mathbf{W}^{i+1} \leftarrow \text{diag}\left(\left|\hat{\mathbf{r}}^i\right| + \epsilon\right)^{-1}$
3. $i \leftarrow i + 1$

end while

in this case the sparse structure of the predictor (in this case high order) allows a joint estimation of a short-term and a long-term predictor [3, 10]. This optimization problem can be posed as a linear programming problem and can be solved using an interior-point algorithm [8].

3 Iteratively Reweighted 1-norm Minimization

Our general goal is to determine a linear predictor that yields a sparse residual. As mentioned before, for $0 \leq p < 1$, the problem cannot be solved using convex optimization. To overcome this problem, an iteratively reweighted 1-norm minimization may be used for estimating $\mathbf{a}$ and enhancing the sparsity on $\mathbf{r}$, while keeping the problem solvable with convex tools [7]. The algorithm is shown in Algorithm 1. The parameter $\epsilon > 0$ is used to provide stability when a component of $\hat{\mathbf{r}}$ goes to zero. ϵ does not need to be too small; as empirically demonstrated in [7], it should be in the order of the expected nonzero magnitude of $\mathbf{r}$. It can be shown that $\|\hat{\mathbf{r}}^{i+1}\|_1 \leq \|\hat{\mathbf{r}}^i\|_1$, meaning that this is a descent algorithm [7]. The halting criterion can therefore be chosen as either a maximum number of iterations or as a convergence criterion.

When we impose sparsity both on the residual and on the predictor, as in (D.5), the algorithm is modified as shown in Algorithm 2. As mentioned before, the high order sparse predictor estimated in (D.5) is found to show a structure similar to the convolution between a short-term and a long-term predictor, usually estimated in two different stages. In previous approaches [3, 10], the predictor shows a clear sparse structure but also some spurious components, i.e., small components in the predictor that are irrelevant to our analysis. In [3], we have used a model order selection criterion to locate the spurious quasi-zero components in the predictor which are then put to zero. The reweighted 1-norm minimization seems to be more effective in removing these spurious components, as the new predictor is iteratively re-estimated, rather than just "cleaned up".

Algorithm 2 Iteratively Reweighted 1-norm Minimization of Residual and Predictor

Inputs: speech segment $\mathbf{x}$
Outputs: predictor $\hat{\mathbf{a}}^i$, residual $\hat{\mathbf{r}}^i$
$i = 0$, initial weights $\mathbf{W}^{i=0} = \mathbf{I}$ and $\mathbf{D}^{i=0} = \mathbf{I}$
while halting criterion false **do**
1. $\hat{\mathbf{a}}^i, \hat{\mathbf{r}}^i \leftarrow \arg\min_{\mathbf{a}} \|\mathbf{W}^i\mathbf{r}\|_1 + \gamma\|\mathbf{D}^i\mathbf{a}\|_1$
   s.t. $\mathbf{r} = \mathbf{x} - \mathbf{X}\mathbf{a}$
2. $\mathbf{W}^{i+1} \leftarrow \text{diag}\left(\left|\hat{\mathbf{r}}^i\right| + \epsilon\right)^{-1}$
3. $\mathbf{D}^{i+1} \leftarrow \text{diag}\left(\left|\hat{\mathbf{a}}^i\right| + \epsilon\right)^{-1}$
4. $i \leftarrow i + 1$

end while

4 Statistical Interpretation

The linear prediction solution defined in (D.4) and (D.5) can be seen respectively as the *Maximum Likelihood* (ML) and *Maximum A Priori* (MAP) estimate of an AR process driven by a Laplacian noise sequence $\mathbf{r}$. In the MAP approach, a prior on $\mathbf{a}$ as a Laplacian variable is also imposed. The Laplacian distribution has already been considered to provide a more appropriate fitting for speech [12] than the Gaussian distribution, due to the heavier tails that admit larger errors in the residual. For the case $p \leq 1$, the density functions will have even heavier tails and a sharper slope near zero. In particular, this means that the maximization will encourage small values to become smaller while leaving unchange the larger values. The limit case for $p = 0$ will have an infinitely sharp slope in zero and equally weighted larger slopes. This will force the maximization to include as many zeros as possible as they are infinitely weighted.

The mismatch between the 0-norm and the 1-norm minimization that we are trying to compensate for, can be seen more clearly in Figure D.1, where larger coefficients are penalized more heavily by the 1-norm than small ones. In this sense, the 0-norm can be seen as more "impartial" by penalizing every nonzero coefficient equally. It is clear that if a very small value would be weighted as much as a large value, the minimization process will try to eliminate the smaller ones and enhance the larger ones.

This explains the choice of the weights as the inverse of the magnitude of the residual. In fact, this weighting will balance the dependence on the magnitude of the 1-norm, changing the cost function and moving the problem towards the 0-norm minimization.

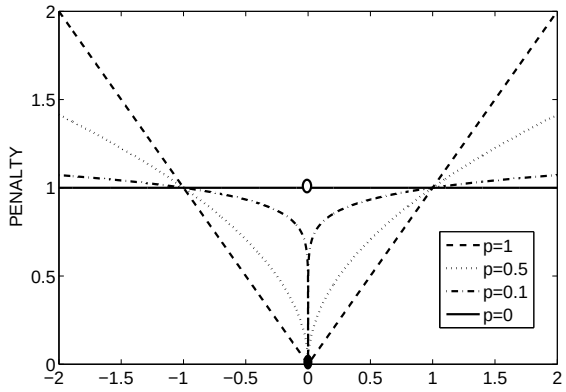

Fig. D.1: *Comparison between cost functions for $p \leq 1$. The 0-norm can be seen as more "democratic" than any other norm by weighting all the nonzero coefficients equally.*

5 Experimental Analysis

To illustrate the effects of the algorithm, we first analyze a segment of stationary voiced speech. The reweighted 1-norm minimization helps to reduce the emphasis on the outliers due to the pitch excitation, as we can see clearly in Figure D.2. The ability of easily spotting the main components in the residual, as we shall see in the next section, have a great impact on coding applications.

An even more interesting case, is the reweighted 1-norm minimization of both residual and predictor. In this case, the use of the high order predictor removes also the long-term redundancies, what is left is almost just an impulse as shown in Figure D.3. This basically means that all the information of the signal is transferred to the predictor which also show a very clear sparse structure, similar to the convolution between the coefficients of short-term and long-term predictors. The examples were obtained analyzing the vowel /a/ uttered by a female speaker using $N = 160$, $f_s = 8$ kHz and order $K = 10$ for Algorithm 1 and $K = 110$ for Algorithm 2. In both cases $\epsilon = 0.01$. The choice of the regularization term γ is given by the L-curve where a trade-off between the sparsity of the residual and the sparsity of the predictor is found [3, 11]. Both algorithms converge rapidly, three to five iteration are sufficient to reach a point where $\|\hat{\mathbf{r}}^{i+1}\|_1 \approx \|\hat{\mathbf{r}}^{i}\|_1$ and, in the joint case, $\|\hat{\mathbf{a}}^{i+1}\|_1 \approx \|\hat{\mathbf{a}}^{i}\|_1$.

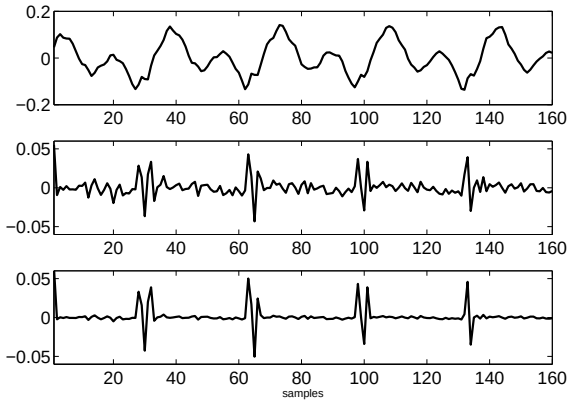

Fig. D.2: *Comparison between true 1-norm 10th order LP residual (middle) and iteratively reweighted 1-norm LP residual (bottom) according to Algorithm 1. The original voiced speech is shown on top. Three iteration where performed, sufficient to reach convergence.*

6 Validation

To validate our method, we have analyzed about one hour of clean speech coming from several different speakers with different characteristics (gender, age, pitch, regional accent) taken from the TIMIT database, re-sampled at 8 kHz. The frame length is $N = 160$ (20 ms). We will consider now the two cases, with the reweighted minimization of the residual and with the reweighted minimization of both residual and predictor. The parameter ϵ, used to avoid division by zero, is chosen to be $\epsilon = 0.01$.

6.1 Reweighted Residual

In order to code the residual sequence when Algorithm 1 is used, after the reweighted scheme we use an Analysis-by-Synthesis to optimize the amplitudes of the $M = 20$ largest pulses (therefore constraining the positions). The order of the predictor is $K = 10$, a long-term predictor is not used for immediacy of the results. Our method (**MPE1r**) is compared with the classic MPE scheme where the linear predictor is found with a 1-norm minimization (**MPE1**), with a 2-norm minimization (**MPE2r**) [4] and using a 2-norm re-weighted minimization (**MPE1r**) [13]. In the reweighted cases, five iterations are done (enough to reach reasonable convergence). The quantization process uses 20 bits to encode the predictor using 10 Line Spectral Frequencies using the procedure in [14], in the case the filter is unstable the poles outside the unit circle are reflected inside of it. A 3 bits uniform quantizer that goes from the lowest to the highest

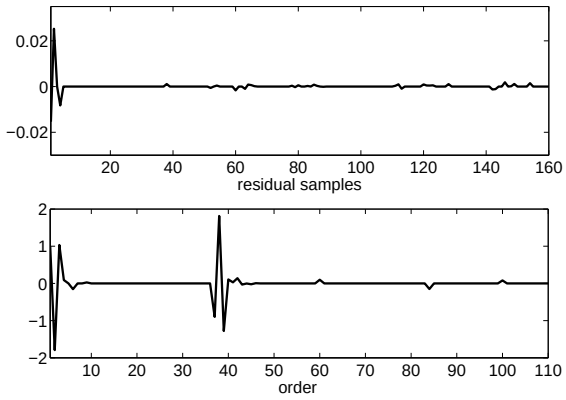

Fig. D.3: *Residual and 110th order linear predictor at the convergence of Algorithm 2 after three iterations. The speech segment analyzed is the same as Figure D.2.*

magnitude of the residual pulses is used to code the residual, 5 bits are used to code the lowest magnitude and 2 bits are used to code the difference between lowest and highest magnitude. The signs are coded with 1 bit per each pulse. We postpone the efficient encoding of the positions to further investigation, for now we just use the information content of the pulse location which is $\log_2 \binom{160}{10}$ bits. This produces a bit rate of 9500 bits/s. The results are shown in Table D.1. We would like to highlight that in **MPE1r** it is not necessary to calculate the positions of the nonzero pulses are located (as it is usually done in MPE coding), we simply exploit the information coming out of the predictive analysis. We are then clearly moving our problem towards a more synergistic way to code a signal.

Table D.1: Comparison between the MPE residual estimation methods in terms of Segmental SNR and Mean Opinion Score (PESQ evaluation). A 95% confidence intervals is given for each value.

METHOD	SSNR	MOS
MPE1r	20.9±1.9	3.24±0.03
MPE1	20.0±3.2	3.20±0.12
MPE2r	19.3±2.9	3.17±0.10
MPE2	18.5±2.1	3.17±0.22

6.2 Reweighted Residual and Predictor

The most interesting case is when both predictor and residual are processed in the reweighted minimization. As shown in our previous work [10], the high order predictor is split into the long-term and short-term component through a simple deconvolution. The short-term predictor $A_{stp}(z)$ will have order $N_{stp} = 10$ and the long term predictor (pitch predictor) $P(z) = 1 - g_p z^{-T_p}$ will have order one. The choice of $K = 110$ in (D.5) means that we can cover accurately pitch delays in the interval $[N_{stp} + 1, K - N_{stp} - 1]$, including the usual range for the pitch frequency $[70\text{Hz}, 500\text{Hz}]$.

In the coding process, we can make a distinction between the voiced case and the unvoiced case. In particular, when the pitch gain g_p is lower than a certain threshold, we will not code the long term informations and we will allocate more pulses for the residual, usually less sparse than the voiced residual. In our experimental analysis we have set the threshold to $TH_{g_p} = 0.05$. $M = 5$ and $M = 10$ pulses are used respectively in the voiced and unvoiced case. Just like we did in Section 6.1, the positions of the M pulses of largest magnitude are used in the Analysis-by-Synthesis to define the only nonzero samples. The quantization procedure is also the same as in Section 6.1, except for the quantization of T_p and g_p for which we use respectively 7 and 6 bits. This produces a bit rate of 5450 bit/s in the voiced case and 4900 bit/s in the unvoiced case, and an approximate average bit rate of 5175 bit/s. We will compare our method (**J11r**) with the scheme without the reweighting (**J11**) presented in Equation (D.5) and the method where the significant coefficients are chosen using a model order selection procedure [3] (**J11os**), we also compared the method with both reweighting and model order selection. In the reweighting cases, only three iteration were needed to reach convergence in all the analyzed frames. The results shown in Table D.2, demonstrate a net improvement over the traditional method (**J11**) and a slight improvement also over (**J11os**), without the costly model order selection procedure. The combinations of both methods (**J11r+os**), shows the best results. This is due to the combination of the reweighting procedure that "concentrates" the nonzero parts in the high order polynomial with the model order selection that "spots" the important ones.

7 Conclusions

In this paper, we have proposed a method to enhance sparsity in linear prediction based on the reweighted 1-norm error minimization. With just few iterations, we were able to move the error minimization criterion toward the 0-norm solution, showing general improvements over conventional 1-norm minimization in coding purposes. Statistical reasons supporting the new criterion have also been provided. A concluding remark would also be that in the cases analyzed, we

Table D.2: Comparison between the coding methods with joint estimation of residual and predictor in terms of Segmental SNR and Mean Opinion Score (PESQ evaluation). A 95% confidence intervals is given for each value.

METHOD	SSNR	MOS
J11r+os	27.9±0.9	3.59±0.02
J11r	25.3±1.3	3.43±0.03
J11os	24.7±1.0	3.40±0.09
J11	23.9±1.9	3.22±0.09

have no prior knowledge of where the residual should be nonzero. This brings the bit allocated to describe the position of few samples to significantly increase the rate. An interesting case, that would subject to further analysis would be to *structure* the reweighting process by imposing where we would like to have the nonzero pulses located. First experiments have shown to be promising and will be subject of our future work.

References

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

[2] D. Giacobello, M. G. Christensen, J. Dahl, S. H. Jensen and M. Moonen, "Sparse Linear Predictors for Speech Processing," *Proc. Interspeech*, 2008.

[3] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen "Speech Coding Based On Sparse Linear Prediction", to appear in *Proc. European Signal Processing Conference*, 2009.

[4] 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.

[5] 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.

[6] M. R. Schroeder and B. S. Atal,"Code-excited linear prediction (CELP): high-quality speech at very low bit rates," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 10, pp. 937Ű940, 1985.

[7] E. J. Candés, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *Journal of Fourier Analysis and Applications*, vol. 14(5-6), pp. 877–905, 2008.

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

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

[10] 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", in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 2009.

[11] 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.

[12] S. Gazor and W. Zhang, "Speech probability distribution", *IEEE Signal Processing Letters*, vol. 10, no. 7, pp. 204–207, 2003.

[13] J. Lansford and R. Yarlagadda, "Adaptive L_p approach to speech coding," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 1, pp. 335–338, 1988.

[14] A. D. Subramaniam, B. D. Rao, "PDF optimized parametric vector quantization of speech line spectral frequencies", *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 2, 2003.

Paper E

Sparse Linear Prediction and Its Applications to Speech Processing

D. Giacobello, M. G. Christensen, M. N. Murthi,
S. H. Jensen, and M. Moonen

This paper has been submitted to
IEEE Transactions on Audio, Speech, and Language Processing,
2010.

The layout has been revised.

Abstract

The aim of this paper is to provide an overview of Sparse Linear Prediction*, a set of speech processing tools created by introducing sparsity constraints into the linear prediction framework. These tools have shown to be effective in several issues related to modeling and coding of speech signals. From a speech analysis perspective, we provide predictors that are accurate in modeling the speech production process and overcome problems related to traditional linear prediction. In particular, the predictors obtained offer a more effective decoupling between the vocal tract transfer function and its underlying excitation, making it a very efficient method for the analysis of voiced speech. From a speech coding perspective, we provide predictors that shape the residual according to the characteristics of the sparse encoding techniques creating more synergistic and straightforward coding strategies. Furthermore, encouraged by the promising application of compressed sensing in signal compression, we investigate its formulation and application to sparse linear predictive coding. The proposed estimators are all solutions to convex optimization problems, which can be solved efficiently and reliably using, e.g., interior-point methods. Extensive experimental results are provided to support the effectiveness of the proposed methods, showing the improvements over traditional linear prediction in both speech analysis and coding.*

1 Introduction

Linear prediction (LP) is widely considered as one of the most prominent ways to model speech signals and has been successfully applied in many modern speech processing systems ranging from such diverse applications as coding, analysis, synthesis and recognition (see, e.g., [1]). The speech model used in many of these applications is the source-filter model where the speech signal is generated by passing an excitation through an all-pole filter (the predictor). Typically, the prediction coefficients are identified such that the 2-norm of the residual, the difference between the observed signal and the predicted signal, is minimized. This works well when the excitation signal is Gaussian and independent and identically distributed (i.i.d.) [2], consistent with the equivalent maximum likelihood approach to determine the coefficients [3]. However, when the excitation signal does not satisfy these assumptions, problems arise [2]. This is the case for voiced speech where the excitation can be considered to be of a quasi-periodic nature with a spiky excitation [1]. In this case, the spectral cost function associated with the minimization of the 2-norm of the residual can be shown to suffer from certain well known problems such as overemphasis on peaks and cancellation of errors [2]. In general, the shortcomings of LP in spectral envelope modeling can be traced back to the 2-norm minimization approach: by minimizing the

2-norm, the LP filter tries to cancel the input voiced speech harmonics causing the envelope to have a sharper contour than desired with poles close to the unit circle. 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 [4], [5], [6], and [7]) while some others are based on changing the statistical assumptions made on the prediction error in the minimization process (notably [8], [9], and [10]).

The above mentioned deficiencies of the 2-norm minimization in LP modeling have also repercussions in the speech coding scenario. In fact, while the 2-norm criterion is consistent with achieving minimal variance of the residual for efficient coding[1], sparse techniques are employed to encode the residual. Examples of this can be seen since early GSM standards with the introduction of multi-pulse excitation (MPE [12]) and regular-pulse excitation (RPE [13]) methods and, more recently, in sparse algebraic codes in code-excited linear prediction (ACELP [14]). In these cases, we can reasonably assume that the best 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 residual*. Early contributions (notably [9], [15], and [16]) have followed this line of thought questioning the fundamental validity of the 2-norm criterion with regards to speech coding. Nevertheless, to the authors' best knowledge, 2-norm minimization is the only criterion used in commercial speech codecs.

Traditional usage of LP is confined to modeling only the the spectral envelope capturing 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. The usual approach is then to employ a cascaded structure where LP is initially applied to determine the short-term prediction coefficients to model the spectral envelope and, subsequently, a long-term predictor is determined to model the harmonic behavior of the spectrum [1]. Such a structure is inherently suboptimal since it ignores the interaction between the two different stages. Also in this case, while early contributions have outlined gains in performance in jointly estimating the two filters (notably [17]), the common approach is to distinctly separate the two steps.

The recent developments in the field of sparse signal processing, backed up by significant improvements in convex optimization algorithms (e.g., interior point methods [18] [19]), have recently encouraged the authors to explore the concept of sparsity in the LP minimization framework [20]. In particular, while rein-

[1]The fundamental theorem of predictive quantization [11] 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.

troducing well known methods to seek a short-term predictor that produces a residual that is sparse rather than minimum variance, we have also introduced the idea of employing high order sparse predictors to model the cascade of short-term and long-term predictors, engendering a joint estimation of the two [21]. This preliminary work has led the way for the exploitation of the sparse characteristics of the high order predictor and the residual to define more efficient coding techniques. Specifically, in [22], we have demonstrated that the new model achieves a more parsimonious description of a speech segment with interesting direct applications to low bit-rate speech coding. While in these early works, the 1-norm has been reasonably chosen as a convex approximation of the so-called 0-norm[2], in [23] we have applied the reweighted 1-norm algorithm in order to produce a more focused solution to the original problem that we are trying to solve. In this work, we move forward, introducing the novelty of a compressed sensing formulation [24] in sparse LP, that will not only offer important information on how to retrieve the sparse structure of the residual, but will also help reduce the size of the minimization problem, with a clear impact on the computational complexity.

The contribution of this paper is then twofold. Firstly, we put our earlier contributions in a common framework giving an introductory overview of Sparse Linear Prediction and we also introduce its compressed sensing formulation. Secondly, we provide a detailed experimental analysis of its usefulness in modeling and coding applications transcending the well known limitations related to traditional LP.

The paper is organized as follows. In Section 2, we provide a prologue that defines the mathematical formulations of the proposed sparse linear predictors. In Section 3, we define the sparse linear predictors and, in Section 4, we provide their compressed sensing formulation. The results of the experimental evaluation of the analysis properties of the short-term predictors are outlined in Section 5, while the experimental results of the coding properties and applications are outlined in Section 6. We provide a discussion on some of the drawbacks of sparse linear prediction in Section 7. Finally, Section 8 concludes our work.

2 Fundamentals

We consider the following speech production model, where a sample of speech $x(n)$ is written as a linear combination of K past samples:

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

[2]The 0-norm is not technically a norm since it violates the triangle inequality.

where $\{a_k\}$ are the prediction coefficients and $r(n)$ is the prediction error. In particular, 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 [18]. Considering the speech production model for a segment of N speech samples $x(n)$, for $n = 1, \ldots, N$, in matrix form:

$$\mathbf{x} = \mathbf{X}\mathbf{a} + \mathbf{r}, \tag{E.2}$$

the problem becomes:

$$\mathbf{a} = \arg\min_{\mathbf{a}} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_p^p + \gamma \|\mathbf{a}\|_k^k, \tag{E.3}$$

where:

$$\mathbf{x} = \begin{bmatrix} x(N_1) \\ \vdots \\ x(N_2) \end{bmatrix}, \mathbf{X} = \begin{bmatrix} x(N_1 - 1) & \cdots & x(N_1 - K) \\ \vdots & & \vdots \\ x(N_2 - 1) & \cdots & x(N_2 - K) \end{bmatrix}. \tag{E.4}$$

The p-norm operator $\|\cdot\|_p$ is defined as $\|\mathbf{x}\|_p = (\sum_{n=1}^{N} |x(n)|^p)^{\frac{1}{p}}$. 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* [25]. The introduction of the regularization term γ in (E.3) can be seen as being related to the prior knowledge of the coefficients vector $\mathbf{a}$, problem (E.3) then corresponds to the *maximum a posteriori* (MAP) approach for finding $\mathbf{a}$ under the assumptions that $\mathbf{a}$ has a Generalized Gaussian Distribution [26]:

$$\begin{aligned} \mathbf{a_{MAP}} &= \arg\max_{\mathbf{a}} f(\mathbf{x}|\mathbf{a}) g(\mathbf{a}) \\ &= \arg\max_{\mathbf{a}} \{\exp(-\|\mathbf{x} - \mathbf{X}\mathbf{a}\|_p^p) \exp(-\gamma \|\mathbf{a}\|_k^k)\}. \end{aligned} \tag{E.5}$$

This reduces to the *maximum likelihood* (ML) approach when $\gamma = 0$, under the assumption that the residual is a vector of i.i.d. Generalized Gaussian variables:

$$\mathbf{a_{ML}} = \arg\max_{\mathbf{a}} f(\mathbf{x}|\mathbf{a}) = \arg\max_{\mathbf{a}} \{\exp(-\|\mathbf{x} - \mathbf{X}\mathbf{a}\|_p^p)\}. \tag{E.6}$$

The question now is how to choose p, k and γ in our minimization problem (E.3) and how to perform the associated minimization, depending on the kind of estimator we wish to implement.

In finding a sparse signal representation, there is the somewhat subtle problem of how to measure sparsity. Sparsity is often measured as the cardinality, corresponding to the so-called 0-norm $\|\cdot\|_0$. Our optimization problem (E.3) would then become:

$$\mathbf{a} = \arg\min_{\mathbf{a}} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_0 + \gamma \|\mathbf{a}\|_0, \tag{E.7}$$

with the particular case in which we are only considering the sparsity in the residual ($\gamma = 0$):

$$\mathbf{a} = \arg\min_{\mathbf{a}} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_0. \tag{E.8}$$

Unfortunately, these are combinatorial problems which generally cannot be solved in polynomial time. Instead of the cardinality measure, we will then use the more tractable 1-norm $\|\cdot\|_1$, which has shown to perform well as a linear programming relaxation of the 0-norm [27]. We will also consider a variation of the 1-norm minimization criterion such as the reweighted 1-norm [28] to enhance the sparsity measure and moving the solution closer to the original 0-norm problem (E.7). Throughout the paper, we will see the application scenarios and performances of the various predictors obtained by introducing sparsity in the LP framework.

3 Sparse Linear Predictors

In this section, we will define the different sparse linear predictors and show their application in the context of speech processing. In particular, we will introduce the problem of determining a short-term predictor that engenders a sparse residual and the problem of finding a high order sparse predictor that also engenders a sparse residual. This second formulation is, as we shall see, particularly relevant in providing a robust joint estimation of short-term and long-term predictors. Since in Section 2, we have introduced the 1-norm minimization as the sparsity measure, here we will also introduce the reweighted 1-norm algorithm to enhance this sparsity measure, moving closer to the original problem (0-norm minimization).

3.1 Finding a Sparse Residual

We consider the problem of finding a prediction coefficient vector $\mathbf{a}$ such that the resulting residual is sparse. Having identified the 1-norm as a suitable convex relaxation of the cardinality, the cost function for this problem is a particular case of (E.3). By setting $p = 1$ and $\gamma = 0$ we obtain the following optimization problem:

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

The 1-norm minimization criterion has already been shown to outperform 2-norm minimization in finding a proper linear predictive model in speech analysis [9] [15] [16]. In particular, comparing the cost functions associated with 2-norm minimization and 1-norm minimization, it can be easily shown that, when solving (E.9), lower emphasis is forced on the larger values present in the underlying excitation sequence. This property becomes particularly relevant when analyzing voiced speech, obtaining a more pronounced spiky behavior in the residual vector

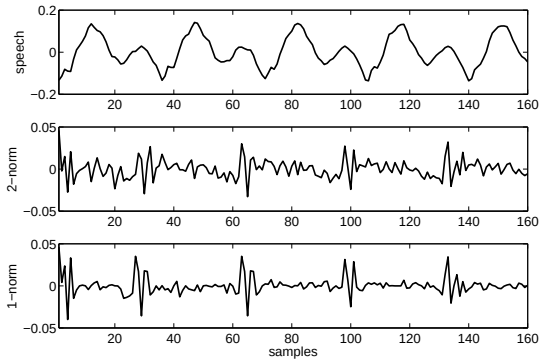

Fig. E.1: *An example of prediction residuals obtained by 2-norm and 1-norm error minimization. The speech segment analyzed is shown in the top box. The prediction order is* $K = 10$ *and the frame length is* $N = 160$*. It can be seen that the spiky pitch excitation is retrieved more accurately when 1-norm minimization is employed.*

consistent with the traditional impulse train representation of the residual. An example of this is shown in Figure E.1. In general, the smaller influence from the large values in the excitation creates a more efficient decoupling of the source excitation from the vocal tract transfer function, creating a more robust analysis tool [8]. This effect can be seen in the spectral envelope that will avoid the over-emphasis on peaks generated in the effort to cancel the pitch harmonics or, equivalently, the large spikes present in the residual. An example of this property is shown in Figure E.2.

The 1-norm minimization criterion, is also equivalent to the ML estimator when the residual is assumed to be i.i.d. Laplacian. This statistical interpretation is also meaningful, since it is well known that the distribution of speech samples is better described by a Laplacian distribution [29]. In the case of unvoiced speech, the Gaussian and Laplacian distributions both seem to provide appropriate models. However, by using the 1-norm minimization, we provide a residual that is sparser. In particular in [30] it is shown that, the residual vector provided by 1-norm minimization will have at least K components equal to zero.

3.2 Finding a High Order Sparse Predictor

We now consider the problem of finding a high order sparse predictor that also engenders a sparse residual. This problem is particularly relevant when considering the usual modeling approach adopted in low bit-rate predictive coding for voiced speech segments. This corresponds to a cascade of a short-term lin-

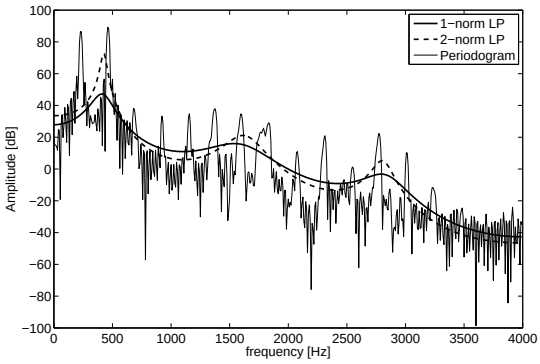

Fig. E.2: *An example of LP spectral model obtained by 1-norm and 2-norm error minimization for a segment of voiced speech. The prediction order is* $K = 10$ *and the frame length is* $N = 160$*. The lower emphasis on peaks in the envelope, when 1-norm minimization is employed, is a direct consequence of the ability to retrieve the spiky pitch excitation.*

ear predictor $F(z)$ and a long-term linear predictor $P(z)$ to remove respectively near-sample redundancies, due to the presence of formants, and distant-sample redundancies, due to the presence of a pitch excitation. The cascade of the predictors corresponds to the multiplication in the z-domain of the their transfer functions:

$$\begin{aligned} A(z) = F(z)P(z) &= 1 - \sum_{k=1}^{K} a_k z^{-k} \\ &= (1 - \sum_{k=1}^{N_f} f_k z^{-k})(1 - \sum_{k=1}^{N_p} g_k z^{-(T_p+k-1)}). \end{aligned} \tag{E.10}$$

The resulting prediction coefficient vector $\mathbf{a} = \{a_k\}$ of the high order polynomial $A(z)$ will therefore be highly sparse[3]. Taking this into account in our minimization process, and again considering the 1-norm as convex relaxation of the 0-norm, our original problem (E.7) becomes:

$$\min_{\mathbf{a}} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_1 + \gamma \|\mathbf{a}\|_1, \tag{E.11}$$

where the dimension of the prediction coefficient vector $\mathbf{a}$ (the order of the predictor) has to be sufficiently large to model the filter cascade ($K > N_f + T_p + N_p$) in (E.10). This approach, while maintaining resemblances to (E.9)

[3]Traditionally, for speech sampled at 8 kHz, $N_f = 10$, $N_p = 1$, and T_p usually belongs in the range $[16, 120]$.

looking for a sparse residual, is fundamentally different. While the predictor in (E.9) aims at modeling the spectral envelope, the purpose of the high order sparse predictor is to model the *whole* spectrum, i.e., the spectral envelope and the spectral harmonics. This can be easily achieved due to the strong ability of high order LP to resolve closely spaced sinusoids [31]. Nevertheless, with an appropriate choice of γ, the sparse high order predictor will still retain the above mentioned structure and can then be easily factorized into the original form of short-term and long-term components. The opportunity given by the high order predictor to jointly find these two components, translates into finding a short-term predictor that will be remarkably uncorrupted by the fine structure belonging to the pitch excitation and also a robust initial estimation of the pitch lag T_p. Furthermore, since the presence of the pitch excitation is taken into account by the high order sparse predictor, the sparse residual will present a very low mutual information among the samples without the characteristic train of pulses found when using only short-term prediction. An example of the predictor obtained as solution of (E.11) is shown in Figure E.3. An example of the spectral modeling properties is shown in Figure E.4.

The minimization problem in (E.11) also has a statistical meaning being equivalent to MAP estimation (E.5) under the assumptions that both the residual and the predictor are sets of i.i.d. Laplacian variables. However, while the assumption on the residual is still meaningful, the assumption on the high order predictor does not have any significant statistical interpretation. In this case, the 1-norm minimization should be considered merely as a convex relaxation of the 0-norm, then the prior on the coefficients vector $\mathbf{a}$ can be seen as a sparsity constraint, where the regularization term γ plays the role of a Lagrange multiplier. γ therefore controls *how sparse* the predictor should be and the trade-off between the sparsity of the predictor and the sparsity of the residual.

There are mainly two problems associated with exploiting the modeling properties of the sparse high order predictor: determining an appropriate value of γ to solve (E.11) and using an approximate factorization to obtain again the initial formulation composed by the two predictors (E.10). Below we address these two issues.

Selection of γ

It is clear from (E.11) that if the regularization term γ is too small or too large the obtained solution may be useless. In particular, by increasing γ, we increase the sparsity of the prediction coefficient vector, until all its entries are zero ($A(z) = 1$) for $\gamma \geq \|\mathbf{X}^T\mathbf{x}\|_\infty$ (where $\|\cdot\|_\infty$ denotes the dual norm to $\|\cdot\|_1$). More precisely, for $0 < \gamma < \|\mathbf{X}^T\mathbf{x}\|_\infty$, the solution vector $\mathbf{a}$ is a linear function of γ. However, in general, the number of nonzero elements in $\mathbf{a}$ is not necessarily a monotonic function of γ.

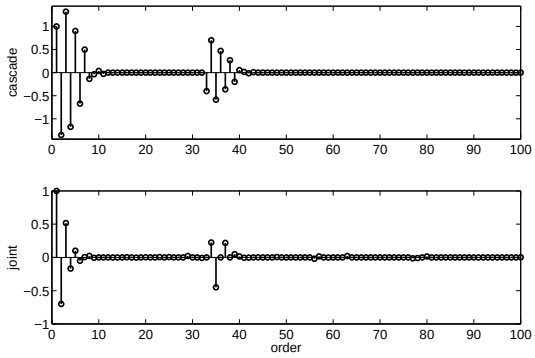

Fig. E.3: *An example of the high order predictor coefficient vector resulting from a cascade of long-term and short-term predictors (top box) and the solution of (E.11) for* $\gamma = 0.1$ *and order* $K = 100$. *The order is chosen sufficiently large to accommodate the filter cascade (E.10). It can be seen that the sparse prediction coefficient vector resembles accurately the cascade of the two predictors.*

There are obviously several ways of determining γ. In our previous work [21] [22], we have found the modified L-curve [32] as an efficient tool to find a balanced sparse representation between the two descriptions. We find the modified L-curve $(\|\mathbf{x} - \mathbf{X}\mathbf{a}_\gamma\|_1, \|\mathbf{a}_\gamma\|_1)$ by solving the minimization problem (E.11) for several values of γ in the interval $0 < \gamma < \|\mathbf{X}^T\mathbf{x}\|_\infty$. The optimal value of γ (in the L-curve sense) is found as the point of maximum curvature of this curve. Considering the 1-norm as a convex relaxation of the 0-norm, then clearly the γ chosen with the L-curve is an efficient way to determine an appropriate sparse representation of the predictor and the residual in (E.11). We have also observed that, in general, a constant value of γ, chosen for example as the average value of the set of γ's found with the L-curve based approach for a large set of speech frames, is an appropriate choice in the predictive problems considered. In the experimental analysis we will consider both approaches to defining γ.

Factorization of the High Order Polynomial

If γ is chosen appropriately, the considered formulation (E.11) results in a high order predictor $\hat{A}(z)$ with a clear structure that resembles the cascade of the short-term and long-term predictor (Figure E.3). We can therefore bring $\hat{A}(z)$ to the original formulation in (E.10), by applying a simple and effective ad-hoc method to factorize the solution [22]. In particular, we use the first N_f

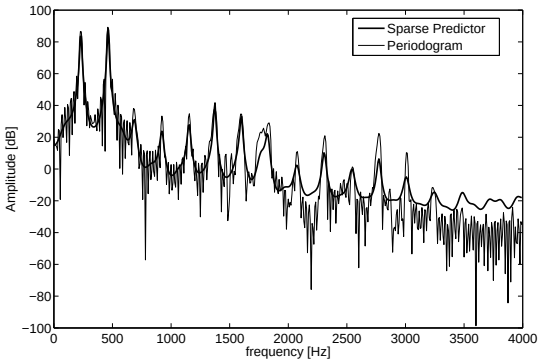

Fig. E.4: *Frequency response of the high order predictor of Figure E.3. The order of the predictor is $K = 100$ and we consider only the nine nonzero coefficients of largest magnitude modeling the short-term and long-term predictors cascade.*

coefficients as the estimated coefficients of the short-term predictor:

$$\hat{F}(z) = 1 - \sum_{k=1}^{N_f} \hat{a}_k z^{-k}, \tag{E.12}$$

and then compute the quotient polynomial $\hat{Q}(z)$ of the division of $\hat{A}(z)$ by $\hat{F}(z)$ so that:

$$\hat{A}(z) = \hat{Q}(z)\hat{F}(z) + E(z) \approx \hat{Q}(z)\hat{F}(z), \tag{E.13}$$

where the deconvolution remainder $E(z)$ is considered to be negligible. From the polynomial $\hat{Q}(z)$ we can then extract the N_p taps predictor. In this paper, we will consider the most common pitch predictor where $N_p = 1$ ($P(z) = 1 - g_p z^{-T_p}$), then we merely identify the minimum value and its position in the coefficients vector of $\hat{Q}(z)$:

$$\begin{aligned} g_p &= \min\{q_k\}, \\ T_p &= \arg\min\{q_k\}. \end{aligned} \tag{E.14}$$

It is clear that, while heuristic, this factorization procedure is highly customizable. A different numbers of taps for both the short-term and long-term can be selected and also a voiced/unvoiced classification can be included, based on the presence or absence of long-term information, as described in [21, 22].

It should be noticed that the structure of the cascade can also be incorporated into the minimization scheme and can be potentially beneficial in reducing the size of the problem. This approach is then similar to the *One-Shot Combined Optimization* presented in [17]. Nevertheless, to obtain the same result

as (E.11), we require prior knowledge on the position of the pitch contributions and the model order of both the short-term and long-term predictors, making this approach impractical.

3.3 Enhancing Sparsity by Reweighted 1-norm Minimization

As shown throughout this section, the 1-norm is used as a convex relaxation of the 0-norm, because 0-norm minimization yields a combinatorial problem (NP-hard). We are therefore interested in adjusting the error weighting difference between the 1-norm and the 0-norm. A variety of recently introduced methods have dealt with this issue relying on iterative reweighted 1-norm minimization (see, e.g., [33] and references therein). In particular, the iteratively reweighted 1-norm minimization may be used for estimating $\mathbf{a}$ and enhancing the sparsity of $\mathbf{r}$ (and $\mathbf{a}$), while keeping the problem solvable with convex tools [28] [23]. The predictor can then be seen as a solution of the following minimization problem:

$$\mathbf{a} = \arg\min_{\mathbf{a}} \lim_{p\to 0} \lim_{k\to 0} \{\|\mathbf{x} - \mathbf{X}\mathbf{a}\|_p^p + \gamma \|\mathbf{a}\|_k^k\}. \tag{E.15}$$

From a optimization point of view, for the case $p \leq 1$, the cost functions will have even lower emphasis on large values and a sharper slope near zero. In particular, this means that the minimization will encourage small values to become smaller while enhancing the amplitude of larger values. The limit case for $p = 0$ will have an infinitely sharp slope in zero and equally weighted tails. This will introduce as many zeros as possible as these are infinitely weighted.

The algorithm to obtain a short-term predictor engendering a sparser residual, a reweighted formulation of (E.9), is shown in Algorithm 3. This approach, as we shall see becomes beneficial in finding a predictor that produces a sparser residual, providing a tighter coupling between the prediction estimation and the search for the approximated sparse excitation. An example is shown in Fig. E.5.

When we impose sparsity both on the residual and on the high order predictor, as in (E.11), the algorithm is modified as shown in Algorithm 4. The formulation in Algorithm 4, while enhancing the sparsity of the residual similarly to Algorithm 3, is particularly relevant due to the presence of near-zero components in the high order predictor obtained (see Fig. E.3). We are therefore interested in putting to zero these spurious components, while enhancing the larger components that contain information of near-end and far-end redundancies. This will be beneficial in finding also a better estimate of the short-term and long-term contributions through the approximate factorization presented in 3.2.

In both algorithms, the parameter $\epsilon > 0$ is used to provide stability when a component of $\hat{\mathbf{r}}$ goes to zero. ϵ does not need to be too small; as empirically

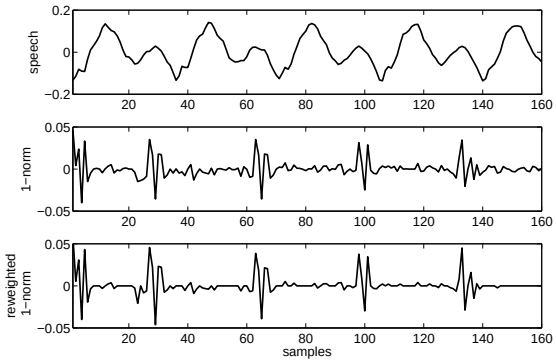

Fig. E.5: *An example of prediction residuals obtained through 1-norm and reweighted 1-norm error minimization using Algorithm 3. The speech segment analyzed is shown in the top box. The prediction order is* $K = 10$ *and the frame length is* $N = 160$*. Five iterations where made with* $\epsilon = 0.01$*.*

demonstrated in [28], it should be in the order of the expected nonzero magnitude of $\mathbf{r}$. Also, it has been shown in [28] that $\|\hat{\mathbf{r}}^{i+1}\|_1 \leq \|\hat{\mathbf{r}}^{i}\|_1$, meaning that this is a descent algorithm. The halting criterion can therefore be chosen as either a maximum number of iterations or as a convergence criterion. The choice of the weights, as the inverse of the magnitude of the residual, is made to penalize every nonzero coefficient equally, as done by the 0-norm. In the experimental analysis we will give details on how many iterations are required in our setting. In [28] and [33], it is also shown that the reweighted 1-norm algorithm, at convergence, is equivalent to the minimization of the log-sum penalty function. This is relevant to what we are trying to achieve in (E.15): the log-sum cost function has a sharper slope near zero compared to the 1-norm, providing more effective sparsity inducing properties. Furthermore, since the log-sum is not convex, the iterative algorithm corresponds to minimizing a sequence of linearizations of the log-sum around the previous solution estimate, providing at each step a sparser solution (until convergence).

4 Compressed Sensing Formulation for Sparse Linear Prediction

In this section, we explore the compressed sensing (CS) formulation for the sparse linear predictors introduced in Section 3. The CS formulation is particularly interesting in our problems: by exploiting prior knowledge about the sparsity of the signal $\mathbf{x}$ we will show that a limited number of random projections

Algorithm 3 Iteratively Reweighted 1-norm Minimization of the Residual

Inputs: speech segment $\mathbf{x}$
Outputs: predictor $\hat{\mathbf{a}}^i$, residual $\hat{\mathbf{r}}^i$
$i = 0$, initial weights $\mathbf{W}^{i=0} = \mathbf{I}$
while halting criterion false **do**
1. $\hat{\mathbf{a}}^i, \hat{\mathbf{r}}^i \leftarrow \arg\min_{\mathbf{a}} \|\mathbf{W}^i\mathbf{r}\|_1 \quad \text{s.t.} \quad \mathbf{r} = \mathbf{x} - \mathbf{X}\mathbf{a}$
2. $\mathbf{W}^{i+1} \leftarrow \text{diag}\left(\left|\hat{\mathbf{r}}^i\right| + \epsilon\right)^{-1}$
3. $i \leftarrow i + 1$

end while

Algorithm 4 Iteratively Reweighted 1-norm Minimization of Residual and Predictor

Inputs: speech segment $\mathbf{x}$
Outputs: predictor $\hat{\mathbf{a}}^i$, residual $\hat{\mathbf{r}}^i$
$i = 0$, initial weights $\mathbf{W}^{i=0} = \mathbf{I}$ and $\mathbf{D}^{i=0} = \mathbf{I}$
while halting criterion false **do**
1. $\hat{\mathbf{a}}^i, \hat{\mathbf{r}}^i \leftarrow \arg\min_{\mathbf{a}} \|\mathbf{W}^i\mathbf{r}\|_1 + \gamma\|\mathbf{D}^i\mathbf{a}\|_1$
   $\text{s.t.} \quad \mathbf{r} = \mathbf{x} - \mathbf{X}\mathbf{a}$
2. $\mathbf{W}^{i+1} \leftarrow \text{diag}\left(\left|\hat{\mathbf{r}}^i\right| + \epsilon\right)^{-1}$
3. $\mathbf{D}^{i+1} \leftarrow \text{diag}\left(\left|\hat{\mathbf{a}}^i\right| + \epsilon\right)^{-1}$
4. $i \leftarrow i + 1$

end while

are sufficient to recover our predictors and sparse residual with high accuracy. In particular, it has been shown [24] [34] that a random projection of a high-dimensional but sparse or compressible signal vector onto a lower-dimensional space contains enough information to be able to reconstruct the signal with small or zero error.

In recent work, notably [35] and [36], CS formulations in the context of speech analysis and coding have been formulated in order to find a sparse approximation of the residual, given the predictor. It is then interesting to extend this work to the case where we want to find directly the predictor that engenders intrinsically a sparse residual. In particular, given the sparsity level of the sparse representation that we wish to retrieve in a given domain, we can determine an efficient *shrinkage* of the minimization problem in a lower dimensional space, with a clear impact on the computational complexity.

If we wish to perform CS, two main ingredients are needed: a domain where the analyzed signal is sparse and the sparsity level of this signal T. In our case, the residual is the domain where the signal is sparse, while the linear transform that maps the original speech signal to the sparse residual is the sparse predictor.

The sparsity in the residual domain is then imposed by our needs [34]. Let us now review the formulation presented in [36]:

$$\hat{\mathbf{r}} = \arg\min_{\mathbf{r}} \|\mathbf{r}\|_1 \quad \text{s.t.} \quad \mathbf{\Phi x} = \mathbf{\Phi H r} \tag{E.16}$$

where $\mathbf{x}$ is the $N \times 1$ analyzed segment of speech, $\mathbf{H}$ the $N \times (N+K)$ synthesis matrix, constructed from the truncated impulse response of the *known* predictor [37], $\mathbf{r}$ is the residual vector to be estimated (supposedly sparse) and $\mathbf{\Phi}$ is the sensing matrix of dimension $M \times N$. The dimensionality of the random linear projection M stems from the sparsity level T that one wishes to impose on the residual. In particular, based on empirical results, the number of projections is set equal to four times the sparsity, i.e. $M = 4T$. Furthermore, when the incoherence between the synthesis matrix and the random basis matrix $\mathbf{\Phi}$ holds ($\mu(\mathbf{\Phi}, \mathbf{H}) \approx 1$), even if $\mathbf{H}$ is not orthogonal the recovery of the sparse residual $\mathbf{r}$ is still possible and the linear program in (E.16) gives an accurate reconstruction of $\mathbf{x}$ with very high probability [24, 36].

To adapt CS principles to the estimation of the predictor as well, let us now consider the relation between the synthesis matrix $\mathbf{H}$ and the analysis matrix $\mathbf{A}$ where one is the pseudo-inverse of the other [38]:

$$\mathbf{A} = \mathbf{H}^{+}. \tag{E.17}$$

We can now replace the constraint $\mathbf{\Phi x} = \mathbf{\Phi H r}$ in (E.16) as

$$\mathbf{\Phi r} = \mathbf{\Phi A x}, \tag{E.18}$$

where $\mathbf{A}$ is the $(N+K) \times N$ analysis matrix that performs the whitening of the signal, constructed from the coefficients of the predictor $\mathbf{a}$ of order K [38], the dimension of the sensing matrix $\mathbf{\Phi}$ is now adjusted accordingly to $M \times (N+K)$. Notice that, due to the structure of $\mathbf{A}$ this can be rewritten equivalently to:

$$\mathbf{\Phi r} = \mathbf{\Phi A x} = \mathbf{\Phi}\left[\mathbf{x}|\mathbf{X}\right]\left[1, \mathbf{a}^T\right]^T, \tag{E.19}$$

where $[\mathbf{x}|\mathbf{X}]$ is the matrix obtained by stacking the vector $\mathbf{x}$ to the left of $\mathbf{X}$ in (E.4). The minimization problem can then be rewritten as:

$$\min_{\mathbf{a},\mathbf{r}} \|\mathbf{r}\|_1 \quad \text{s.t.} \quad \mathbf{\Phi r} = \mathbf{\Phi}(\mathbf{x} - \mathbf{X a}). \tag{E.20}$$

We can now see that (E.20) is *equivalent* to (E.9), the only difference being the projection onto the random basis in the constraint. The results obtained will then be similar to our initial formulation (E.9), as long as the choice of $\mathbf{\Phi}$ is appropriate. In this case, the formulation in (E.20) will not only provide hints on the T pulses to be selected in the residual, but also a dimensionality reduction

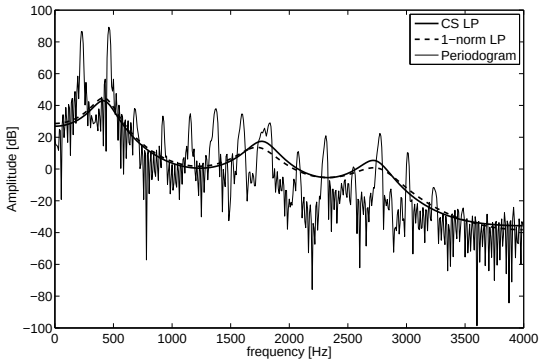

Fig. E.6: *An example of LP spectral model obtained through 1-norm minimization (E.9) and through CS based minimization (E.20) for a segment of voiced speech. The prediction order is $K = 10$ and the frame length is $N = 160$, for the CS formulation the dimension of the sensing matrix is $M = 80$, corresponding to the sparsity level $T = 20$.*

that will simplify the calculations. This computational complexity reduction, resulting from the dimensionality reduction given by the projection onto random basis has been also observed in [39] and arises from the Johnson-Lindestrauss lemma [40]. An example of an envelope estimation using the formulation in (E.20) is presented in Figure E.6 while the recovered sparse residual is shown in Figure E.7.

Similarly, if we are looking for a high order sparse predictor, the problem (E.11) can be cast into a CS framework leading to:

$$\arg\min_{\mathbf{a},\mathbf{r}} \|\mathbf{r}\|_1 + \gamma\|\mathbf{a}\|_1 \quad \text{s.t.} \quad \mathbf{\Phi r} = \mathbf{\Phi}(\mathbf{x} - \mathbf{X a}). \tag{E.21}$$

Both formulations (E.20) and (E.21), can also be modified to involve iterative reweighting (Algorithm 5 shows the general case for $\gamma > 0$). In [28] applications of the reweighted 1-norm minimization in a CS framework are provided.

5 Properties of Sparse Linear Prediction

As mentioned in the introduction, many problems appearing in traditional 2-norm LP modeling of voiced speech can be traced back to the inability of the predictor to decouple the vocal tract transfer function from the pitch excitation. This results in a lower spectral modeling accuracy and a strong dependence on the placement of the analysis window. In this section we provide some experiments to illustrate how the sparse linear predictors presented in the previous sections manage to overcome these problems. As a general remark, it is well

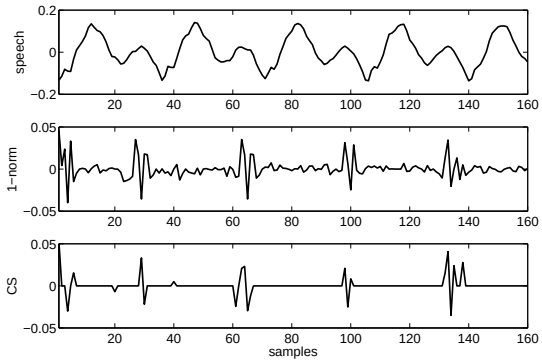

Fig. E.7: *An example of prediction residuals obtained through 1-norm minimization and CS recovery. The speech segment analyzed is shown in the top box. The prediction order is* $K = 10$ *and the frame length is* $N = 160$*. For the CS formulation, the imposed sparsity level is* $T = 20$*, corresponding to the size* $M = 80$ *for the sensing matrix.*

known that the $p-$norm LP estimate with $p \neq 2$ is not guaranteed to be stable [41]. Nevertheless, the results presented in this section concentrate on the spectral modeling properties of sparse LP, thus the stability of the predictor is simply imposed by pole reflection which stabilizes the filter without modifying the magnitude of the frequency response. We will provide a thorough discussion of the stability issues in the Section 7 and in Section 6 where the speech coding properties are analyzed and stability is critical.

The experimental analysis has been done on 20,000 frames of length $N = 160$ (20 ms) of clean voiced speech coming from several different speakers with different characteristics (gender, age, pitch, regional accent) taken from the TIMIT database, downsampled at 8 kHz. The prediction method we will compare in this section are shown in Table E.1. The optimality of the methods **BE** and **RLP**, presented in [6], comes from the selection of the parameters which provide the lowest distortion compared with the reference envelope. For brevity and clarity of the presented results, we have omitted the predictors obtained as solutions of the iterative reweighted algorithms presented in Section 3.3 and the CS formulation presented in Section 4. These methods, while presenting very similar modeling properties to **SpLP10** and **SpLP11**, produce predictors estimates with slightly higher variance, thus requiring few more bits to be encoded. Therefore, while it is hard to provide a fair comparison in terms of modeling, their properties become more interesting in the coding scenario that will thoroughly analyzed in Section 6; in particular, the differences in their bit allocation necessary for efficient coding and the information required in the residual will be analyzed.

Algorithm 5 CS Formulation of the Iteratively Reweighted 1-norm Minimization of Residual and Predictor

Inputs: speech segment $\mathbf{x}$, desired residual sparsity level T
Outputs: predictor $\hat{\mathbf{a}}^i$, residual $\hat{\mathbf{r}}^i$
$i = 0$, initial weights $\mathbf{W}^{i=0} = \mathbf{I}$ and $\mathbf{D}^{i=0} = \mathbf{I}$,
random matrix $\mathbf{\Phi}$ of size $M \times (N + K)$, $M = 4T$
while halting criterion false **do**

1. $\hat{\mathbf{a}}^i, \hat{\mathbf{r}}^i \leftarrow \arg\min_{\mathbf{a}} \|\mathbf{W}^i\mathbf{r}\|_1 + \gamma\|\mathbf{D}^i\mathbf{a}\|_1$
   s.t. $\mathbf{\Phi r} = \mathbf{\Phi}(\mathbf{x} - \mathbf{X a})$
2. $\mathbf{W}^{i+1} \leftarrow \text{diag}\left(\left|\hat{\mathbf{r}}^i\right| + \epsilon\right)^{-1}$
3. $\mathbf{D}^{i+1} \leftarrow \text{diag}\left(\left|\hat{\mathbf{a}}^i\right| + \epsilon\right)^{-1}$
4. $i \leftarrow i + 1$

end while

5.1 Spectral Modeling

In this section, we provide results to the modeling properties of the short-term predictors. As a reference, we use the envelope obtained through a cubic spline interpolation between the harmonics peaks of the logarithmic periodogram. This method was presented in [6] and provides an approximation of the vocal tract transfer function, without the fine structure corresponding to the pitch excitation. We then calculate the log spectral distortion between our reference envelope $S_{int}(\omega)$ and the estimated predictive model $S(\omega, \mathbf{a})$ as:

$$\text{SD}_m = \sqrt{\frac{1}{2\pi}\int_{-\pi}^{\pi} \left[10\log_{10} S_{int}(\omega) - 10\log_{10} S(\omega, \mathbf{a})\right]^2 d\omega}. \tag{E.22}$$

where the numerator gain is calculated as the variance of the residual.

The coefficients of the short-term predictors presented have also shown to be smoother and therefore they have a lower sensitivity to quantization. We will also compare the log spectral distortion between our reference envelope $S_{int}(\omega)$ and the quantized predictive model $S(\omega, \hat{\mathbf{a}})$ for every predictor obtained with the presented methods. The quantizer used is the one presented in [42], with the number of bits fixed at 20 for the different prediction orders, providing in all the method presented a *transparent coding*[4]. The results are shown in Table E.2 for different prediction orders. A critical analysis of the results show the superior modeling properties of **SpLP11**. This is given by its ability to

[4]According to [43], transparent coding of LP parameters is achieved when the two versions of coded speech, obtained using unquantized LP parameters and quantized LP parameters, are indistinguishable through listening. This is usually achieved with an average log distortion between quantized and unquantized spectra lower that 1 dB, with no outliers with log distortion greater than 4 dB and a number of outliers with 2-4 dB distortion lower than 2%.

Table E.1: Prediction methods compared in the modeling properties evaluation.

Method	Description
LP	Traditional 2-norm LP with 10Hz bandwidth expansion ($\gamma = 0.996$) and Hamming windowing.
SpLP10	1-norm LP presented in (3.1), solution of (E.9). Stability is imposed by pole reflection if unstable. No windowing is performed.
SpLP11	1-norm LP presented in (3.2). The order of (E.11) is $K = 110$ (covering accurately pitch delays in the interval $[N_f + 1, K - N_f - 1]$). γ is chosen as the point of maximum curvature in the L-curve. The short-term predictor coefficients are the first N_f coefficients of the high order polynomial. Stability is imposed by pole reflection if unstable. No windowing is performed.
BE	Optimally bandwidth expanded 2-norm LP as shown in [6]. Hamming window is used.
RLP	Optimally regularized 2-norm LP as shown in [6]. Hamming window is used.

take into consideration the whole speech production model, thus decoupling more effectively the short-term contribution that provides the spectral envelope from the contribution given by the pitch excitation. **SpLP10** and **RLP** achieve similar performances, providing evidence supporting the generally good spectral modeling properties of the minimization problem in (E.9).

5.2 Shift Invariance

In speech analysis, a desirable property for an estimator is to be invariant to the small shifts of the analysis window, since speech, and voiced speech in particular, is assumed to be short-term stationary. However, standard LP is well known not to be shift invariant [8]. This is a direct consequence of the coupling between the vocal tract transfer function and the underlying pitch excitation that standard LP introduces in the estimate. To analyze the invariance of the LP methods to window shifts, we take the same 20,000 frames of clean voiced speech and we

Table E.2: Average spectral distortion for the considered methods in the unquantized case (SD_m) and quantized case (SD_q). A 95% confidence interval is given for each value.

METHOD	K	SD_m	SD_q
LP	8	2.11±0.06	3.24±0.11
	10	1.97±0.03	2.95±0.09
	12	1.98±0.05	2.72±0.12
SpLP10	8	1.91±0.01	2.92±0.02
	10	1.78±0.01	2.53±0.02
	12	1.61±0.01	2.31±0.04
SpLP11	8	1.64±0.00	2.65±0.01
	10	1.69±0.00	2.37±0.01
	12	1.39±0.01	2.13±0.01
BE	8	2.04±0.03	3.11±0.08
	10	1.88±0.02	2.92±0.07
	12	1.83±0.10	2.71±0.04
RLP	8	1.89±0.02	2.93±0.04
	10	1.72±0.01	2.51±0.03
	12	1.53±0.02	2.22±0.04

expand them to the left and to the right with 20 samples, giving a total length $N = 200$. In each frame of length $N = 200$ we define a $M = 160$ samples boxcar window and we shift the window by $s = 1, 2, 5, 10, 20$ samples. The average difference of the 10^{th} order AR estimate between $S_0(\omega)$ and $S_s(\omega)$ is analyzed. The average differences obtained for the methods in Table E.1 are shown in Table E.3. In Figure E.8, we show an example of the shift invariance property. The results obtained indicate clearly the sparse predictor robustness to small shifts in the analyzed window. Also in this case, the change in the frequency response in traditional LP is clearly given by the pitch bias in the estimate of the predictor, particularly dependent on the location of the spikes of the pitch excitation. The approaches **SpLP11** and **SpLP10**, since they do not try to cancel this characteristic spiky excitation, are less dependent from its location and provide a more robust estimate of the true envelope.

5.3 Pitch Independence

The ability of the sparse linear predictors to decouple the pitch excitation from the vocal tract transfer function is reflected also in the ability to have estimates of the envelope that are not affected by the pitch excitation. In this experiment, we calculate the envelope using 10^{th} order regularized LP (**RLP**) and we model

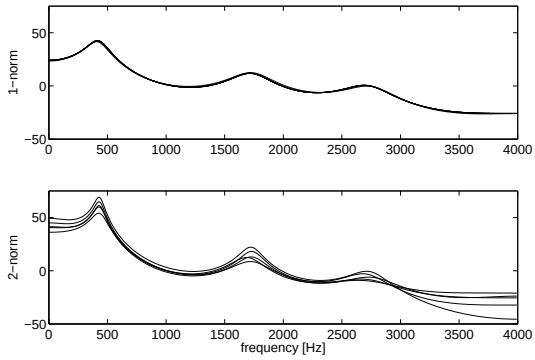

Fig. E.8: *An example of the shift invariance property of the sparse linear predictor (***SpLP11***) (top box), compared to traditional LP (***LP***). Ten envelopes are analyzed by shifting a the analysis window (160 samples) of* $s = 1, 2, 5, 10, 20$ *samples over a stationary voiced speech segment (length 200 samples).*

Table E.3: Average spectral distortion for the considered methods with shift of the analysis window $s = 1, 2, 5, 10, 20$.

METHOD	SD_1	SD_2	SD_5	SD_{10}	SD_{20}
LP	0.113	0.128	0.223	0.452	1.262
SpLP10	0.003	0.003	0.011	0.017	0.032
SpLP11	0.001	0.002	0.005	0.006	0.009
BE	0.097	0.117	0.197	0.238	0.328
RLP	0.015	0.089	0.180	0.201	0.323

the underlying pitch excitation with an impulse train with different spacing. We then analyze the synthetic speech applying the different LP methods in Table E.1. We divide the analysis into three subsets: high pitched $T_p \in [16, 35]$ ($f_0 \in [228\text{Hz}, 500\text{Hz}]$), mid pitched $T_p \in [36, 71]$ ($f_0 \in [113\text{Hz}, 222\text{Hz}]$) and low pitched $T_p \in [72, 120]$ ($f_0 \in [67\text{Hz}, 111\text{Hz}]$). The shortcomings of LP can be particularly seen in high pitched speech, as shown in the results of Table E.4. Because high pitched speakers have fewer harmonics within a given frequency range, modeling of the spectral envelope is more difficult and particularly problematic for traditional LP. The sparse linear predictors are basically unaffected by the underlying pitch excitation, which results in an improved spectral modeling. In particular for **SpLP11**, since the high order structure of the initial estimate includes the pitch harmonic structure, the extracted short-term predic-

Table E.4: Average spectral distortion for the considered methods with different underlying pitch excitation. A 95% confidence interval is given for each value.

METHOD	low	mid	high
LP	0.81±0.12	1.04±0.23	1.32±0.56
SpLP10	0.02±0.00	0.09±0.00	0.11±0.01
SpLP11	0.00±0.00	0.00±0.00	0.01±0.00
BE	0.45±0.07	0.65±0.19	0.89±0.34
RLP	0.05±0.02	0.16±0.10	0.19±0.09

tor is particularly robustly independent from the underlying excitation.

6 Coding Applications of Sparse Linear Prediction

By introducing sparsity in the residual, we can reasonably assume that only a small portion of the residual samples are sufficient to reconstruct the speech signal with high accuracy. We will corroborate our intuition by providing some experiments on the coding applications of sparse linear prediction. Specifically, in 6.1, we will first give experimental proof of the sparsity inducing effectiveness of the short-term predictors in the Analysis-by-Synthesis (AbS) scheme [37]. In this case, we use a very simple excitation model coding without long-term prediction where we exploit directly the information on the location of the nonzero samples. In 6.2, we will present a simple coding procedure that exploits the properties of the combined high order sparse LP and sparse residual. As we shall see in 6.3, this approach presents interesting properties such as noise robustness for which we give both objective and subjective evaluation.

As a general remark, since the stability of the short-term predictors is not assured, we will consistently perform a stability check and, if the short-term predictor is found to be unstable, we will perform a pole reflection. Note that this approach will necessarily modify the time domain behavior of the residual as well as the predictor coefficients. Nevertheless, since the rate of unstable filters is low and the instability is very mild (i.e., the magnitude of the poles is only very slightly higher than one), this can be considered as an adequate solution to this problem. We will return to the stability issue in Section 7.

Table E.5: Prediction methods compared in the coding properties evaluation.

Method	Description
LP	Traditional 2-norm LP with a fixed bandwidth expansion of 60 Hz (done by lag-windowing the autocorrelation function) and Hamming windowing.
SpLP10	1-norm LP solution of (E.9).
RWLP10	Reweighted 1-norm LP presented in Section 3.3 using Algorithm 3. Four reweighting iterations are performed (sufficient for convergence).
CSLP10	Compressed sensing formulation presented in Section 4, solution of (E.20). The size of the sensing matrix is given by the number of samples we want to retrieve in the residual.
RWCSLP10	Reweighed compressed sensing formulation of **CSLP10** using Algorithm 3. Four reweighting iterations are performed (sufficient for convergence).

6.1 Coding Properties of the Short-Term Sparse Linear Predictor

The first experiment regards the use of the short-term predictor in speech coding. In particular we will compare the use of the multipulse encoding procedure in the case of bandwidth expanded linear prediction (**LP**) with a fixed bandwidth expansion of 60 Hz (done by lag-windowing the autocorrelation function). We will compare this approach with our introduced sparse linear predictors. The only difference is that, instead of performing the multipulse encoding, we perform the AbS procedure straight after selecting the T positions of the T largest samples that are located in the residual. In this experiment, we will not perform long-term prediction, focusing only on the coding properties of the sparsity inducing short-term predictors.

We consider the formulation **SpLP10**, reweighted 1-norm **RWLP10**, and their CS formulations **CSLP10** and **RWCSLP10**. The methods compared

Table E.6: Comparison between the sparse predictor estimation methods. A 95% confidence interval is given for each value.

METHOD	T	$\hat{\mathbf{a}}$	SSNR	MOS	t
LP	5	19	14.1±3.2	2.85±0.23	0.1±0.1
	10	19	19.1±2.9	3.01±0.16	0.9±0.3
SpLP10	5	18	15.3±2.1	2.87±0.12	1.3±0.2
	10	18	20.1±1.7	3.11±0.11	1.3±0.2
RWLP10	5	22	17.2±1.6	3.01±0.06	4.1±0.3
	10	22	21.4±1.5	3.19±0.03	4.1±0.3
CSLP10	5	19	16.9±1.9	2.97±0.04	0.4±0.0
	10	19	20.9±1.5	3.25±0.03	0.6±0.2
RWCSLP10	5	24	20.2±0.9	3.15±0.03	1.3±0.3
	10	24	24.4±0.4	3.43±0.01	1.9±0.2

are summarized in Table E.5. As mentioned in Section 5, all these methods achieve similar modeling performances to **SpLP10**, although their estimate of the predictor requires a slightly larger number of bits. Here we will show this providing a comparison also in terms of bits needed for transparent quantization of the predictor. The methods **BE** and **RLP**, presented in the previous section (Table E.1) while offering better modeling properties than traditional LP, do not provide any significant improvement in the coding scenario, thus they will be omitted from the current experimental analysis.

We have performed the analysis on the same speech signals database considered in Section V. The frame size is $N = 40$, the 10^{th} order predictors are quantized transparently using the LSFs coding method in [42] while the T pulses are left unquantized. In the CS formulations the sensing matrix has $M = 4T$ rows; this means that just a slight reduction in the size of the problem is obtained when $T = 10$. Nevertheless we are able to obtain important information on the location of the pulses. In the reweighted schemes, the number of iterations is four, which is sufficient to reach convergence in all the analyzed frames.

In Table E.6, we present the results in terms of Segmental SNR, Mean Opinion Score (obtained through PESQ evaluation) and empirical computational time t in elapsed CPU seconds for $T = 5$ and $T = 10$, and number of bits necessary to transparently encode the predictor ($\hat{\mathbf{a}}$) using LSFs [42]. The results demonstrate the effectiveness of the sparse linear predictors. These results also show that the predictors in the reweighted cases (**RWLP10** and **RWCSLP10**), need a larger number of bits for transparent quantization due to the larger variance of their estimates. This result is particularly interesting when considering the model in (E.2). In particular, the description of a segment of speech is distributed

between its predictive model and the corresponding excitation. Thus, we can observe that the complexity of the predictor necessarily increases when the complexity of the residual decreases (less significant pulses). This also leaves open questions on the *optimal* bit distribution between the two descriptions. As a proof of concepts, the results show how only 5 bits of difference between **LP** and **RWCSLP10** in the representation of the filter result in a significant improvement in performance: only 5 pulses in the residual are necessary in **RWCSLP10** to obtain similar performances to **LP** using 10 pulses.

A critical analysis of the results leads to another interesting conclusion. In fact, while 1-norm based minimization, with or without the *shrinkage* of the problem provided by the CS formulation in (E.20), is computationally more costly, than 2-norm minimization, it greatly simplifies the next stage where the excitation is selected in a closed-loop AbS scheme. In particular, the empirical computational time in Table E.6 refers to both the LP analysis stage and the search for the MPE excitation. Since the MPE search for the location is not performed in our sparse LP methods and we exploit directly the information regarding the T pulses of largest magnitude, the AbS procedure is merely a small least square problem where we find the T pulse amplitudes. We will come back to the discussion regarding complexity in 7.2. Furthermore, it should be noted that the CS formulation improves the selection of the T largest pulses. This is remarkable since while the predictor obtained with or without the random projection is similar, the reduction of the constraints helps us find a more specific solution for the level of sparsity T that we would like to retrieve in the residual. As mentioned above, the price to pay is a slightly higher bit allocation for the predictors obtained through CS formulation.

6.2 Speech Coding Based on Sparse Linear Prediction

As a proof of concepts, we will now present a very simple coding scheme that summarizes all the previously introduced methods. We will use the method presented in Section 3.2, exploiting the sparse characteristics of the high order predictor and the sparse residual. In order to reduce the number of constraints, we cast the problem in a CS formulation (E.21) that provides a shrinkage of the constraint according to the number of samples we wish to retrieve in the residual. Furthermore, in order to refine the initial sparse solution, we apply the reweighting algorithm. The core scheme is summarized in Algorithm 3. Differently from multistage coders, this method, with its joint estimation of a short-term and a long-term predictor and the presence of a sparse residual, provides a synergistic one-step approach to speech coding. In synthesis, given a segment of speech, a way to encode the speech signal can be as follows:

1. Define the desired level of sparsity of the residual T and define the sensing matrix dimensionality accordingly $M = 4T$.

2. Perform n steps of the CS reweighted minimization process (Algorithm 5).
3. Factorize the prediction coefficients into a short-term and long-term predictor using the procedure in 3.2.
4. Quantize short-term and long-term predictors.
5. Select the T positions where the values of largest magnitude are located.
6. Solve the analysis-by-synthesis equation keeping only the T nonzero positions.
7. Quantize the residual.

We have again analyzed about one hour of clean speech taken from the TIMIT database. In order to obtain comparable results, the frame length is now $N = 160$ (20 ms). The order of the high order predictor in (E.21) is $K = 110$ (meaning that we can cover accurately pitch delays in the interval $[N_f + 1, K - N_f - 1]$, including the usual range for the pitch frequency $[70\text{Hz}, 500\text{Hz}]$). the fixed regularization parameter is $\gamma = 0.12$ and the defined level of sparsity is $T = 20$. Four iterations of the reweighting minimization process are performed, sufficient to reach convergence in all the analyzed frames. The orders of the short-term and long-term predictors obtained from the factorization of the high order predictor are $N_f = 10$ and $N_p = 1$, respectively. 25 bits are used to transparently encode the LSF vector, 7 bits are used to quantize the pitch period T_p and 6 bits to quantize the pitch gain g_p. The stability of the overall cascade is imposed by pole reflection on the short-term predictor, and by limiting the pitch gain to be less that unity. As for the residual, the quantizer normalization factor is logarithmically encoded with 6 bits while a 8 levels uniform quantizer is used to quantize the normalized amplitudes; the signs are coded with 1 bit per each pulse. The upper bound given by the information content of the pulse location ($\log_2 \binom{160}{20}$ bits) is used as an estimate of the number of bits used for distortionless encoding of the location. No perceptual weighting is performed in our case. The total number of bits per frame used are 202, producing a 10.1 kbps rate. We will compare this method (**SpLP**) with the AMR coder in the 10.2 kbps mode (**AMR102**) [44]. The results in terms of MOS (obtained through PESQ evaluation) and empirical computation time are shown in Table E.7 and demonstrate similar performances but with a more straightforward approach to coding than AMR. The CS formulation also helps to generally keep the problem solvable in reasonable time.

6.3 Noise Robustness

This study is motivated by the ability of a sparse coder to identify more effectively the features of the residual signal that are important for its reconstruction,

Table E.7: Comparison between the coding properties of the **AMR102** and the coder based on sparse linear prediction **SpLP**. A 95% confidence interval is given for each value.

METHOD	rate	MOS	t
AMR102	10.2 kbps	4.02±0.11	0.1±0.0
SpLP	10.1 kbps	4.13±0.13	1.2±0.1

Table E.8: Performances of **AMR102** and the coder based on sparse linear prediction (**SpLP**) for different values of SNR (white gaussian noise). A 95% confidence interval is given for each value.

METHOD	clean	30dB	20dB	10dB
AMR102	4.02±0.11	3.88±0.21	3.25±0.19	2.76±0.23
SpLP	4.13±0.13	3.94±0.15	3.52±0.14	3.21±0.19

discarding those which probably are a result of the noise. The traditional encoding formulation, based on minimum variance analysis and residual encoding through pseudo-random sequences (i.e., algebraic codes), makes the identification of these important features basically impossible and requires, for low SNRs, noise reduction in the preprocessing. Interestingly enough, sparse LP based coding appears to be quite robust in the presence of noise. An example of the different performances in terms of MOS for different SNR under additive white Gaussian noise is given in table E.8.

6.4 Subjective Assessment of Speech Quality

To further investigate the properties of our methods, we have conducted two MUSHRA listening tests [45] with 16 non-expert listeners. Ten speech clips were used in the listening test. In the first MUSHRA test we investigate what we have shown in 6.2, about the similarity in quality between the AMR coder and our method. In the second MUSHRA test the noise robustness of our method, discussed in 6.3, is proved. The test results are presented in Figure E.9 where the score 100 corresponds to "Imperceptible" and the score 0 corresponds to "Very annoying" according to the 6-grade impairment scale. From the results, we can see that our method does not affect greatly the quality of the signal, given that our method is conceptually much simpler and substantially less optimized compared to AMR. In clean condition the average score was 89 for **AMR102**, and 82 for **SpLP**. The most significant results though, are the one related to the coding of noisy signals. In particular, we can see from Figure E.9 that our

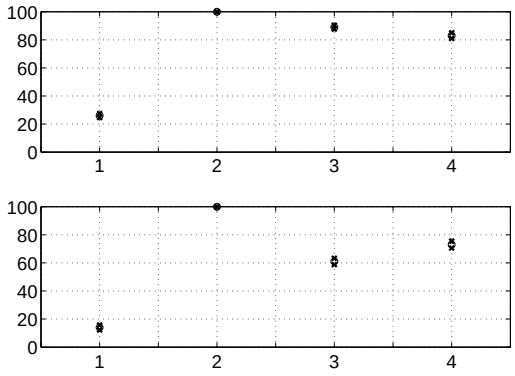

Fig. E.9: *MUSHRA test results. In the box above we show the results for clean speech and in the box below for speech corrupted by white noise (SNR=10dB). The four versions of the clips appear in the following order: Anchor, Hidden reference,* ***AMR102****, and* ***SpLP****. The anchor is the NATO standard 2400 bps LPC coding [46]. A 95% confidence interval is given for each value (upper and lower star).*

method scores considerably better than the AMR showing how a sparse encoding technique can be more effective in noise robust speech coding. In fact, in noisy conditions, the average score was 62 for **AMR102**, and 75 for **SpLP**.

7 Discussion

7.1 Stability

Although the predictor stability in the analysis of speech signals may not be required, it is a fundamental requirement in speech synthesis and coding, where an unstable filter can generate saturations in the synthesized speech. According to [41], the roots of the monic polynomial solution to the 1-norm minimization belongs to the numerical range of the shift operator matrix $\mathbf{B}$ used to generate $\mathbf{x}$ and $\mathbf{X}$ in (E.4), which, in turn, belongs to an open circular disk with radius $2\|\mathbf{B}\|_2 = 2$ and centered in the origin. This means that, for all the presented sparse linear predictors, we can define an upper bound on the maximum absolute value of the obtained roots. Even if this is an interesting result, it does not really help us create a minimization problem that is intrinsically stable (like it has been done, for example, in [47]) since the maximum absolute value for a root found in all our considered predictors is $\rho_{max} = 1.0259$.

The stability problem in (E.9) was already tackled in [9] by introducing the Burg method for prediction parameters estimation based on the least ab-

solute forward-backward error. In this approach, however, the sparsity is not preserved. This is mostly due to the decoupling of the main K-dimensional minimization problem in K one-dimensional minimization sub-problems. Therefore this method is suboptimal and produces results, as we have observed, somewhere in between those of the 2-norm and 1-norm approach. Also, the approach is only valid in (E.9) and not in all the other minimization schemes presented.

In the presented applications of sparse linear predictors, the percentage of unstable filters was found to be low (around 2%) and the instability "mild" ($\rho_{max} = 1.0259$). This suggests the use of a simple stability check and pole reflection. This approach, while leaving the magnitude of the spectrum unchanged would slightly modify both the coefficients of the predictor and the residual (affecting the sparsity properties), but not significantly enough to invalidate the presented results. Another approach to obtain stability, since $\rho_{max} = 1.0259$, could have been a mild bandwidth expansion of about 60 Hz, sufficient to bring the percentage of unstable filters to zero. However, this approach does not guarantee to find intrinsically stable solution. Also, even if no windowing has been applied in our scheme to the analyzed speech frame, we have observed that the use of a Hamming window eliminates completely the presence of unstable filters: this can be explained by the modified behavior of the analyzed speech signal segment. For example, we have observed that unstable filters are almost uniquely obtained when modeling the beginning of a strongly vocalized phoneme, where the waveform exhibits an "explosive" behavior. The impulse response of the all-pole filter used in the AbS equations will then mimic this behavior by not converging to zero but growing indefinitely. Therefore, the enhanced modeling properties given by our sparse linear predictors comes with a potential of instability. In particular, by properly modeling the behavior, we are able to find a better representation of a segment of speech that may be unstable, as has been observed in [48]. It is therefore interesting to continue to investigate into the subject and try to find a way to obtain, if not the optimal 1-norm solution, a good approximate that retains its properties at all time. Further work will concentrate on this issue.

7.2 Computational Cost

As for the computational cost, finding the solution of the overdetermined system of equations in (E.9) using a modern interior point algorithm [19] can be shown to be equivalent to solving around 20-30 least square problems. Nevertheless, implementing this procedure in an AbS coder, as done in Section 6.1, is shown to greatly simplify the search for the sparse approximation of the residual in a closed-loop configuration, without compromising the overall quality. Furthermore, in the case of (E.11), the advantage is that a one step approach is taken to calculate both the short-term and the long-term predictors while the encoding

of the residual is facilitated by its sparse characteristics.

The introduction of a compressed sensing formulation for the prediction problem has helped reduce dramatically the computational costs. An example of this can be seen in the coding scheme presented in 6.2. Retrieving $T = 20$ samples reduces the number of constraints of the minimization problem from 270 $(N + K)$ to 80 $(M = 4T)$. Since for each constraint we have a dual variable, by reducing the number of the constraints we also reduce the number of the dual variables [18]. In turn, the whole coding scheme, as shown empirically, is only about one order of magnitude more expensive than a 2-norm LP based coder, although with added improvements such as noise robustness and a fairly high conceptual simplicity.

7.3 Uniqueness

The minimization problems considered do not necessarily have a unique solution. In these rare cases with multiple solutions, due to the convexity of the cost function, we can immediately state that all the possible multiple solutions will still be optimal [18]. Viewing the non-uniqueness of the solution as a weakness is also arguable: in the set of possible optimal solutions we can probably find one solution that offers better properties for our modeling or coding purposes. A theorem to verify uniqueness is discussed in [49].

7.4 Frequency Domain Interpretation

The standard linear prediction method exhibits spectral matching properties in the frequency domain due to Parseval's theorem [2]:

$$\sum_{n=-\infty}^{\infty} |e(n)|^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} |E(e^{j\omega})|^2 d\omega. \tag{E.23}$$

It is also interesting to note that minimizing the squared error in the time domain and in the frequency domain leads to the same set of equations, namely the Yule-Walker equations [25]. To the best of our knowledge, the only relation existing between the time and frequency domain error using the 1-norm is the trivial Hausdorff-Young inequality [50]:

$$\sum_{n=-\infty}^{\infty} |e(n)| < \frac{1}{2\pi} \int_{-\pi}^{\pi} |E(e^{j\omega})| d\omega, \tag{E.24}$$

which implies that time domain minimization does not corresponds to frequency domain minimization. It is therefore difficult to say if the 1-norm based approach is always advantageous compared to the 2-norm based approach for spectral

modeling, since the statistical character of the frequency errors is not clear. However, the numerical results in Tables E.2, E.3 and E.4 clearly show better spectral modeling properties of the sparse formulation.

8 Conclusions

In this paper, we have given an overview of several linear predictors for speech analysis and coding obtained by introducing sparsity into the linear prediction framework. In speech analysis, the sparse linear predictors have been shown to provide a more efficient decoupling between the pitch harmonics and the spectral envelope. This translates into predictors that are not corrupted by the fine structure of the pitch excitation and offer interesting properties such as shift invariance and pitch invariance. In the context of speech coding, the sparsity of residual and of the high order predictor provides a more synergistic new approach to encode a speech segment. The sparse residual obtained allows a more compact representation, while the sparse high order predictor engenders joint estimation of short-term and long-term predictors. A compressed sensing formulation is used to reduce the size of the minimization problem, and hence to keep the computational costs reasonable. The sparse linear prediction based robust encoding technique provided a competitive approach to speech coding with a synergistic multistage approach and a slower decaying quality for decreasing SNR.

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, no. 4, pp. 561–580, April 1975.

[3] F. Itakura and S. Saito, "Analysis synthesis telephony based on the maximum likelihood method," *Rep. 6th Int. Congr. Acoustics*, pp. C17–C20, Paper C-5-5, 1968.

[4] A. El-Jaroudi and J. Makhoul, "Discrete All-Pole Modeling," *IEEE Trans. Signal Processing*, vol. 39, pp. 411–423, 1991.

[5] M. N. Murthi and B. D. Rao, "All-pole modeling of speech based on the minimum variance distortionless response spectrum," *IEEE Trans. Speech and Audio Processing*, vol. 8, pp. 221–239, 2000.

[6] L. A. Ekman, W. B. Kleijn and M. N. Murthi, "Regularized Linear Prediction of Speech," *IEEE Trans. Audio, Speech, Language Processing*, vol. 16, no. 1, pp. 65–73, 2008.

[7] H. Hermansky, H. Fujisaki, Y. Sato, "Spectral envelope sampling and interpolation in linear predictive analysis of speech," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 9, pp. 53–56, 1984.

[8] C.-H. Lee, "On Robust Linear Prediction of Speech," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. 36, No. 5, pp. 642–650, 1988.

[9] E. Denoël and J.-P. Solvay, "Linear prediction of speech with a least absolute error criterion," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 33, no. 6, pp. 1397–1403, 1985.

[10] J. Schroeder, R. Yarlagadda, "Linear predictive spectral estimation via the L_1 norm," *Signal Processing*, Vol. 17, No. 1, pp. 19–29, 1989.

[11] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, Kluwer Academic Publishers, 1993.

[12] 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, pp. 614–617, 1982.

[13] 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.

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

[15] J. Lansford and R. Yarlagadda, "Adaptive L_p approach to speech coding," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, pp. 335–338, 1988.

[16] M. N. Murthi and B. D. Rao, "Towards a synergistic multistage speech coder," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, pp. 369–372, 1998.

[17] P. Kabal and R. P. Ramachandran, "Joint Optimization of Linear Predictors in Speech Coders," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 642 – 650, May 1989.

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

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

[20] D. Giacobello, M. G. Christensen, J. Dahl, S. H. Jensen, and M. Moonen, "Sparse Linear Predictors for Speech Processing," *Proc. Interspeech*, pp. 1353–1356, 2008.

[21] 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.

[22] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen, "Speech Coding Based on Sparse Linear Prediction," *Proc. European Signal Proc. Conf.*, pp. 2524–2528, 2009.

[23] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen, "Enhancing sparsity in linear prediction of speech by iteratively reweighted 1-norm minimization," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2010.

[24] D. L. Donoho, "Compressed sensing," *IEEE Trans. on Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

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

[26] S. Nagarajah, "A generalized normal distribution," *Journal of Applied Statistics*, vol. 32, no. 7, pp. 685–694, 2005.

[27] D. L. Donoho and M. Elad, "Optimally sparse representation from overcomplete dictionaries via ℓ^1-norm minimization," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 100, no. 5, pp. 2197–2202, 2002.

[28] E. J. Candés, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5, pp. 877–905, 2008.

[29] S. Gazor and W. Zhang, "Speech probability distribution," *IEEE Sig. Proc. Letters*, vol. 10, no. 7, pp. 204–207, 2003.

[30] J. A. Cadzow, "Minimum ℓ_1, ℓ_1, and ℓ_∞ norm approximate solutions to an overdetermined system of linear equations," *Digital Signal Processing*, vol. 12, no. 4, pp. 524–560, 2002.

[31] P. Stoica and T. Söderström, "High Order Yule-Walker equations for estimating sinusoidal frequencies: the complete set of solutions," *Signal Processing*, vol. 20, pp. 257–263, 1990.

[32] 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.

[33] D. Wipf, S. Nagarajan, "Iterative reweighted ℓ_1 and ℓ_2 methods for finding sparse solutions," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 2, pp. 317–329, 2010.

[34] E. J. Candés, and M. B. Wakin, "An introduction to compressive sampling," *IEEE Sig. Proc. Magazine*, vol. 25, no. 2, pp. 21–30, 2008.

[35] T. V. Sreenivas and W. B. Kleijn, "Compressive sensing for sparsely excited speech signals," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 4125–4128, 2009.

[36] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen, "Retrieving sparse patterns using a compressed sensing framework: applications to speech coding based on sparse linear prediction," *IEEE Sig. Proc. Letters*, vol. 17, no. 1, pp. 103–106, 2010.

[37] 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., ch. 3, pp. 79–119, 1995.

[38] L. Scharf, *Statistical Signal Processing*, Addison-Wesley, 1991.

[39] M. G. Christensen, J. Østergaard, and S. H. Jensen, "On compressed sensing and its applications to speech and audio signals," in *Rec. Asilomar Conf. Sig., Sys., and Comp.*, 2009.

[40] W. B. Johnson and J. Lindenstrauss, "Extensions of Lipschitz mapping into Hilbert space," *Conf. in modern analysis and probability*, vol. 26, pp. 189–206, 1984.

[41] L. Knockaert, "Stability of linear predictors and numerical range of shift operators in normed spaces," *IEEE Trans. on Inf. Theory*, vol. 38, no. 5, pp. 1483–1486, 1992.

[42] A. D. Subramaniam, B. D. Rao, "PDF optimized parametric vector quantization of speech line spectral frequencies", *IEEE Trans. Speech and Audio Processing*, vol. 11, no. 2, 2003.

[43] K. K. Paliwal, B. S. Atal, "Efficient vector quantization of LPC parameters at 24 bits/frame," *IEEE Trans. Speech and Audio Processing*, vol. 1, no. 1, pp 3–14, 1993.

[44] "Adaptive Multi-Rate (AMR) speech codec; Transcoding functions," 3GPP TS 26.190, 2004.

[45] Method for the Subjective Assessment of Intermediate Quality Level of Coding Systems, ITU-R BS.1534-1 2003.

[46] NATO (unclassified), "Parameters and coding characteristics that must be common to assure interoperability of 2400 bps linear predictive encoded digital speech," Annex X to AC/302 (NBDS) R/2.

[47] C. Magi, J. Pohjalainen, T. Bäckström and P. Alku, "Stabilised weighted linear prediction," *Speech Communication*, vol. 51, no. 5, pp. 401–411, 2009.

[48] W. F. G. Mecklenbrauker, "Remarks on the minimum phase property of optimal prediction error filters and some related questions," *IEEE Sig. Proc. Letters*, vol. 5, no. 4, pp. 87–88, 1998.

[49] P. Bloomfield and W. Steiger, "Least absolute deviations curve-fitting", *SIAM J. on Scientific and Statistical Computing*, vol. 1, no. 2, pp. 290–301, 1980.

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

Paper F

Stable Solutions for Linear Prediction of Speech Based on 1-norm Error Criterion

D. Giacobello, M. G. Christensen, M. N. Murthi,
S. H. Jensen, and M. Moonen

This paper will be submitted to
IEEE Transactions on Audio, Speech, and Language Processing,
2010.

The layout has been revised.

Abstract

In linear prediction of speech, the 1-norm error minimization criterion has already been shown to provide a valid alternative to the 2-norm criterion for both analysis and coding of speech signals. However, unlike 2-norm minimization, the 1-norm minimization does not guarantee stability of the corresponding all-pole model and, in coding applications, this can generate saturations in the synthesized speech. In this paper we introduce two new methods to obtain intrinsically stable solutions to the 1-norm minimization problem. The first method is based on the reduction of the numerical range of the shift operator associated with the particular prediction problem considered. The second method is based on imposing a constraint, given by the alternative Cauchy bound, in the 1-norm error minimization. The methods are compared with two well known stable methods: the Burg algorithm, based on the 1-norm minimization of the forward and backward prediction error, and the iteratively reweighted 2-norm minimization. The evaluation gives proof of the effectiveness of the new methods, performing very similarly to traditional 1-norm based linear prediction in terms of modeling and coding behavior.

1 Introduction

Linear Prediction (LP) is widely used in a diverse range of speech processing algorithms for analysis, coding and recognition [1]. The traditional approach is to find the prediction coefficients through the 2-norm minimization of the difference between the predicted and observed signal. This works well when the excitation signal is i.i.d. Gaussian [2], however, when this assumption is not satisfied, problems arise. This is the case for voiced speech where the pitch excitation can be considered to be of quasi-periodic nature with spiky excitation. In this case, the approach based on the 1-norm minimization of the prediction error has shown to offer better modeling properties thanks to its ability to decouple the pitch excitation from the vocal tract transfer function [3].

The improved statistical fitting of the 1-norm minimization shows also to be beneficial in speech coding applications. In particular, seeing the 1-norm as a convex relaxation of the 0-norm, the minimization process will offer a residual that is sparser, providing tighter coupling between the multiple stages of time-domain speech coders, and thereby enabling more efficient coding [4]. Nevertheless, unlike those obtained through 2-norm minimization, the predictors obtained through 1-norm minimization are not intrinsically stable [5] and, in coding application, unstable filters may create problems, generating saturations in the synthesized speech.

The problem of stability in 1-norm LP was already tackled in [6] by intro-

ducing the Burg method for AR parameter estimation based on 1-norm forward and backward error minimization. However, in this approach the sparsity is not preserved [3]. We are therefore interested in finding new ways to determine stable solutions for the 1-norm LP problem that allow for a improved spectral modeling but also allowing more efficient coding.

The paper is organized as follows. In Section 2, we provide a brief review of linear prediction. In Section 3, the core of the paper, we introduce our two new methods to obtain intrinsically stable solution to the 1-norm minimization problem. In Section 4, we compare the spectral modeling and coding performances of the predictors. Finally, Section 5 concludes the paper.

2 Fundamentals of Linear Prediction

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

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

where $\{a_k\}$ are the prediction coefficients, $e(n)$ is the driving noise process (also referred to as prediction residual or excitation) and we assume that $x(n) = 0$ for $n < 1$ and $n > N$. The speech production model (F.1) in matrix form becomes:

$$\mathbf{x} = \mathbf{X}\mathbf{a} + \mathbf{e}. \tag{F.2}$$

The problem considered in this paper is 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 [7]:

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_p^p, \tag{F.3}$$

where

$$\mathbf{x} = \begin{bmatrix} x(N_1) \\ \vdots \\ x(N_2) \end{bmatrix}, \mathbf{X} = \begin{bmatrix} x(N_1-1) & \cdots & x(N_1-K) \\ \vdots & & \vdots \\ x(N_2-1) & \cdots & x(N_2-K) \end{bmatrix}, \tag{F.4}$$

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 $n > N$ [8]. We will consider the case $N_1 = 1$ and $N_2 = N + K$, equivalent, when $p = 2$, to the autocorrelation method:

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_2^2 = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{x}, \tag{F.5}$$

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

The case we would like to consider is when $p = 1$, which corresponds to minimizing the sum of absolute values:

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_1. \tag{F.6}$$

This formulation is relevant particularly in LP of voiced speech signals where the prediction residual is usually modeled by an impulse train. The 1-norm, intended as a convex relaxation of the 0-norm, will offer an approximate solution to the minimization of the cardinality, i.e., the sparsest prediction prediction residual. This translates into the ability of the predictor to preserve the structure of the underlying sparse pulse-like excitation. The spectral envelope will benefit from this by avoiding the over-emphasis on peaks generated in the effort to cancel the voiced speech harmonics [3, 6].

The 1-norm minimization criterion, is also equivalent to the ML estimator when the prediction error is assumed to be i.i.d. Laplacian:

$$\mathbf{a_{ML}} = \arg\max_{\mathbf{a}} f(\mathbf{x}|\mathbf{a}) = \arg\max_{\mathbf{a}}\{\exp(-\|\mathbf{x} - \mathbf{X}\mathbf{a}\|_1)\}. \tag{F.7}$$

This statistical interpretation is also meaningful, since it is well known that the distribution of speech samples is better described by a Laplacian distribution [9].

The minimization problem in (F.6) does not allow for a closed form solution and so a linear programming formulation is required [7]. In particular, interior point methods [10] have been proved to solve the minimization problem efficiently.

3 Methods for Obtaining Stable Solutions

3.1 Reducing the Numerical Range of the Shift Operator

First of all, we have to consider a more general prediction framework where the columns of the matrix obtained concatenating $\mathbf{x}$ and $\mathbf{X}$, defined in (F.4):

$$[\mathbf{x}|\mathbf{X}] = [\mathbf{x}_0\ \mathbf{x}_1\ \ldots\ \mathbf{x}_K] \in \mathbb{R}^{(N+K)\times(K+1)}, \tag{F.8}$$

can be generated via the formula:

$$\mathbf{x}_{k+1} = \mathbf{B}\mathbf{x}_k. \tag{F.9}$$

In this formulation, (F.6) is a particular case where:

$$\mathbf{x}_0 = [x_1\ x_2\ \ldots\ x_N\ 0\ \ldots\ 0]^T \in \mathbb{R}^{N+K}, \tag{F.10}$$

and $\mathbf{B}$ is an upper shift matrix of size $(N+K)\times(N+K)$:

$$\mathbf{B}=\begin{bmatrix} 0 & 0 & \cdots & \omega \\ 1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}. \tag{F.11}$$

In our case ω plays no useful role and thus we will set $\omega = 0$ (noncirculant shift matrix).

Let us now consider the p-norm LP problem (F.3), where the column $[\mathbf{x}|\mathbf{X}]$ are constructed using the formula in (F.9) where $\mathbf{B}$ is generalized to any matrix in $\mathbb{R}^{(N+K)\times(N+K)}$. It has been shown that, in this case, the roots $\{z_i\}$ of the monic polynomial solution to the p-norm minimization problem (F.3) belong to the numerical range $\eta_p(\mathbf{B})$ of the matrix $\mathbf{B}$, which, in turn, belongs to an open circular disk $\rho(\mathbf{B})$ of radius $2\|\mathbf{B}\|_2$ and center in the origin [11]. It is then clear that the roots of the predictor, obtained solving (F.6), will be contained in a closed circle of radius $2\|\mathbf{B}\|_2 = 2$. This result can be generalized for any shift matrix with nonzero entries different from the unity:

$$\mathbf{B}=\begin{bmatrix} 0 & 0 & \cdots & 0 \\ B_{2,1} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & 0 & B_{N+K,N+K-1} & 0 \end{bmatrix}. \tag{F.12}$$

In this case, the radius of the circle $\rho(\mathbf{B})$ that contain the numerical range $\eta_1(\mathbf{B})$ is defined as:

$$2\|\mathbf{B}\|_2 = 2\max|B_{i,i+1}|. \tag{F.13}$$

We will then change the nonzero values of $\mathbf{B}$ (and subsequently the construction of $[\mathbf{x}|\mathbf{X}]$), in order to reduce the radius of the circle containing $\eta_1(\mathbf{B})$ to be equal or less than one, therefore guaranteeing the stability of the linear predictor. In particular, having $\max|B_{i,j}| \leq 1/2$ will be sufficient for stability. We can also consider a more general formulation of the predictive scheme, where we apply a weighting $\mathbf{w} \in \mathbb{R}_+^{N+K}$ on the analyzed speech signal. The effect of the weighting can be moved to the shift matrix and the analyzed speech segment by defining:

$$\tilde{\mathbf{B}}=\begin{bmatrix} 0 & 0 & \cdots & 0 \\ w_2/w_1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & 0 & w_{N+K}/w_{N+K-1} & 0 \end{bmatrix}, \tag{F.14}$$

and

$$\tilde{\mathbf{x}}_0 = [w_1 x_1 \; w_2 x_2 \; w_N x_N \; 0 \; \dots \; 0]^T. \tag{F.15}$$

Constructing all the other columns of the new matrix $[\tilde{\mathbf{x}}|\tilde{\mathbf{X}}]$ using the relation in (F.9), the minimization problem (F.6) then becomes:

$$\min_{\mathbf{a}} \|\tilde{\mathbf{x}} - \tilde{\mathbf{X}}\mathbf{a}\|_1. \tag{F.16}$$

The 2-norm of the matrix can be defined as the maximum value of the entries of the shift matrix $|w_{n+1}/w_n|$ and the circle containing the roots of the predictor will have radius:

$$\rho(\mathbf{B}) = 2 \max \frac{w_{n+1}}{w_n}. \tag{F.17}$$

Knowing this we can construct a weighting vector that stabilizes the predictor. A smart way to choose the weights can be done using the method in [12] and [13], where the weight function in (F.14) is chosen based on the short-time energy (STE):

$$w_n = \sqrt{\sum_{i=0}^{M-1} x_{n-i-1}^2} \tag{F.18}$$

where M is the length of the STE window. The STE window tends to weight more those section of the speech signal which consist of samples of large magnitude providing robust signal selection especially for the analysis of voiced speech. Considering now the radius of the numerical range where the roots are contained (F.17), we can define the entries of the matrix $\tilde{\mathbf{B}}$ in (F.14) so that:

$$\tilde{B}_{i+1,i} = \begin{cases} (w_{i+1}/w_i) & \text{if} \quad (w_{i+1}/w_i) \leq 1/2, \\ 1/2 & \text{if} \quad (w_{i+1}/w_i) > 1/2. \end{cases} \tag{F.19}$$

Finally, we can solve our modified 1-norm problem in (F.16) obtaining an intrinsically stable solution. Clearly, the window, and thus the weights, can be chosen *ad libitum*, we will use the STE windowing that provides important signal selection properties to retrieve the underlying spiky structure for of the speech signal, as done in [13].

3.2 Constrained 1-norm Minimization

Let us now consider the univariate polynomial $A(z)$, corresponding to the solution of (F.6):

$$A(z) = 1 + \sum_{k=1}^{K} a_k z^{-k}, \tag{F.20}$$

so that $\mathbf{a} \in \mathbb{R}^K$. According to [14], the alternative Cauchy bound state that all zeros of (F.20) lie in the disk:

$$|z| \leq \lambda, \quad \text{where} \quad \lambda = \max\left\{1, \sum_{k=1}^{K} |a_k|\right\}. \tag{F.21}$$

This bound, refinement of the famous Cauchy bound [15], gives precious hints on how to modify the formulation of (F.6), so to guarantee a minimum phase solution. In particular, we can rewrite the problem as:

$$\min_{\mathbf{a}} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_1, \quad \text{s.t.} \quad \|\mathbf{a}\|_1 < 1, \tag{F.22}$$

where the constrained $\|\mathbf{a}\|_1 < 1$, according to (F.21), provides a sufficient (not necessary) condition for the zeros of (F.20) to belong to the open unit disk, and can be easily incorporated in the linear program to solve (F.6) [7]. Note that, if the constraint would have been $\|\mathbf{a}\|_1 \leq 1$ the zeros could be located on the the disk ($|z| \leq 1$). The strict constraint guarantees that $|z| < 1$.

3.3 Iteratively Reweighted 2-norm Minimization

A known method to obtain a minimum phase solution to the 1-norm minimization problem is based on iteratively reweighted 2-norm minimization [16]. The algorithm is shown in Algorithm 6. This method is guaranteed to output a polynomial with roots contained in the unit circle since the only difference is the projection in the weighted domain by the matrix $\mathbf{W}^i$, not changing the construction of $\mathbf{x}$ and $\mathbf{X}$, as discussed in Section 3.1. In [16] a proof that $\|\hat{\mathbf{r}}^{i+1}\|_2 \leq \|\hat{\mathbf{r}}^i\|_2$ (where $\hat{\mathbf{r}}^i = \mathbf{x} - \mathbf{X}\hat{\mathbf{a}}^i$) is provided, meaning that this is a descent algorithm. In Algorithm 6, the halting criterion can be chosen as either a maximum number of iterations or as a convergence criterion. The parameter $\epsilon > 0$ is used to avoid problems when a component of $\hat{\mathbf{r}}$ goes to zero. The weighting done by the square root of the inverse of the amplitude of the residual increases the influence of the small values in the residual while the influence of the large residual values decreases, consistently with the Laplacian probability density functions, for which (F.6) is the maximum likelihood approach.

3.4 Burg Method Based on 1-norm Minimization

This method, proposed in [6], stands as a generalization of the Burg method where the reflection coefficients of the lattice filter are obtaining by minimizing the 2-norm of the forward and backwards prediction error. In this case the 1-norm is minimized instead. The algorithm is shown in Algorithm 7. Once the K reflection coefficient are found, the prediction polynomial and the prediction

Algorithm 6 Iteratively Reweighted 2-norm Minimization

Inputs: speech segment $\mathbf{x}$
Outputs: predictor $\hat{\mathbf{a}}^i$, residual $\hat{\mathbf{r}}^i$
$i = 0$, initial weights $\mathbf{W}^{i=0} = \mathbf{I}$
while halting criterion false **do**
1. $\hat{\mathbf{a}}^i \leftarrow \arg\min_{\mathbf{a}} \|\mathbf{W}^i(\mathbf{x} - \mathbf{X}\mathbf{a})\|_2^2$
2. $\mathbf{W}^{i+1} \leftarrow \text{diag}\left(\left|\mathbf{x} - \mathbf{X}\hat{\mathbf{a}}^i\right| + \epsilon\right)^{-1/2}$
3. $i \leftarrow i + 1$
end while

Algorithm 7 1-norm Burg Method

Inputs: speech segment $\mathbf{x}$
Outputs: reflection coefficients $\{k_i\}$
Initialize forward $\mathbf{f}_0 = \mathbf{x}$ and backward $\mathbf{b}_0 = \mathbf{x}$ error
for $i = 1, \ldots, K$ **do**
1. $k_i \leftarrow \arg\min_{k_i} \|\mathbf{f}_{i-1} + k_i\mathbf{b}_{i-1}\|_1 + \|k_i\mathbf{f}_{i-1} + \mathbf{b}_{i-1}\|_1$
update forward error
2. $f_i(n) \leftarrow f_{i-1}(n) + k_i b_{i-1}(n-1)$
update backward error
3. $b_i(n) \leftarrow k_i f(n)_{i-1} + b_{i-1}(n-1)$
end for

error can be easily calculated. This method is also guaranteed to be stable since all the reflection coefficients obtained have amplitude less than one. A simple proof is shown in [6]. This method is however suboptimal due to the decoupling of the main K-dimensional minimization problem (F.6) in K one-dimensional minimization sub-problems. This is in contrast with all the other methods that try to find a minimum in the $K-$dimensional cost function.

4 Experimental Analysis

In this section, we analyze and compare the performances of the stable predictors presented in the previous section with traditional 2-norm LP and 1-norm LP. An overview of the methods compared is shown in Table F.1. In the case of 1-norm LP, a stability check takes place once the solution is obtained, the stabilization is performed through pole reflection when the filter is unstable. Notice that pole reflection is the only way to have the amplitude of the frequency response of the all-pole model that is exactly the same as the one of the unstable filter. In all other method, no stability check is performed and the predictor is calculated directly with the intrinsically stable method.

Table F.1: Description of the different prediction methods compared in our evaluation.

METHOD	DESCRIPTION
LP2	Traditional 2-norm minimization (F.5) with 10Hz bandwidth expansion ($\gamma = 0.996$) and Hamming windowing.
LP1	Traditional 1-norm minimization (F.6). Stability is imposed by pole reflection if unstable. No windowing is performed.
STW	Stable 1-norm minimization through reduction of the numerical range of the shift operator (F.16). The weigths in (F.14) and (F.15) are chosen from the STE (F.18).
CS1	Constrained 1-norm minimization as shown in (F.22). No windowing is performed.
BU1	Burg method based on the 1-norm minimization of forward and backward error (as shown in Algorithm 7). No windowing is performed.
RW2	Reweighted 2-norm minimization (as shown in Algorithm 6). No bandwidth expansion is performed. No windowing is performed.

4.1 Modeling Performances

In this section we analyze the modeling performances of the predictors in case of voiced speech. The experimental analysis has been done on 5,000 frames of length $N = 40$ (5 ms) of clean voiced speech coming from several different speakers with different characteristics (gender, age, pitch, regional accent) taken from the TIMIT database, downsampled at 8 kHz. As a reference, we used the envelope obtained through a cubic spline interpolation between the harmonics peaks of the logarithmic periodogram. This method was presented in [17] and provides an approximation of the vocal tract transfer function, "cleaned" from the fine structure belonging to the pitch excitation. We then calculate the log spectral distortion between our reference envelope $S_{int}(\omega)$ and the estimated AR

Table F.2: Average spectral distortion for the considered methods in the unquantized case SD_m and quantized case SD_q for different prediction orders K. A 95% confidence intervals is given for each value.

METHOD	K	SD_m	SD_q
LP2	10	1.97±0.03	2.95±0.09
	12	1.98±0.05	2.92±0.12
LP1	10	1.78±0.01	2.53±0.02
	12	1.61±0.01	2.31±0.04
STW	10	1.71±0.02	2.47±0.01
	12	1.52±0.01	2.19±0.09
CS1	10	1.88±0.01	2.64±0.01
	12	1.65±0.01	2.22±0.01
BU1	10	1.91±0.06	2.71±0.09
	12	1.84±0.11	2.59±0.10
RW2	10	1.83±0.01	2.51±0.02
	12	1.69±0.03	2.37±0.05

model $S(\omega, \mathbf{a})$ as:

$$\mathrm{SD}_m = \sqrt{\frac{1}{2\pi}\int_{-\pi}^{\pi} \left[10\log_{10} S_{int}(\omega) - 10\log_{10} S(\omega, \mathbf{a})\right]^2 d\omega}. \tag{F.23}$$

In general, the linear predictors obtained through 1-norm minimization provide smoother all-pole models of the vocal tract, therefore more robust to quantization. We will then also compare the log spectral distortion between our reference envelope $S_{int}(\omega)$ and the quantized AR model $S(\omega, \hat{\mathbf{a}})$:

$$\mathrm{SD}_q = \sqrt{\frac{1}{2\pi}\int_{-\pi}^{\pi} \left[10\log_{10} S_{int}(\omega) - 10\log_{10} S(\omega, \hat{\mathbf{a}})\right]^2 d\omega}. \tag{F.24}$$

The quantizer used is the one presented in [18], with the number of bits fixed at 20 for the different prediction orders, providing in all the method presented a *transparent coding* [19]. A critical analysis of the results shows how 1-norm based LP (**LP1**) offers substantially better modeling of the envelope than traditional LP (**LP2**). All the other methods achieve similar performances to **LP1**, nevertheless **STW** offers even better modeling performances, thanks also to the choice of weights. It should be noted that **CS1** increase its performances considerably from order $K = 10$ to $K = 12$. This is due to the stringent constraint on the prediction coefficients ($\|\mathbf{a}\|_1 < 1$) that necessarily needs a larger K in order to grasp the spectral information as well as the other methods.

Table F.3: Comparison between the considered predictors in coding application through multipulse encoding (T pulses). A 95% confidence interval is given for each value.

METHOD	T	$\hat{\mathbf{a}}$	SSNR
LP2	5	19	14.1±3.2
	10	19	19.1±2.9
LP1	5	18	15.3±2.1
	10	18	21.1±1.7
STW	5	17	14.9±1.6
	10	17	20.6±0.9
CS1	5	15	13.9±1.9
	10	15	19.2±1.5
BU1	5	19	14.2±0.9
	10	19	19.4±0.4
RW2	5	21	15.2±1.2
	10	21	20.9±1.7

4.2 Coding Performances

The second objective is to extend these algorithms to the contest of speech coding. The experimental analysis has been conducted on about one hour of clean speech (both voiced and unvoiced) coming from several different speakers with different characteristics (gender, age, pitch, regional accent) taken from the TIMIT database, re-sampled at 8 kHz. We propose a simple scheme to evaluate the coding performances of the proposed prediction method. A 10^{th} order predictive analysis is first done on a segment of speech of $N = 40$. Then a multipulse encoding procedure [20] is performed to code T pulses in the residual, with $T = 5$ and $T = 10$. Multipulse encoding is used to obtain a sparse residual, rather than a pesudo-random one like algebraic codes, therefore matching the characteristics of the 1-norm minimization. In Table F.3, we present the results in terms of Segmental SNR and number of bits necessary to transparently encode the predictor ($\hat{\mathbf{a}}$) using the method presented in [18]. The best performances in coding are achieved by **RW2**, consistently with the "guidance" in the reweighting algorithm given by the square root of inverse of the residual amplitude, although it requires a larger number of bits to transparently encode the predictor. As mentioned in the introduction, **BU1** does not preserve the sparsity of the residual and the coding characteristics of the 1-norm, performing very similarly to the 2-norm. The methods we have introduce seem to have good coding characteristics. The very smooth spectrum obtain with **CS1** allows considerably less bits than any other methods to achieve transparent coding of the

prediction coefficients, achieving performances comparable to **LP2** and **BU1**. **STW** performs just slightly worse than **RW2**, but with a significant saving in the bit budget of the predictor.

5 Conclusions

This paper has presented two new methods for finding intrinsically stable solution to the 1-norm linear prediction problem. The stable methods introduced, one based on constrained 1-norm minimization and one of the reduction of the numerical range of the shift operator, have both shown as valid alternatives to the original 1-norm linear prediction problem preserving the well known properties of the 1-norm minimization criterion. The experimental analysis has shown that both methods have attractive performances for the analysis and coding of speech signals offering comparable performances to the original problem without a significant increase in complexity. This two methods have also shown to be slightly better in modeling performance compared to two well-known stable 1-norm methods: the 1-norm Burg method and the reweighted 2-norm.

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, no. 4, pp. 561–580, 1975.

[3] 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 Trans. Speech, Audio and Language Processing*, 2010.

[4] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen "Speech coding based on sparse linear prediction", *Proc. European Signal Processing Conference*, 2009.

[5] W. F. G. Mecklenbrauker, "Remarks on the minimum phase property of optimal prediction error filters and some related questions," *IEEE Signal Processing Letters*, vol. 5, no. 4, pp. 87–88, 1998.

[6] 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.

[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] S. Gazor and W. Zhang, "Speech probability distribution," *IEEE Sig. Proc. Letters*, vol. 10, no. 7, pp. 204–207, 2003.

[10] S. J. Wright, *Primal-Dual Interior-Point methods*, SIAM, 1997.

[11] L. Knockaert, "Stability of linear predictors and numerical range of shift operators in normed spaces," *IEEE Trans. on Information Theory*, vol. 38, no. 5, pp. 1483–1486, 1992.

[12] C. Ma, Y. Kamp and L. F. Willems, "Robust signal selection for linear prediction analysis of voiced speech," *Speech Communication*, vol. 12, no. 1, pp. 69–81, 1993.

[13] C. Magi, J. Pohjalainen, T. Bäckström, and P. Alku, "Stabilized weighted linear prediction," *Speech Communication*, vol. 51, pp. 401–411, 2009.

[14] H. P. Hirst and W. T. Macey, "Bounding the roots of polynomials," *The College Mathematics Journal*, vol. 18, no. 4, pp. 292–295, 1997.

[15] A. L. Cauchy, *Exercise de mathematique*, Oeuvres 2, vol. 19, 1829.

[16] Y. Li, "A globally convergent method for L_p problems," *SIAM J. Optimization*, vol. 3, no. 3, pp. 609–629, 1993.

[17] L. A. Ekman, W. B. Kleijn, and M. N. Murthi, "Regularized Linear Prediction of Speech," *IEEE Trans. Audio, Speech, Language Processing*, vol. 16, no. 1, pp. 65–73, 2008.

[18] A. D. Subramaniam and B. D. Rao, "PDF optimized parametric vector quantization of speech line spectral frequencies," *IEEE Trans. Speech and Audio Processing*, vol. 11, no. 2, pp. 87–89, 2003.

[19] K. K. Paliwal and B. S. Atal, "Efficient vector quantization of LPC parameters at 24 bits/frame," *IEEE Trans. Speech and Audio Processing*, vol. 1, no. 1, pp 3–14, 1993.

[20] 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. Acoustics, Speech, Signal Processing*, vol. 7, pp. 614–617, 1982.

Paper G

High-Order Sparse Linear Predictors for Audio Processing

D. Giacobello, T. van Waterschoot, M. G. Christensen,
S. H. Jensen, and M. Moonen

This paper has been accepted for publication in
Proceedings of the 18th European Signal Processing Conference (EUSIPCO),
2010.

The layout has been revised.

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., [6–9]). 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]).

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:

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

[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.

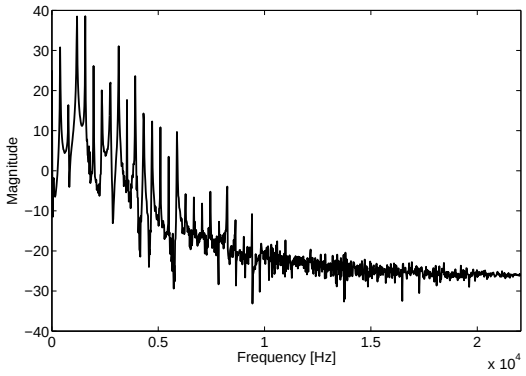

Fig. G.1: *Magnitude spectrum of the monophonic audio signal of Example 1.*

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.

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 G.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 G.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:

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

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:

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

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

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

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

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 (G.1):

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

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 G.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 $F(z)$.

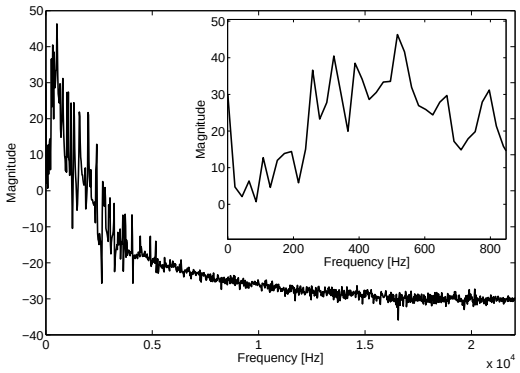

Fig. G.2: *Magnitude spectrum of the polyphonic audio signal of Example 3. In the smaller frame, we show a detail of the frequency range* $[0, 800]$ *Hz where the first harmonics of each of the four monophonic signals are located (*$f_{0,n} = \{258.4, 323.0, 387.6, 516.8\}$*).*

2.3 Polyphonic Audio Signals

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

$$x(n) = \sum_{m=1}^{M} \left(\sum_{q=1}^{Q_m} \alpha_{m,q} \cos(q\omega_{0,m} n + \phi_{m,q}) \right) + r(n), \qquad n = 1, \ldots, L \tag{G.6}$$

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 G.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 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 (G.2), will have to be modeled by:

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

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:

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

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 (G.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:

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

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:

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

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:

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

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*:

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

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 (G.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. (G.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 (G.1) is harmonic, while the polyphonic signal model in (G.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 (G.4) and for the polyphonic audio (G.9), we use the minimization problem in (G.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:

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

The high-order LP in (G.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 (G.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 (G.12), a number of spurious spectral peaks start 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 (G.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].

Monophonic Audio Signal

The frequency response of the filters is shown in Figure G.4 while the two predictors are shown in Figure G.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. G.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.

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 G.5. For the sake of brevity the predictor structure is not shown.

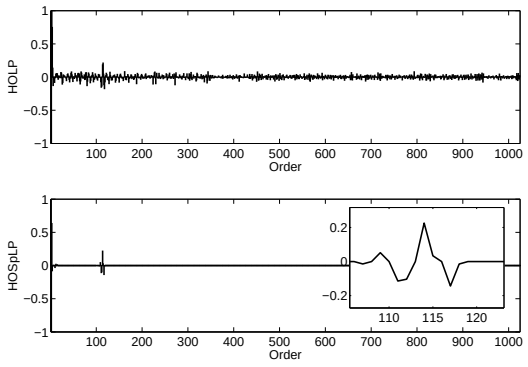

Fig. G.3: *High-order 2-norm LP (HOLP, above) and high-order sparse LP (HOSpLP, below) for a monophonic audio signal. A detail of the coefficients of order 105-125 is shown in the frame. The number of nonzero samples in the sparse predictor is 25.*

Polyphonic Audio Signal

The frequency response of the filters is shown in Figure G.6 while the two predictors are shown in Figure G.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 (G.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 G.1 we describe the $\mathbf{\Delta}$SFM'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-predictor, while performing with

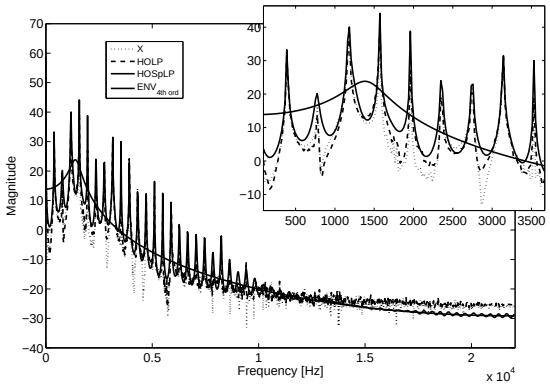

Fig. G.4: *Monophonic audio signal. Frequency response for the all-pole high-order 2-norm LP (HOLP), high-order sparse LP (HOSpLP) and the 4th order smooth spectral envelope (ENV). A detail of the first nine harmonics and the predictors modeling behavior is shown in the smaller frame.*

Table G.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).

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

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 1/100th 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

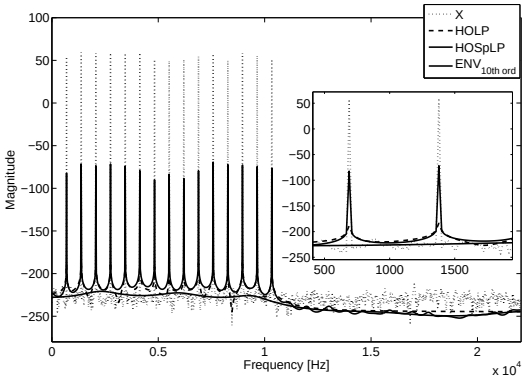

Fig. G.5: *Synthetic sum of sinusoids. Frequency response for the all-pole high-order 2-norm LP (HOLP), high-order sparse LP (HOSpLP) and the 10th order smooth spectral envelope (ENV). A detail of the first two harmonics and the predictors behavior is shown in the smaller frame.*

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.

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.

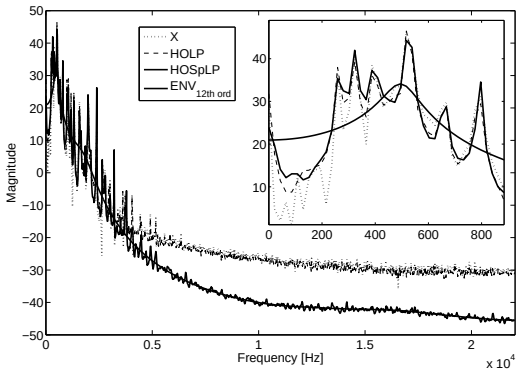

Fig. G.6: *Polyphonic audio signal. Frequency response for the all-pole high-order 2-norm LP (HOLP), high-order sparse LP (HOSpLP) and the 12th order smooth spectral envelope (ENV). A detail of the first four harmonics (each belonging to a different signal) and the predictors behavior is shown in the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

[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.

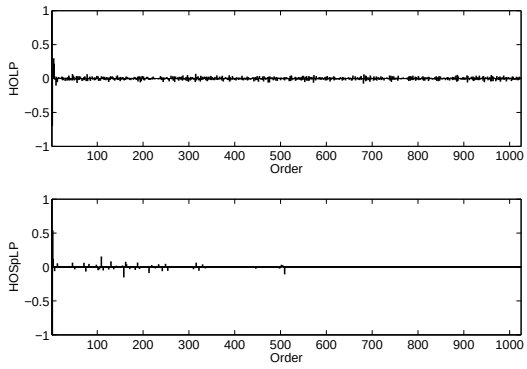

Fig. G.7: *High-order 2-norm LP (HOLP, above) and high-order sparse LP (HOSpLP, below) for polyphonic audio signal. The number of nonzero samples in the sparse predictor is 53.*

[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.

[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. 863–875, 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.

Paper H

Retrieving Sparse Patterns Using a Compressed Sensing Framework: Applications to Speech Coding Based on Sparse Linear Prediction

D. Giacobello, M. G. Christensen, M. N. Murthi,
S. H. Jensen, and M. Moonen

This paper has been published in
IEEE Signal Processing Letters,
vol. 17, no. 1, pp. 103–106, 2010.

The layout has been revised.

Abstract

Encouraged by the promising application of compressed sensing in signal compression, we investigate its formulation and application in the context of speech coding based on sparse linear prediction. In particular, a compressed sensing method can be devised to compute a sparse approximation of speech in the residual domain when sparse linear prediction is involved. We compare the method of computing a sparse prediction residual with the optimal technique based on an exhaustive search of the possible nonzero locations and the well known Multi-Pulse Excitation, the first encoding technique to introduce the sparsity concept in speech coding. Experimental results demonstrate the potential of compressed sensing in speech coding techniques, offering high perceptual quality with a very sparse approximated prediction residual.

1 Introduction

Finding a sparse approximation of the prediction residual in Linear Predictive Coding (LPC) has been an active field of research for the past thirty years. A significant result was found with the introduction of the Multi-Pulse Excitation (MPE) technique [1] providing a suboptimal solution to a problem of combinatorial nature. The purpose of this scheme is to find a prediction residual approximation with a minimum number of nonzero elements, still offering a high perceptual quality. MPE quickly evolved to Code Excited Linear Prediction (CELP), where the best residual approximation is selected from a codebook populated with pseudo-random white sequences. This choice was motivated by the statistic of the residual, ideally a sequence of i.i.d. Gaussian samples (due to the use of 2-norm minimization in the LP analysis).

In our recent work, we have utilized recent developments in convex optimization to define a new synergistic predictive framework that aims for a sparse prediction residual rather than the usual minimum variance residual [2, 3]. We have also shown that MPE techniques are better suited in this framework for finding a sparse approximation of the residual. Considering that MPE is itself a sub-optimal approach to modeling prediction residuals, a natural question is whether one can improve upon the performance of MPE by moving towards a more optimal approach of capturing prediction residuals without increasing complexity. Recent work on sparse solutions to linear inverse problems, commonly referred to as compressive sensing (CS), should be able to provide methods for tackling such issues [4]. While CS has been mainly applied to signals such as images with a natural underlying sparse structure, CS methods also seem to be appropriate for signals that are almost sparse, or for which sparsity is imposed [5]. Consequently, one expects that CS methods can be utilized to esti-

mate a sparse residual within a suitably modeled predictive coding framework. In [6] the authors examined the use of CS within speech coding, resulting in a restricted approach in which a codebook of impulse response vectors is utilized in tandem with an orthonormal basis. In [7], one can find a CS formulation of sinusoidal coding of speech.

In this paper, we examine CS within predictive coding of speech. In contrast to the work in [6], we do not utilize a codebook of impulse response vectors, and instead examine the more familiar approach to predictive coding in which the impulse response matrix is specified. In particular, we demonstrate how a CS formulation utilizing the Least Absolute Shrinkage and Selection Operator (LASSO [8]) method allows for a trade-off between the sparsity of the residual and the waveform approximation error. Moreover, this CS approach leads to a reduction in complexity in obtaining sparse residuals, moving closer to the optimal 0-norm solution while keeping the problem tractable through convex optimization tools and projection onto a random basis. In addition, this paper also shows the successful extension of the CS formulation to the case where the basis is not orthogonal, a case which is rarely examined in the CS literature. In simulations, the CS-based predictive coding approach provides better speech quality than that of MPE-based methods at roughly the same complexity.

The paper is structured as follows. In Section 2 we briefly review the general CS theory. In Section 3 we introduce the CS formulation for the case of speech coding, providing some significant results in Section 4. Section 5 will then conclude our work.

2 Compressed Sensing Principles

Compressed sensing (CS) has arguably represented a shift in paradigm in the way we acquire, process and reconstruct signals. In essence, CS exploits prior knowledge about the sparsity of a signal $\mathbf{x}$ in a linear transform domain in order to develop efficient sampling and reconstruction. Let $\mathbf{x} \in \mathbb{R}^N$ be the signal for which we would like to find a sparse representation and $\mathbf{\Psi} = \{\psi_1, \ldots, \psi_N\}$ be the orthonormal basis (or *orthobasis*). Considering the expansion of $\mathbf{x}$ onto the basis $\mathbf{\Psi}$ as:

$$\mathbf{x} = \mathbf{\Psi}\mathbf{r} = \sum_{i=1}^{N} r_i \psi_i \tag{H.1}$$

where $\mathbf{r}$ is the vector of the scalar coefficients of $\mathbf{x}$ in the orthobasis. The assumption of sparsity means that only K coefficients, with $K << N$, of $\mathbf{r}$ are significant to represent $\mathbf{x}$. In particular, $\mathbf{x}$ is said to be K-sparse if only K nonzero samples in $\mathbf{r}$ are sufficient to represent $\mathbf{x}$ exactly.

In CS we do not observe the K-sparse signal $\mathbf{x}$ directly, instead we record

$M < N$ nonadaptive linear measurements:

$$\mathbf{y} = \mathbf{\Phi x} = \sum_{i=1}^{N} \phi_m(i)x(i), \quad 1 \leq m \leq M < N, \tag{H.2}$$

where $\mathbf{\Phi} \in \mathbb{R}^{M \times N}$ is a measurement matrix made up of random orthobasis vectors. CS theory states that we can reconstruct $\mathbf{x}$ (or, equivalently $\mathbf{r}$) accurately from $\mathbf{y}$ if $\mathbf{\Phi}$ and $\mathbf{\Psi}$ are incoherent ($\mu(\mathbf{\Psi}, \mathbf{\Phi}) \approx 1$, where $\mu(\mathbf{\Psi}, \mathbf{\Phi})$ is the coherence measure, the largest correlation between any two columns of the basis matrix and the random matrix). This property is easily achievable when the entries of the random matrix $\mathbf{\Phi}$ are i.i.d. Gaussian variables. In this case, the recovery works with high probability if M is in the order of $K \log(N)$ [9]. If the incoherence holds, the following linear program gives an accurate reconstruction with very high probability:

$$\min_{\mathbf{r} \in \mathbb{R}^N} \|\mathbf{r}\|_1 \quad \text{s.t.} \quad \mathbf{y} = \mathbf{\Phi\Psi r}, \tag{H.3}$$

where $\|\mathbf{r}\|_1 = (\sum_{n=1}^{N} |r(n)|)$ is the 1-norm and it is used as a convex relaxation of the so-called 0-norm, the cardinality of a vector.

A very interesting property of CS is that if $\mathbf{x}$ is not K-sparse (or, not exactly K-sparse), the quality of the recovered signal $\mathbf{r}$ (or, equivalently $\mathbf{x}$) is as good as if we were to select only the K largest values before the calculations, and measure them directly. To quote [9]:

> the reconstruction is nearly as good as that provided by an oracle which, with full and perfect knowledge about $\mathbf{r}$, would extract the K most significant pieces of information for us.

This important property, stated elegantly in [10], extends the use of CS to all kinds of signals for which we would like to find a sparse representation. In particular, it allows us to apply CS to signals where K is not defined by the signal $\mathbf{x}$ but by our "need" for sparsity, therefore allowing an approximation error:

$$\mathbf{e} = \mathbf{y} - \mathbf{\Phi\Psi r}. \tag{H.4}$$

The formulation in (H.4) will then become:

$$\min_{\mathbf{r} \in \mathbb{R}^N} \|\mathbf{r}\|_1 \quad \text{s.t.} \quad \|\mathbf{y} - \mathbf{\Phi\Psi r}\|_2^2 \leq \epsilon, \tag{H.5}$$

where ϵ is the bound for the approximation error. This inequality constrained convex problem can also be rewritten using Lagrange multipliers as:

$$\min_{\mathbf{r} \in \mathbb{R}^N} \|\mathbf{r}\|_1 + \gamma \|\mathbf{y} - \mathbf{\Phi\Psi r}\|_2^2. \tag{H.6}$$

This latest formulation, also called *Least Absolute Shrinkage and Selection Operator* (LASSO [8]), shows more clearly the robustness of CS to signals that are not necessarily sparse and in particular, the trade-off between the sparsity of $\mathbf{r}$ and the approximation error $\mathbf{e}(\gamma, \mathbf{r})$.

Summarizing, if we wish to perform CS, two main ingredients are needed: a domain where the analyzed signal is sparse and the sparsity of this signal. The domain is found through a linear transform while the level of sparsity can be either known or assumed. In the next section we will see how can we define the CS formulation in speech coding.

3 Compressed Sensing Formulation for Speech Coding

3.1 Definition of the Transform Domain

In speech coding, the transform domain where the representation is required to be sparse is the prediction residual. In our previous work, we have indeed found very few nonzero samples in the residual when sparse linear prediction is involved [2, 3]. Considering the simple case in which we would like to find a linear predictor $\mathbf{a}$ of order P that provides a sparse residual, the formulation becomes:

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}\in\mathbb{R}^P} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_1; \tag{H.7}$$

where:

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

and $\|\cdot\|_1$ is the 1-norm. The start and end points N_1 and N_2 can be chosen in various ways assuming that $x(n) = 0$ for $n < 1$ and $n > N$. An appropriate choice is $N_1 = 1$ and $N_2 = N + P$ (in the case of 2-norm minimization, this leads to the autocorrelation and to the Yule-Walker equations). The more tractable 1-norm is used as a linear programming relaxation of the sparsity measure, just like in (H.4). Given a prediction filter $\mathbf{a}$ the residual vector can be expressed as:

$$\mathbf{r} = \mathbf{A}\mathbf{x}, \tag{H.8}$$

where $\mathbf{A}$ is the $N \times N$ matrix that performs the whitening of the signal, constructed from the coefficients of the predictor $\mathbf{a}$ of order P [11].

Equivalently, we can write:

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{r} = \mathbf{H}\mathbf{r}, \tag{H.9}$$

where $\mathbf{H}$ is the $N \times N$ inverse matrix of $\mathbf{A}$ and is commonly referred to as the synthesis matrix [11] that maps the residual representation to the original speech domain. In practice, this inversion is not computed explicitly and $\mathbf{H}$ is constructed directly from the impulse response $\mathbf{h}$ of the all pole filter that corresponds to $\mathbf{a}$. Furthermore, the usual approach is to have $N + P$ columns in $\mathbf{H}$ bringing in the effects of P samples of the residual of the previous frame (the filter state/memory).

It is important to notice that the column vector $\mathbf{r}$ will be now composed of $N + P$ rows, but the first P elements belong to the excitation of the previous speech frame and therefore are fixed and do not affect the minimization process.

It is now clear that the basis vectors matrix is the synthesis matrix $\mathbf{\Psi} = \mathbf{H}$. We can now write:

$$\mathbf{x} = \sum_{i=1}^{K} r_{n_i} \mathbf{h}_{n_i}, \quad \{n_1, n_2, \ldots, n_K\} \subset \{1, \ldots, N + P\}. \tag{H.10}$$

where $\mathbf{h}_i$ represents the $i-$th column of the matrix $\mathbf{H}$. The formulation then becomes:

$$\hat{\mathbf{r}} = \arg \min_{\mathbf{r} \in \mathbb{R}^N} \|\mathbf{r}\|_1 + \gamma \|\mathbf{y} - \mathbf{\Phi H r}\|_2^2 \tag{H.11}$$

where $\mathbf{y} = \mathbf{\Phi x}$ is the speech signal compressed through the projection onto the random basis $\mathbf{\Phi}$ of dimension $M \times N$. The second term is now the 2-norm of the difference between the original speech signal and the speech signal with the sparse representation, projected onto the random basis. Assuming that:

$$\|\mathbf{y} - \mathbf{\Phi H r}\|_2^2 = \|\mathbf{\Phi}(\mathbf{x} - \mathbf{H r})\|_2^2 \approx \|\mathbf{x} - \mathbf{H r}\|_2^2, \tag{H.12}$$

the problem in (H.11) can now be seen as a trade-off between the sparsity in the residual vector and the accuracy of the new speech representation $\hat{\mathbf{x}} = \mathbf{H}\hat{\mathbf{r}}$. To ensure simplicity in the preceding and following derivations, we have assumed that no perceptual weighting is performed. The results can then be generalized for an arbitrary weighting filter.

An important aspect that should be taken into consideration is that, if the transformation matrix $\mathbf{\Phi}$ is not exactly orthogonal, such as in the case of $\mathbf{\Phi} = \mathbf{H}$, the recovery is still possible, as long as the incoherence holds ($\mu(\mathbf{\Phi}, \mathbf{H}) \approx 1$) [4].

3.2 Defining the Level of Sparsity

CS theory states that for a vector $\mathbf{x}$ of length N with sparsity level K ($K << N$), $M = O(K \log(N))$ random linear projections of $\mathbf{x}$ are sufficient to robustly (i.e., with overwhelming probability) recover $\mathbf{x}$ in polynomial time. With a proper random basis, so that $\mathbf{\Phi}$ and $\mathbf{H}$ are incoherent ($\mu(\mathbf{\Phi}, \mathbf{H}) \approx 1$) [12], as a rule of thumb, four times as many random samples as the number of non-zero sparse

samples should be used; therefore, we can simply choose $M = 4K$ [9]. It is now clear that the size of the random matrix $\mathbf{\Phi}$ depends uniquely on the sparsity level K that we expect in the residual vector. Now the question is how sparse do we expect the residual to be? An interesting case for the choice of K is obtained for voiced speech. In this case, the residual $\mathbf{r}$ is a train of impulses. Each impulse is separated by T_p samples, the pitch period of the voiced speech which is inversely proportional to the fundamental frequency f_0. It is now clear that K will depend on T_p; for a segment of voiced speech of length N, we can reasonably assume to find only N/T_p significant samples in the residual, belonging to the impulse train. A coarse estimation of the integer pitch period T_p can be easily obtained by an open-loop search on the autocorrelation function of the vector $\mathbf{x}$. Then the number of random projections sufficient for recovering $\mathbf{x}$ will be $M = 4\frac{N}{T_p}$. In the case of unvoiced speech the choice of K is not direct, however we can use a heuristic approach where $K = k$ is picked when the improvement in the accuracy of the representation between the choice of $K = k$ and $K = k + 1$ is negligible.

3.3 Similarities with Multi-Pulse Excitation

In Multi-Pulse Excitation (MPE) coders the prediction residual consists of K freely located pulses in each segment of length N. This problem is made impractical by its combinatorial nature and a suboptimal algorithm was proposed in [1] where the sparse residual is constructed one pulse at a time. Starting with a zero residual, pulses are added iteratively adding one pulse in the position that minimizes the error between the original and reconstructed speech. The pulse amplitude is then found in an Analysis-by-Synthesis (AbS) scheme. The procedure can be stopped either when a maximum fixed number of amplitudes is found or when adding a new pulse does not improve the quality. MPE provides an approximation to the optimal approach, when all possible combinations of K positions in the approximated residual of length N are analyzed, i.e.:

$$\hat{\mathbf{r}} = \arg\min_{\mathbf{r}\in\mathbb{R}^N} \|\mathbf{x} - \mathbf{H}\mathbf{r}\|_2^2 \quad \text{s.t.} \quad \|\mathbf{r}\|_0 = K. \tag{H.13}$$

The compressive sensing formulation in (H.11) can then be seen to approximate (H.13), finding a trade-off between the information content of the prediction residual and the quality of the synthesized speech.

4 Experimental Results

To evaluate our method, we have analyzed about one hour of clean speech coming from several different speakers with different characteristics (gender, age, pitch, regional accent) taken from the TIMIT database, re-sampled at 8 kHz. The order

Table H.1: Comparison between the sparse residual estimation methods. A 95% confidence intervals is given for each value.

METHOD	K	AS-SNR	MOS	t
OPT	10	21.2±3.1	3.25±0.13	343±5
	20	27.2±1.6	3.52±0.09	581±3
CS	10	20.6±2.6	3.13±0.16	0.3±0.1
	20	25.9±1.9	3.49±0.13	0.5±0.1
MPE	10	17.2±4.1	3.03±0.15	0.1±0.2
	20	20.3±3.2	3.22±0.12	0.9±0.3

of the sparse linear predictor is $P = 10$, the length of the speech frame is $N = 160$ (20 ms). Three methods are compared: the MPE, the CS based approach in (H.11) and the optimal combinatorial approach (OPT) in (H.13). For simplicity, no long-term pitch prediction is performed. In the CS formulation, the random matrix $\mathbf{\Phi}$ is populated with Gaussian samples with distribution $N(0, 1)$ and the size is chosen according to the level of sparsity we want to retrieve using the relation $M = 4K$. The regularization parameter γ is chosen as the point of maximum curvature of the $L-$curve, using the method presented in [13].

In Figure H.1, we present the unquantized results of the three methods in term of the normalized error $\|\mathbf{x} - \hat{\mathbf{x}}\|_2/\|\mathbf{x}\|_2$, with $\hat{\mathbf{x}} = \mathbf{H}\hat{\mathbf{r}}^{(K)}$ averaged over all frames, choosing different levels of sparsity K. It is clear that for $K > 10$, the CS solution performs similarly to the optimal solution. While for very few samples $K < 5$, the performance is comparable to that of MPE.

In the quantized case, we concentrate our experimental analysis for the two most significant cases ($K = 10$ and $K = 20$). The quantization process uses 20 bits to encode the predictor using 10 line spectral frequencies (providing transparent coding) using split vector quantization. A 3 bit uniform quantizer that goes from the lowest to the highest magnitude of the residual pulses is used to code the residual pulses; 5 bits are used to code the lowest magnitude and 2 bits are used to code the difference between the lowest and highest magnitudes. The signs are coded with 1 bit per each pulse. We postpone the efficient encoding of the positions to further investigation, for now we just use the information content of the pulse location $\log_2 \binom{N}{K}$ bits. The bit rate produced is respectively 5900 bits/s for $K = 10$ and 9500 bits/s for $K = 20$. In Table H.1, we present the results in terms of Average-Segmental SNR, MOS and empirical computational time t in elapsed CPU seconds of the three methods for the quantized case. It is now clear that the CS formulation achieves similar performances to the optimal case, in a computational time similar to that of MPE.

As mentioned in the previous section, the CS recovery seems also particularly

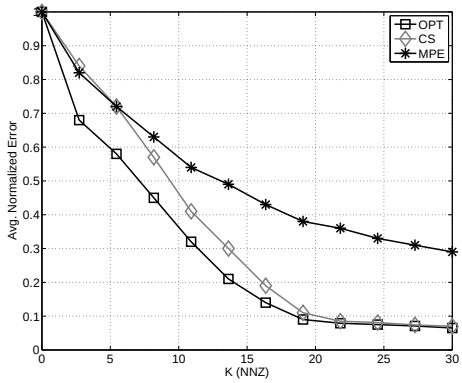

Fig. H.1: *Number of nonzero samples K versus the average normalized reconstruction error $\|\mathbf{x} - \hat{\mathbf{x}}\|_2 / \|\mathbf{x}\|_2$ for a speech segment $\mathbf{x}$. The values corresponding to $K > 30$ are not shown for clarity as the error rates converge to zero.*

attractive for the analysis and coding of stationary voiced signals. In Figure H.2 we see an example of CS recovery of pitch excitation. The open-loop pitch search gives us a coarse approximation of the pitch period of $T_p = 35$ ($f_0 \approx 229$Hz). We then impose $K = \lceil N/T_p \rceil = 10$ and $M = 40$, using the relation $M = 4K$. From the solution we take the $K = 10$ pulses with largest magnitude. We can clearly see that this kind of approximation works very well in the case of voiced speech, retrieving the K pulses belonging to the train of impulses with very high accuracy. The distance between pulses is then approximately T_p.

5 Conclusions

In this paper we have introduced a new formulation in the context of speech coding based on compressed sensing. The CS formulation based on LASSO has shown to provide an efficient approximation of the 0-norm for the selection of the residual allowing a trade-off between the sparsity imposed on the residual and the waveform approximation error. The convex nature of the problem, and its dimensionality reduction through the projection onto random basis, makes it also computationally efficient. The residual obtained engenders a very compact representation, offering interesting waveform matching properties with very few samples, making it an attractive alternative to common residual encoding procedures. The results obtained also show clearly that CS performs quite well when the basis are not orthogonal, as anticipated in some CS literature.

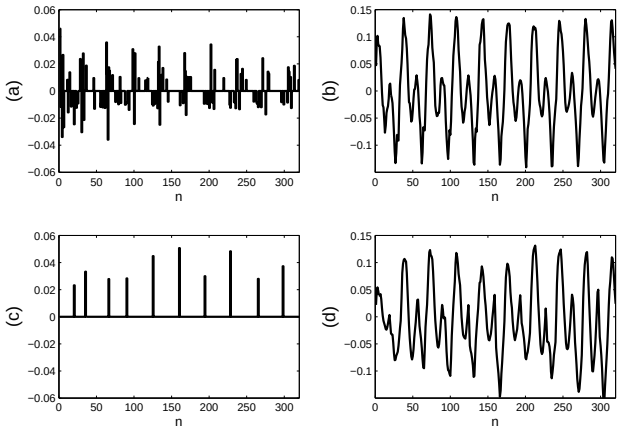

Fig. H.2: *Example of CS recovery of the pitch excitation for a segment of stationary voiced speech. In (a) we show the estimated excitation using (H.7) and in (b) the original speech segment. In (c) we show the CS recovered excitation with* $K = N/T_p = 320/35 \approx 10$ *and in (d) the reconstructed speech segment.*

References

[1] B. S. Atal and J. R. Remde, "A new model of LPC excitation for producing natural sounding speech at low bit rates," *Proc. IEEE ICASSP*, vol. 7, pp. 614–617, 1982.

[2] D. Giacobello, M. G. Christensen, J. Dahl, S. H. Jensen, and M. Moonen, "Sparse Linear Predictors for Speech Processing," *Proc. Interspeech*, pp. 1353–1356, 2008.

[3] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen, "Speech Coding Based on Sparse Linear Prediction," *Proc. EUSIPCO*, pp. 2524–2528, 2009.

[4] D. L. Donoho, "Compressed sensing," *IEEE Trans. on Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[5] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Com. on Pure and App. Math.*, vol. 57, pp. 1413–1457, 2004.

[6] T. V. Sreenivas and W. B. Kleijn, "Compressive sensing for sparsely excited speech signals," *Proc. ICASSP*, pp. 4125–4128, 2009.

[7] M. G. Christensen, J. Østergaard, and S. H. Jensen, "On compressed sensing and its applications to speech and audio signals," to appear in *Asilomar Conf.*, 2009.

[8] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Royal. Statist. Soc.*, vol. 58, no. 1, pp. 267–288, 1996.

[9] E. J. Candés, and M. B. Wakin, "An introduction to compressive sampling," *IEEE Sig. Proc. Mag.*, vol. 25, no. 2, pp. 21–30, 2008.

[10] E. J. Candés, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Com. on Pure and App. Math.*, vol. 59, no. 8, pp. 1207–1223, 2006.

[11] L. Scharf, *Statistical Signal Processing*, Addison-Wesley, 1991.

[12] E. J. Candés and J. Romberg, "Sparsity and incoherence in compressive sampling," *Inv. Problems*, no. 23(3), pp. 969–985.

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

Paper I

Re-estimation of Linear Predictive Parameters in Sparse Linear Prediction

D. Giacobello, M. N. Murthi, M. G. Christensen,
S. H. Jensen, and M. Moonen

This paper has been published in
Conference Record of the 43rd Asilomar Conference on Signals, Systems and Computers,
pp. 1770–1773, 2009

The layout has been revised.

Abstract

In this work, we propose a novel scheme to re-estimate the linear predictive parameters in sparse speech coding. The idea is to estimate the optimal truncated impulse response that creates the given sparse coded residual without distortion. An all-pole approximation of this impulse response is then found using a least square approximation. The all-pole approximation is a stable linear predictor that allows a more efficient reconstruction of the segment of speech. The effectiveness of the algorithm is proved in the experimental analysis.

1 Introduction

The most important speech coding paradigm in the past twenty years has been *Analysis-by-Synthesis* (AbS) [1, 2]. The name signifies analysis of the optimal parameters by synthesizing speech based on these. In other words, the speech encoder mimes the behavior of the speech decoder in order to find the best parameters needed. The usual approach is to first find the linear prediction parameters in a open-loop configuration then searching for the best excitation given certain constraints on it. This is done in a closed-loop configuration where the perceptually weighted distortion between the original and synthesized speech waveform is minimized. The conceptual difference between a quasi-white true residual and its approximated version, where usually sparsity is taken into consideration, creates a mismatch that can raise the distortion significantly. In our previous work we have defined a new synergistic predictive framework that reduces this mismatch by jointly finding a sparse prediction residual as well as a sparse high order linear predictor for a given speech frame [3]. Multipulse encoding techniques [4] have shown to be more consistent with this kind of predictive framework, offering a lower distortion with very few samples [5].

In this work, we propose a method to further reduce the mismatch between sparse linear predictor and approximated residual by re-estimating the linear predictive parameters. This paper is structured as follow. In Section 2, we introduce the coding method based on sparse linear prediction. In Section 3, we introduce the re-estimation procedure and in Section 4 we propose the results to validate our method. Finally, Section 5 concludes our work.

2 Speech Coding Based on Sparse Linear Prediction

In our previous work [3, 5], we have defined a synergistic new predictive framework that jointly finds a sparse prediction residual $\mathbf{r}$ as well as a sparse high

order linear predictor $\mathbf{a}$ for a given speech frame $\mathbf{x}$ as

$$\hat{\mathbf{a}}, \hat{\mathbf{r}} = \arg\min_{\mathbf{a}} \|\mathbf{r}\|_1 + \gamma\|\mathbf{a}\|_1, \quad \text{subject to} \quad \mathbf{r} = \mathbf{x} - \mathbf{X}\mathbf{a}; \tag{I.1}$$

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 $\|\cdot\|_1$ is the 1-norm defined as the sum of absolute values of the vector on which operates. The start and end points N_1 and N_2 can be chosen in various ways assuming that $x(n) = 0$ for $n < 1$ and $n > N$ [6]. The more tractable 1-norm $\|\cdot\|_1$ is used as a linear programming relaxation of the sparsity measure, often represented as the cardinality of a vector, the so-called 0-norm $\|\cdot\|_0$. This optimization problem can be posed as a linear programming problem and can be solved using an interior-point algorithm [7]. The choice of the regularization term γ is given by the L-curve where a trade-off between the sparsity of the residual and the sparsity of the predictor is found [8].

The sparse structure of the predictor allows a joint estimation of short-term and long-term predictor [9]:

$$A(z) \approx \tilde{A}(z) = F(z)P(z), \tag{I.2}$$

where $F(z)$ is the short-term predictor, commonly employed to remove short-term redundancies due to the formants, and $P(z)$ is the pitch predictor that removes the long-term redundancies. The sparse structure of the true residual $\hat{\mathbf{r}}$ allows for a quick and more efficient search of approximated residual $\tilde{\mathbf{r}}$ using sparse encoding procedure, where the approximated residual is given by a regular pulse excitation (RPE) [10]. The problem can be rewritten as:

$$\tilde{\mathbf{r}} = \arg\min_{\mathbf{r}} \|\mathbf{W}(\mathbf{x} - \tilde{\mathbf{H}}\mathbf{r})\|_2, \tag{I.3}$$

by imposing the RPE structure on $\tilde{\mathbf{r}}$:

$$\tilde{r}(n) = \sum_{i=0}^{N/S-1} \alpha_i \delta(n - iS - s) \quad s = 0, 1, \ldots, S-1, \tag{I.4}$$

where α_i are the amplitudes $\delta(\cdot)$ is the Kronecker delta function, N/S are the number of pulses and S is the spacing; only S different configurations of the positions are allowed (s is the shift of the residual vector grid). In (I.3), $\mathbf{W}$ is the perceptual weighting matrix, $\tilde{\mathbf{H}}$ is the $(N) \times (K+N)$ synthesis matrix whose $i-$th row contains the elements of index $[0, K+i-1]$ of the truncated impulse

response $\tilde{\mathbf{h}}$ of the combined prediction filter $\tilde{A}(z) = F(z)P(z)$:

$$\tilde{\mathbf{H}} = \begin{bmatrix} \tilde{h}_K & \cdots & \tilde{h}_0 & 0 & 0 & \cdots & 0 \\ \tilde{h}_{K+1} & \ddots & \ddots & \ddots & 0 & 0 & 0 \\ \vdots & \ddots & \ddots & \cdots & \tilde{h}_0 & 0 & 0 \\ \tilde{h}_{K+N-2} & \ddots & \ddots & \cdots & \tilde{h}_1 & \tilde{h}_0 & 0 \\ \tilde{h}_{K+N-1} & \tilde{h}_{K+N-2} & \cdots & \cdots & \tilde{h}_2 & \tilde{h}_1 & \tilde{h}_0 \end{bmatrix}. \tag{I.5}$$

and $\mathbf{r}$ is composed of the previous residual samples $\tilde{\mathbf{r}}_-$ (the filter memory, already quantized) and the current $\tilde{\mathbf{r}}$ that has to be estimated:

$$\mathbf{r} = \begin{bmatrix} \tilde{\mathbf{r}}_-^T & \tilde{\mathbf{r}}^T \end{bmatrix}^T = [\tilde{r}_{-K}, \cdots, \tilde{r}_{-2}, \tilde{r}_{-1}, \tilde{r}_0, \tilde{r}_1, \tilde{r}_2, \cdots, \tilde{r}_{N-1}]^T. \tag{I.6}$$

In the end a segment of speech can be represented by the sparse predictor $\tilde{A}(z)$ and its approximated excitation $\tilde{\mathbf{r}}$.

3 Re-estimation of the Predictive Parameters

To ensure simplicity in the following derivations, let us assume that no perceptual weighting is performed ($\mathbf{W} = \mathbf{I}$). The results can then be generalized for an arbitrary $\mathbf{W}$. The problem in (I.3) is now just a waveform matching problem. The interesting thing is that, once found a proper sparse excitation, we can re-estimate the matrix $\mathbf{H}$ and therefore the impulse response $\mathbf{h}$ by posing it as a convex optimization problem:

$$\hat{\mathbf{H}} = \arg\min_{\mathbf{H}} \|(\mathbf{x} - \mathbf{H}\tilde{\mathbf{r}})\|_2 \rightarrow \hat{\mathbf{h}} = \arg\min_{\mathbf{h}} \|(\mathbf{x} - \tilde{\mathbf{R}}\mathbf{h})\|_2 \tag{I.7}$$

where:

$$\tilde{\mathbf{R}} = \begin{bmatrix} \tilde{r}_0 & \cdots & \tilde{r}_{-K} & 0 & 0 & \cdots & 0 \\ \tilde{r}_1 & \ddots & \ddots & \ddots & 0 & 0 & 0 \\ \vdots & \ddots & \ddots & \cdots & \ddots & 0 & 0 \\ \tilde{r}_{N-1} & \ddots & \ddots & \cdots & \ddots & \tilde{r}_{-K} & 0 \\ \tilde{r}_N & \tilde{r}_{N-1} & \cdots & \cdots & \cdots & \tilde{r}_{-K+1} & \tilde{r}_{-K} \end{bmatrix}. \tag{I.8}$$

where $\{\tilde{r}_{-K}, \ldots, \tilde{r}_{-1}\}$ is the past excitation (belonging to the previous frame). The problem (I.7) allows for a closed form solution when the 2-norm is employed in the minimization:

$$\hat{\mathbf{h}} = \mathbf{h}_{opt} = \tilde{\mathbf{R}}^T(\tilde{\mathbf{R}}\tilde{\mathbf{R}}^T)^{-1}\mathbf{x}. \tag{I.9}$$

Because the matrix $\tilde{\mathbf{R}}^T(\tilde{\mathbf{R}}\tilde{\mathbf{R}}^T)^{-1}$ in (I.9) is the pseudo-inverse $\tilde{\mathbf{R}}^+$ of $\tilde{\mathbf{R}}$, the new $\mathbf{h}_{opt}$ is then the optimal truncated impulse response that matches the given sparse residual:

$$\|\mathbf{x} - \tilde{\mathbf{R}}\mathbf{h}_{opt}\|_2 = 0. \tag{I.10}$$

It is therefore clear that the optimal sparse linear predictor $A(z)$ is the one that has $\mathbf{h}_{opt}$ as truncated impulse response. The problem now is that the impulse response will include both short-term and long-term contribution. We can split the two contribution and perform a two step optimization.

Assuming $\mathbf{h}_f$ the impulse response of the short-term predictor $1/F(z)$ and $\mathbf{h}_p$ the impulse response of the long-term predictor $1/P(z)$, we can rewrite the problem in (I.7) as:

$$\hat{\mathbf{H}}_f, \hat{\mathbf{H}}_p = \arg\min_{\mathbf{H}_f, \mathbf{H}_p} \|(\mathbf{x} - \mathbf{H}_f\mathbf{H}_p\tilde{\mathbf{r}})\|_2. \tag{I.11}$$

We can then proceed with the re-estimation of the impulse response of the short-term predictor by solving the problem:

$$\hat{\mathbf{h}}_f = \arg\min_{\mathbf{h}_f} \|(\mathbf{x} - (\mathbf{H}_p\tilde{\mathbf{R}})\mathbf{h}_f)\|_2, \tag{I.12}$$

and then find the predictor that approximates $\hat{\mathbf{h}}_f$. The predictor $A(z) = 1 + \sum_{k=1}^{Q} a_k z^{-k}$ can then just be seen as a reduced Q order IIR approximation ($Q << N + K$) to the optimal FIR filter $H_f(z)$. Assuming:

$$H_f(z) = \frac{E(z)}{A(z)} \tag{I.13}$$

where $E(z)$ is the error polynomial and $A(z)$ is the approximating polynomial:

$$E(z) = \sum_{k=0}^{N+Q-1} e_i z^{-i} \tag{I.14}$$

and

$$e_i = h_i^f - \sum_{k=1}^{Q} a_k h_{i-k}^f. \tag{I.15}$$

We recognize this also as a linear predictive problem. Putting (I.15) into matrix form:

$$\hat{\mathbf{e}} = \mathbf{h}_f - \mathbf{H}_f^F\hat{\mathbf{a}}, \tag{I.16}$$

and:

$$\mathbf{h}_f = \begin{bmatrix} h_f(N_1) \\ \vdots \\ h_f(N_2) \end{bmatrix}, \mathbf{H}_f^F = \begin{bmatrix} h_f(N_1 - 1) & \cdots & h_f(N_1 - Q) \\ \vdots & & \vdots \\ h_f(N_2 - 1) & \cdots & h_f(N_2 - Q) \end{bmatrix}$$

we can solve it using common procedures. In particular, rewriting the problem as:

$$\hat{\mathbf{a}} = \arg\min_{\hat{\mathbf{a}}} \|\mathbf{h}_f - \mathbf{H}_f^F \hat{\mathbf{a}}\|_2. \tag{I.17}$$

Choosing $N_1 = 1$ and $N_2 = N+Q$ and assuming $h_f(n) = 0$ for $n < 1$ and $n > N$, we find the well known Yule-Walker equations. This guarantees stability and simplicity of the solution. In more general terms the problem of approximating the impulse response $H_f(z)$ through the linear predictor $A(z)$ falls in the class of the approximation of FIR through IIR digital filters (see, for example, [12, 13]). Using a similar approach we can recalculate the long-term predictor as well.

4 Experimental Analysis

In order to evaluate our method, we have analyzed about one hour of clean speech coming from several different speakers with different characteristics (gender, age, pitch, regional accent) taken from the TIMIT database, re-sampled at 8 kHz. We choose a frame length of $N = 160$ (20 ms) and a order of the optimization problem in (I.1) of $K = 110$. We implement the sparse linear predictive coding using $N_f = 10$ and $N_p = 1$, the residual is encoded using RPE with 20 samples (pulse spacing $S = 8$), a gain and a shift. The gain is coded with 6 bits and the pulse amplitude are coded using a 8 level uniform quantizer, the LSF vector is encoded with 20 bits (providing transparent coding) using the procedure in [14], the pitch period is coded with 7 bits and the gain with 6 bits. This produces a fixed rate of 102 bit/frame (5100 bit/s). No perceptual weighting is employed. The re-estimation is done only on the short-term parameters. The coder that employs re-estimation consists of the following steps:

1. Determine $\tilde{A}(z) = F(z)P(z)$ using sparse linear prediction.
2. Calculate the residual vector $\tilde{\mathbf{r}}$ using RPE encoding.
3. Re-estimate the optimal truncated impulse response $\mathbf{h}_f$.
4. Least square IIR approximation of $\mathbf{h}_f$ using order $N_f = 8, 10, 12$.
5. Optimize the amplitudes of the sparse RPE residual $\tilde{\mathbf{r}}$ using the new synthesis filter $\hat{\mathbf{h}}_f$ (positions and shift stay the same).

We compare two approaches, one with only the re-estimation of $\mathbf{h}_f$ and one with the optimization of the amplitudes of the RPE residual, using (I.3). The results, in comparison with standard Sparse Linear Prediction, are shown in table I.1. An example of the re-estimated impulses responses are shown in Figure I.1.

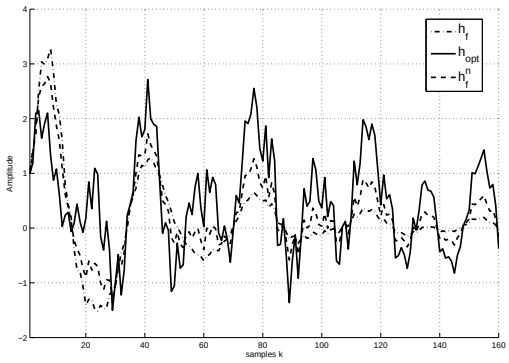

Fig. I.1: *An example of the different impulse response used in the work. The impulse response* $\mathbf{h}_f$ *of the original short-term predictor* $F(z)$*, the optimal re-estimated impulse response adapted to the quantized residual* $\mathbf{h}_{opt}$ *and the approximated impulse response* $\mathbf{h}_f^n$ *of the new short-term predictor* $\hat{F}(z)$*. The order is* $N_f = 10$*.*

5 Conclusions

In this paper, we have proposed a new method for the re-estimation of the prediction parameters in speech coding. In particular, the autoregressive modeling is no more employed as a method to remove the redundancies of the speech segment but as IIR approximation of the optimal FIR filter, adapted to the quantized approximated residual, that is used in the synthesis of the speech segment. The method has shown an improvement in the general performances of the sparse linear prediction framework, but it can be applied also to common methods based on minimum variance linear prediction (e.g. ACELP). The work can be extended for these methods where we expect an even greater increase in performances due to the mismatch between true residual and approximated one.

References

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

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

Table I.1: Improvements over conventional SPARSE LP in the decoded speech signal in terms of reduction of log magnitude segmental distortion (ΔDIST) and Mean Opinion Score (ΔMOS) using PESQ evaluation. A 95% confidence intervals is given for each value.

METHOD	ΔDIST	ΔMOS
N_f=8	+0.12±0.02 dB	+0.01±0.00
N_f=10	+0.35±0.03 dB	+0.05±0.00
N_f=12	+0.65±0.02 dB	+0.04±0.00
N_f=8 + REST	+0.17±0.01 dB	+0.03±0.00
N_f=10 + REST	+0.41±0.02 dB	+0.06±0.00
N_f=12 + REST	+0.71±0.04 dB	+0.07±0.00

[3] D. Giacobello, M. G. Christensen, J. Dahl, S. H. Jensen, and M. Moonen, "Sparse linear predictors for speech processing", *Proc. INTERSPEECH*, 2008.

[4] W. C. Chu, *Speech coding algorithms: foundation and evolution of standardized coders*, Wiley, 2003.

[5] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen "Speech Coding Based On Sparse Linear Prediction", in *Proc. European Signal Processing Conference*, 2009.

[6] P. Stoica and R. Moses, *Spectral analysis of signals*, Pearson Prentice Hall, 2005.

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

[8] 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.

[9] 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", in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 4109–4112, 2009.

[10] P. Kroon, E. F. Deprettere, and R. J. Sluyter, "Regular-pulse excitation - a novel approach to effective and efficient multipulse coding of speech", *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 34, no. 5, pp. 1054–1063, 1986.

[11] M. N. Murthi and B. D. Rao, "All-pole modeling of speech based on the minimum variance distortionless response spectrum," *IEEE Trans. Speech and Audio Processing*, vol. 8, pp. 221–239, 2000.

[12] H. Brandenstein, R. Unbehauen, "Least-squares approximation of FIR by IIR digital filters", *IEEE Trans. on Signal Processing*, vol. 46, pp. 21-30, 1998.

[13] B. Beliczynski, J. Kale, and G. D. Cain, "Approximation of FIR by IIR digital filters: An algorithm based on balanced model reduction", *IEEE Trans. Signal Processing*, vol. 40, pp. 532–542, 1999.

[14] A. D. Subramaniam, B. D. Rao, "PDF optimized parametric vector quantization of speech line spectral frequencies", *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 2, pp. 130–142, 2003.

Paper J

Estimation of Frame Independent and Enhancement Components for Speech Communication over Packet Networks

D. Giacobello, M. N. Murthi, M. G. Christensen,
S. H. Jensen, and M. Moonen

This paper has been published in
Proceedings of the 35th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),
pp. 4682–4685, 2010

The layout has been revised.

Abstract

In this paper, we describe a new approach to cope with packet loss in speech coders. The idea is to split the information present in each speech packet into two components, one to independently decode the given speech frame and one to enhance it by exploiting inter-frame dependencies. The scheme is based on sparse linear prediction and a redefinition of the analysis-by-synthesis process. We present Mean Opinion Scores for the presented coder with different degrees of packet loss and show that it performs similarly to frame dependent coders for low packet loss probability and similarly to frame independent coders for high packet loss probability. We also present ideas on how to make the coder work synergistically with the channel loss estimate.

1 Introduction

With the increasing importance of VoIP (Voice over IP) telephony, alternative methods to improve the robustness of speech codecs to packet loss are required. The approaches presented in literature, notably [1] with the definition of the iLBC (Internet Low Bit Rate Codec), tend to create speech coders that are totally frame independent or, in other words, where each frame is independently decodable and does not depend on the previous frames. On the other hand, in the case of telephony with dedicated circuits, the coding schemes used achieve high quality with low bit rate mostly because of their property to exploit inter-frame dependencies. However, these coding schemes, and in particular the ACELP (Algebraic Code Excited Linear Prediction) based codecs, in the case of packet loss show severe shortcomings [1].

In this paper we introduce a new approach to speech coding over packet networks, creating a coder that has frames with a core that is independently decodable and an enhancement layer that is based on the previously received frames. In particular, we create a coder that can select between two decoding procedures, if the previous frames are received correctly, then it decodes using all the information, otherwise, it uses only the frame independent information. By doing so, we offer the flexibility of a frame independent codec if the loss probability is significant but, if the probability is low (or ideally null), then it will exploit inter-frame dependencies to perform similarly to a frame dependent coder. In our coding scheme, the speech analysis is based on sparse linear prediction which has shown better statistical modeling in creating an output (residual and predictor) that offers better coding properties [2]. Frame independence is achieved through a rethinking of the analysis-by-synthesis (AbS) scheme [3], allowing the possibility of re-estimating the synthesis matrix (and thus the impulse response that generates it) that creates an independently decodable frame of speech given

the residual similarly to what is done in [4].

The paper is organized as follows. Section 2 describes the system architecture of our coder. In Section 3, we provide some experimental results in comparison with G.729a [5] and iLBC, chosen due to their public availability. In Section 4, we discuss how the bit allocation can work synergistically with the channel loss statistics to generally improve the performance of the coder. Section 5 concludes our paper.

2 System Architecture

2.1 Step 1: Prediction Parameters Estimation

The first step is to perform a linear predictive analysis using a sparse linear prediction framework. A sparse linear predictive framework has already shown to offer, not only sparsity properties that make coding more straightforward [2] but also a more compact description of all the features extracted from a speech frame [7]. For a given speech frame $\mathbf{x}$, we obtain an estimate of the underlying autoregressive process by minimizing the prediction error vector $\mathbf{e} = \mathbf{x} - \mathbf{X}\mathbf{a}$ (commonly referred to as the residual):

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_1 + \gamma\|\mathbf{a}\|_1, \tag{J.1}$$

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 $\|\cdot\|_1$ is the 1-norm defined as the sum of absolute values of the vector on which operates. The start and end points N_1 and N_2 can be chosen in various ways assuming that $x(n) = 0$ for $n < 1$ and $n > N$ [8]. The more tractable 1-norm $\|\cdot\|_1$ is used here as a linear programming relaxation of the sparsity measure, often represented as the cardinality of a vector, i.e. the so-called 0-norm $\|\cdot\|_0$. This optimization problem can be posed as a linear programming problem and can be solved using an interior-point algorithm [9]. The choice of the regularization term γ is based on a trade-off between the sparsity of the residual and the sparsity of the predictor, found through by the L-curve [10]. The sparse structure of the predictor, allows a joint estimation of a short-term and a long-term predictors [7]:

$$A(z) \approx \hat{F}(z)\hat{P}(z) \tag{J.2}$$

where $\hat{F}(z)$ is the short-term predictor, commonly employed to remove short-term redundancies due to the formants, and $\hat{P}(z)$ is the long-term pitch predictor that removes the long-term redundancies. The two filters will then be quantized.

2.2 Step 2: Residual Estimation

In order to achieve frame independence, we rethink the analysis-by-synthesis (AbS) scheme used for the estimation of the approximated residual given $A(z)$, estimated in the previous step. In particular, the main equation of AbS coding is the following [3]:

$$\begin{aligned} \hat{\mathbf{r}} = \arg\min_{\mathbf{r}} &\|\mathbf{W}(\mathbf{x} - \hat{\mathbf{H}}\left[\hat{\mathbf{r}}_{-}^{T}, \mathbf{r}^{T}\right]^{T})\|_2, \\ &\text{s.t.} \quad \mathit{struct}(\mathbf{r}), \end{aligned} \tag{J.3}$$

where $\mathbf{x}$ is the $N \times 1$ frame of speech, $\mathbf{W}$ is the $N \times N$ perceptual weighting matrix, $\hat{\mathbf{H}}$ is the $N \times K + N$ synthesis matrix whose $i-$th row contains the elements with index $[0, K + i - 1]$ of the truncated impulse response $\hat{\mathbf{h}}$ of the combined quantized prediction filter $\hat{A}(z) = \hat{F}(z)\hat{P}(z)$:

$$\hat{\mathbf{H}} = \left[\begin{array}{ccc|ccccc} \hat{h}_K & \cdots & & \hat{h}_0 & 0 & 0 & \cdots & 0 \\ \hat{h}_{K+1} & \ddots & & \ddots & \ddots & 0 & 0 & 0 \\ \vdots & \ddots & & \ddots & \cdots & \hat{h}_0 & 0 & 0 \\ \hat{h}_{K+N-2} & \ddots & & \ddots & \cdots & \hat{h}_1 & \hat{h}_0 & 0 \\ \hat{h}_{K+N-1} & \hat{h}_{K+N-2} & & \cdots & \cdots & \hat{h}_2 & \hat{h}_1 & \hat{h}_0 \end{array}\right]. \tag{J.4}$$

The residual term $\left[\hat{\mathbf{r}}_{-}^{T}, \mathbf{r}^{T}\right]^{T}$ is composed of the K previous residual samples $\hat{\mathbf{r}}_{-}$ (the filter memory, already quantized) and the current $N \times 1$ residual vector $\mathbf{r}$ that has to be estimated. It is now clear that the dependence plays a central role in the estimation of the residual. The operator $\mathit{struct}(\cdot)$, that we will leave undefined at the moment, imposes the structure on the residual (e.g., MPE, RPE, CELP). Also, for the sake of simplicity, we will assume that no perceptual weighting is performed ($\mathbf{W} = \mathbf{I}$). The results can then be generalized for an arbitrary $\mathbf{W}$.

We now look for two estimates of the residual in (J.3), one where we take into consideration the previous residual $\hat{\mathbf{r}}_{-}$, one where we do not take it into consideration, therefore setting it to zero. The frame independent is then obtained considering only the $N \times N$ right side of the synthesis matrix in (J.4). The two residuals $\hat{\mathbf{r}}^{FI}$ and $\hat{\mathbf{r}}^{FD}$ will then be quantized.

2.3 Step 3: Re-estimation of the Prediction Coefficients

Once we have the two estimated residuals $\hat{\mathbf{r}}^{FI}$ and $\hat{\mathbf{r}}^{FD}$, we can calculate the truncated impulse response that generates them. In particular, we can rewrite the problem in (J.3) as:

$$\tilde{\mathbf{H}} = \arg\min_{\mathbf{H}} \|(\mathbf{x} - \mathbf{H}\hat{\mathbf{r}})\|_2 \rightarrow \tilde{\mathbf{h}} = \arg\min_{\mathbf{h}} \|(\mathbf{x} - \hat{\mathbf{R}}\mathbf{h})\|_2, \tag{J.5}$$

where

$$\hat{\mathbf{R}} = \begin{bmatrix} \hat{r}_0 & \cdots & \hat{r}_{-K} & 0 & 0 & \cdots & 0 \\ \hat{r}_1 & \ddots & \ddots & \ddots & 0 & 0 & 0 \\ \vdots & \ddots & \ddots & \cdots & \ddots & 0 & 0 \\ \hat{r}_{N-1} & \ddots & \ddots & \cdots & \ddots & \hat{r}_{-K} & 0 \\ \hat{r}_N & \hat{r}_{N-1} & \cdots & \cdots & \cdots & \hat{r}_{-K+1} & \hat{r}_{-K} \end{bmatrix}, \tag{J.6}$$

is the $N \times N + K$ matrix constructed with the frame dependent residual vector $\left[\hat{\mathbf{r}}_-^T, \mathbf{r}^T\right]$. The problem (J.5) allows for a closed form solution when the 2-norm is employed in the minimization:

$$\tilde{\mathbf{h}} = \hat{\mathbf{R}}^T(\hat{\mathbf{R}}\hat{\mathbf{R}}^T)^{-1}\mathbf{x}, \tag{J.7}$$

with

$$\|\mathbf{x} - \hat{\mathbf{R}}\tilde{\mathbf{h}}\|_2 = 0. \tag{J.8}$$

We can now see that the optimal sparse linear predictor (frame dependent and frame independent) is the one that has $\tilde{\mathbf{h}}$ as truncated impulse response. The problem now is that the impulse response will include both short-term and long-term contribution. We can split the two contribution as:

$$\hat{A}(z) = \hat{F}(z)\hat{P}(z) \rightarrow \hat{\mathbf{H}} = \hat{\mathbf{H}}_f\hat{\mathbf{H}}_p, \tag{J.9}$$

and re-estimate only the short-term impulse response, assuming that the long-term impulse response will not vary significantly, we can rewrite (J.5) using (J.9):

$$\tilde{\mathbf{h}}_f = \arg\min_{\mathbf{h}_f} \|(\mathbf{x} - \hat{\mathbf{H}}_p\hat{\mathbf{R}}\mathbf{h}_f)\|_2. \tag{J.10}$$

We can then obtain two estimates of the impulse responses, a frame dependent one $\tilde{\mathbf{h}}_f^{FD}$ and a frame independent one $\tilde{\mathbf{h}}_f^{FI}$. In the frame independent case, the matrix $\hat{\mathbf{R}}$ in (J.6) will be $N \times N$ and it will be constructed using only $\hat{\mathbf{r}}^{FI}$.

Using an autoregressive modeling of both $\tilde{\mathbf{h}}^{FD}$ and $\tilde{\mathbf{h}}^{FI}$, we obtain two new short-term predictive filters $\tilde{F}^{FI}(z)$ and $\tilde{F}^{FD}(z)$, that not only generate a better approximate of the impulse response but are also stable [4]. We will then quantize them.

2.4 Definition of an Enhancement Layer

For a given frame of speech we have calculated two residuals ($\hat{\mathbf{r}}^{FI}$ and $\hat{\mathbf{r}}^{FD}$) and two predictors ($\tilde{A}^{FI}(z) = \hat{P}(z)\tilde{F}^{FI}(z)$ and $\tilde{A}^{FD}(z) = \hat{P}(z)\tilde{F}^{FD}(z)$). The reconstructed speech frames are, for the frame independent case:

$$\hat{\mathbf{x}}^{FI} = \hat{\mathbf{H}}_p\tilde{\mathbf{H}}_f^{FI}\hat{\mathbf{r}}^{FI}, \tag{J.11}$$

and, for the frame dependent case:

$$\hat{\mathbf{x}}^{FD} = \hat{\mathbf{H}}_p \tilde{\mathbf{H}}_f^{FD} \left[(\hat{\mathbf{r}}_-^{FD})^T, (\hat{\mathbf{r}}^{FD})^T \right]^T . \tag{J.12}$$

It should be noted that $\hat{\mathbf{H}}_p$ is constructed from the truncated impulse response of $\hat{P}(z)$, that is equal for both cases, but in the frame independent case $\hat{\mathbf{H}}_p$ is $N \times N$ while in the frame dependent case $\hat{\mathbf{H}}_p$ is $N \times N + K$.

What we will do is transmit the frame independent parameters ($\hat{\mathbf{r}}^{FI}$, $\tilde{A}^{FI}(z) = \hat{P}(z)\tilde{F}^{FI}(z)$) to robustly construct a frame independent coder then define an enhancement layer based on the frame dependent parameters. To do so, we transmit the differences between the two short-term predictors $\tilde{F}^{\Delta}(z)$ and the differences between the two residuals $\hat{\mathbf{r}}^{\Delta}(z)$. We will specify in the next section how to code the differences and in which domain.

If there is no loss of speech packets, it is clear that the decoder will work in "full" mode, using the frame independent informations together with the enhancement layer, (J.12) would then become:

$$\hat{\mathbf{x}} = \hat{\mathbf{H}}_p (\tilde{\mathbf{H}}_f^{FI} + \tilde{\mathbf{H}}_f^{EN}) \left[(\hat{\mathbf{r}}_-^{FI} + \hat{\mathbf{r}}_-^{EN})^T, (\hat{\mathbf{r}}^{FI} + \hat{\mathbf{r}}^{EN})^T \right]^T , \tag{J.13}$$

where $\tilde{\mathbf{H}}^{EN}$, $\hat{\mathbf{r}}_-^{EN}$ and $\hat{\mathbf{r}}^{EN}$ are functions of the parameters used to define the enhancement layer $\tilde{F}^{\Delta}(z)$ and $\hat{\mathbf{r}}^{\Delta}(z)$.

The interesting case is when a $k-$th frame is missing. In this case, the $k+1-$th frame is self-constructed only from the frame independent parameters, using (J.11). The $k+2-$th frame will then be reconstructed using the frame dependent information but first it is necessary to convert the part of the residual of the $k+1-$th frame $\hat{\mathbf{r}}_-^{FI}$, that will appear in the reconstruction equation (J.13), into the frame dependent one ($\hat{\mathbf{r}}_-^{FI} + \hat{\mathbf{r}}_-^{FE}$).

3 Experimental Analysis

3.1 Setup

Linear predictive analysis

The length of the analyzed speech frames in our scheme is $N = 160$ (20 ms). The order of the optimization problem in (J.1) is $K = 110$, meaning that we can cover accurately pitch delays in the interval $[N_{stp} + 1, K - N_{stp} - 1]$, including the usual range for the pitch frequency [70Hz, 500Hz]. This also means that the dependency from the previous frame is $K = 110$ residual samples. The linear prediction filters $F(z)$ and $P(z)$ are chosen as respectively of order $N_f = 12$ and $N_p = 1$. $F(z)$ is coded initially as an LSF vector with 26 bits (providing transparent coding) using the procedure in [11]. The pitch period is coded with 7 bits and the gain with 6 bits. **Coding of the residual**

The residual coding of both $\hat{\mathbf{r}}^{FI}$ and $\hat{\mathbf{r}}^{FD}$ is implemented using an RPE procedure [12] with fixed shift equal to zero and a sample spacing $Q = 8$. The RPE procedure is slightly modified to have the first 8 pulses as nonzero (27 nonzero pulses in total). This guarantees, other than a full row rank of $\hat{\mathbf{R}}$, also a well conditioned problem in (J.10) in both the frame dependent, where $\hat{\mathbf{R}}$ is $N \times N + K$ and frame independent case, where $\hat{\mathbf{R}}$ is $N \times N$. $\hat{\mathbf{r}}^{FI}$ is calculated first, then we impose the same sign structure when calculating $\hat{\mathbf{r}}^{FD}$. The residuals are also quantized simultaneously with a 8-level uniform quantizer, the peak magnitude is encoded with 6 bits per frame and 1 bit per pulse is used to code the sign.

Re-estimation procedure

In the re-estimation procedure (J.10), we impose the constraint of having $h_f(0) = 1$, this is done to simplify the IIR modeling of $\mathbf{h}_f$, so that the filter has a unit numerator. The new short-term predictive filters are also coded as an LSF vector with 26 bits (providing transparent coding in both cases).

Coding of the Enahncement Layer

The difference vector $\tilde{F}^{\Delta}(z)$ is calculated between $\tilde{F}^{FD}(z)$ and $\tilde{F}^{FI}(z)$ in the quantized LSF domain. A 11 bits vector quantizer has proved to be sufficient to describe the difference between the two polynomial. In particular, the reconstructed polynomial (sum of $\tilde{F}^{FI}(z)$ and $\tilde{F}^{\Delta}(z)$ in the LSF domain) is going to fulfill the spectral transparency performances as $\tilde{F}^{FD}(z)$ does. As for the difference between the two residuals $\hat{\mathbf{r}}^{\Delta}(z)$, we will use 2 bits per pulse, sufficient to code the difference almost without distortion in the quantized domain. Each frame will then be coded with a total of 218 bits, 153 belonging to the frame independent part and 65 belonging to the frame dependent enhancement layer, generating a total bit rate of 10.9 kbps (7.65 kbps for the frame independent information and 3.25 kbps for the enhancement layer).

3.2 Results

In this subsection we present the numerical results of our method compared, in terms of PESQ-MOS [13], to the iLBC in [1] and the G.729a [5], working respectively at 13.33 kbps and 8 kbps.

We have analyzed about one hour of clean speech coming from several different speakers with different characteristics (gender, age, pitch, regional accent) taken from the TIMIT database [14], re-sampled at 8 kHz. In our simulations, we used the Gilbert model for packet loss with parameters $q = P(\text{loss}|\text{loss}) = 0.7$ and $p = P(\text{loss}|\text{noloss})$ varied in order to have an average loss rate of $p/(p+q)$. The analyzed loss rates are 0%, 2.5%, 5%, 7.5%, 10%, and 15%. In our implementation, a simple packet loss concealment (PLC) based on repeating the previously received frames is implemented for our method and also for the G.729a.

As the results suggest in Figure J.1, the coder works well with performances

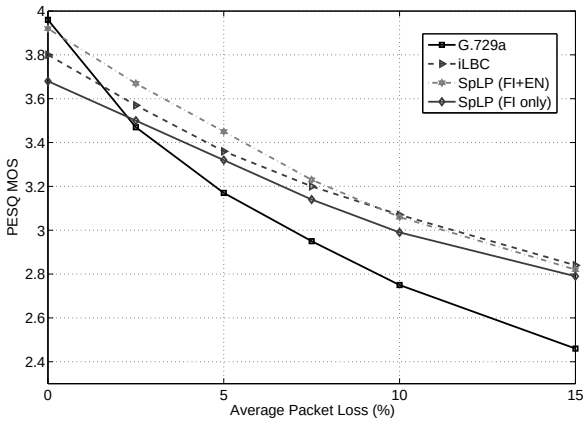

Fig. J.1: *Performances of the compared methods: G.729a (8 kbps), iLBC (13.33 kbps), and our introduced method based on sparse linear prediction (SpLP) with (FI+EN) and without (FI) the frame dependent enhancement layer (respectively 10.9 and 7.65 kbps).*

similar to the G.729a codec at 0% packet losses, where the iLBC fails to do so. The frame dependent layer seems to work well at low packet loss rates and loses its enhancement properties when the loss rate increases, as we may have expected. It should be noted, that our scheme, when only the frame independent part is employed, performs only slightly worse than iLBC with a net decrease in rate and a very simple PLC scheme. This can be explained by the novelty we have introuduced in the re-estimation of the "frame independent linear predictors" and by the compact and robust modeling advantages offered by sparse linear prediction [2]. Our coder performs worse than iLBC for loss percentage higher than 7.5%, mostly due to the more advanced PLC implemented on iLBC. A final comment is that the structured sparsity of the residual can allow guidance in order to generate an excitation sequence when packet loss occurs, for example when the other parameters are estimated in a Hidden Markov Model based PLC [15].

4 Discussion

The coding algorithm we have presented is representative of a more general problem, where we minimize the expected distortion between the analyzed speech and its coded approximation, subject to a rate constraint:

$$\begin{aligned} \text{minimize} \quad & D(\mathbf{x}, \hat{\mathbf{x}}), \\ \text{subject to:} \quad & R(\hat{\mathbf{x}}) \leq R^*; \end{aligned} \tag{J.14}$$

where $D(\mathbf{x}, \hat{\mathbf{x}})$ represent the expected distortion by representing $\mathbf{x}$ with $\hat{\mathbf{x}}$, $R(\hat{\mathbf{x}})$ is the rate (or, equivalently, the bit allocation) to transmit $\hat{\mathbf{x}}$ and R^* is the maximum possible rate (the constraint). In our case, the distortion will be dependent on how the representation of $\hat{\mathbf{x}}$ divided between a frame independent core $\hat{\mathbf{x}}^{FI}$ and a frame dependent enhancement layer $\hat{\mathbf{x}}^{EN}$. In particular, the distortion term can be made dependent on the loss rate and therefore adjusting the bit allocation on the frame dependent and frame independent parts. We see for example from Figure J.1 how the increase in performance given by the enhancement layer tend to reduce itself with the increase of the loss rate, in particular with a 15% of lost packets, there is almost no difference, although there is a 3.25 kbps difference in rate. In this case, what we would then like to do is to reallocate the bits used to define the enhancement layer, to improve the performances of the frame independent coder, the problem in (J.14) can then be rewritten as:

$$\begin{array}{rl} \text{min.} & w_{p_L} D(\mathbf{x}, \hat{\mathbf{x}}^{FI}) + (1 - w_{p_L}) D(\mathbf{x}, \hat{\mathbf{x}}^{FI} + \hat{\mathbf{x}}^{EN}), \\ \text{s.t.:} & R(\hat{\mathbf{x}}^{FI}) + R(\hat{\mathbf{x}}^{EN}) \leq R^*. \end{array} \tag{J.15}$$

where the allocation of the rate is now split between the frame independent part and the enhancement layer that exploits frame dependence. Also the expected distortion will be proportional to the different bit allocation. In (J.15), w_{p_L} is a weight that will be somehow proportional to the packet loss probability p_L ($0 \leq w_{p_L} < 1$), and, on a higher order analysis, it will also depend on other loss statistics such as the burst length. An interesting case, it is also to use the bit allocated for the enhancement layer to bring information for the packet loss concealment on how to reconstruct the missing frames when the loss rate is high. How to implement the problem in (J.15) will be subject of our future work.

5 Conclusion

In this paper, we have introduced a novel formulation for speech coding in packet networks. In particular, we have defined an algorithm that generates parameters that independently decode a speech segment at 7.65 kbps. A 3.25 kbps frame dependent enhancement layer is added to exploit inter-frame dependencies. This allows to reach performances similar to the G.729a coder for 0% packet loss probability while behaving similarly to the iLBC coder for higher packet loss probabilities. Sparse linear prediction has been used to robustly analyze a speech segment, providing a joint estimation of long-term and short-term predictors and a sparse residual. Also, a new formulation of the Analysis-by-Synthesis scheme has been defined by re-estimating a more appropriate synthesis matrix. A definition of the future work on the how to optimally construct a frame dependent/independent coder has also been given.

References

[1] S. V. Andersen, et al., "iLBC - A linear predictive coder with robustness to packet loss", in *Proc. IEEE Workshop on Speech Coding*, pp. 23–25, 2002.

[2] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen "Speech coding based on sparse linear prediction", in *Proc. European Signal Processing Conference*, 2009.

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

[4] D. Giacobello, M. N. Murthi, M. G. Christensen, S. H. Jensen, and M. Moonen, "Re-estimation of Linear Predictive Parameters in Sparse Linear Prediction", to appear in *Rec. 43rd Asilomar Conf. on Signals, Systems, and Computers*, 2009.

[5] ITU-T G.729, "Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP)", 2009.

[6] D. Giacobello, M. G. Christensen, J. Dahl, S. H. Jensen, and M. Moonen, "Sparse linear predictors for speech processing", in *Proc. Interspeech*, pp. 1353–1356, 2008.

[7] 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", in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 4109–4112, 2009.

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

[9] S. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge University Press, 2004.

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

[11] A. D. Subramaniam and B. D. Rao, "PDF optimized parametric vector quantization of speech line spectral frequencies", *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 2, 2003.

[12] P. Kroon, E. F. Deprettere, and R. J. Sluyter, "Regular-pulse excitation - a novel approach to effective and efficient multipulse coding of speech",

IEEE Trans. on Acoustics, Speech, and Signal Processing, vol. 34, no. 5, pp. 1054–1063, 1986.

[13] ITU-T Recommendation P.862, "Perceptual Evaluation of Speech Quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs", 2001.

[14] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dahlgren, "DARPA-TIMIT acoustic-phonetic continuous speech corpus", *Technical Report NISTIR*, no. 4930, 1993.

[15] C. A. Rodbro, M. N. Murthi, S. V. Andersen, S. H. Jensen, "Hidden Markov Model-Based Packet Loss Concealment for Voice Over IP", *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1609–1623, 2006.