SPEECH CODING BASED ON SPARSE LINEAR PREDICTION

Daniele Giacobello[1], *Mads Græsbøll Christensen*[1], *Manohar N. Murthi*[2], *Søren Holdt Jensen*[1], *Marc Moonen*[3]

[1]Dept. of Electronic Systems, Aalborg Universitet, Denmark
[2]Dept. of Electrical and Computer Engineering, University of Miami, USA
[3]Dept. of Electrical Engineering (ESAT-SCD), Katholieke Universiteit Leuven, Belgium
{dg,mgc,shj}@es.aau.dk, mmurthi@miami.edu, marc.moonen@esat.kuleuven.be

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 formulations 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:

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

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:

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

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

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

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

The work of Manohar N. Murthi is supported by the National Science Foundation via awards CCF-0347229 and CNS-0519933.

Daniele Giacobello was with the Dept. of Electrical and Computer Engineering, University of Miami, USA, as a visiting researcher during this work. The authors would like to thank Shaminda Subasingha (University of Miami) and Anders Ekman (KTH) for providing part of the code used in the evaluation procedures as well as useful tips.

© EURASIP, 2009

(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:

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

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 (2) is intimately related to the *a priori* knowledge that we have on the coefficients 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 coefficients vector and the error vector are both i.i.d. Laplacian sets of variables:

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

3. BASIC CODING STRUCTURE

The core of the speech coder is based on the optimization problem (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 (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 (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 (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 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 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 (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 $P(z)$.

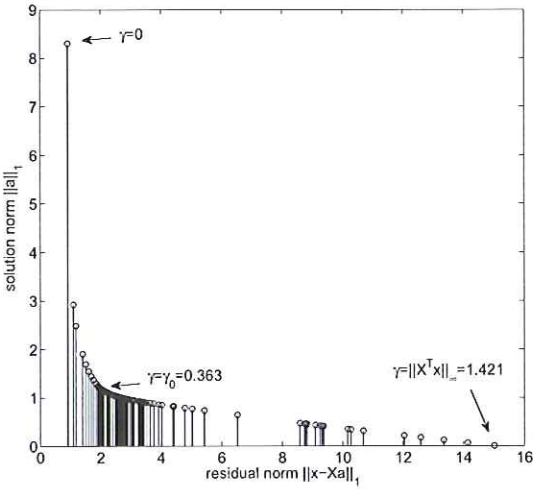


Figure 1: *An example of the L-curve* $(\|\mathbf{x}-\mathbf{X}\mathbf{a}_\gamma\|_1, \|\mathbf{a}_\gamma\|_1)$ *obtained for a segment of 160 samples of speech (20 ms at 8 kHz); the order is* $K=110$. *The lower and upper bounds of* γ *and their respective solution norm and residual norm are also shown.*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

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:

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

where the prediction vector $\mathbf{a}$ is obtained by solving the minimization problem in (3) for different orders k, using the regularization parameter γ_0 found in the previous step. It has been shown [6] that when solving (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 (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 (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

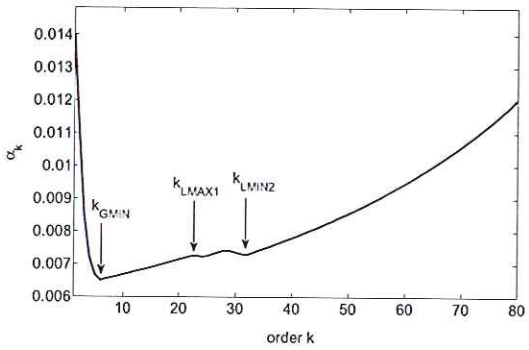


Figure 2: *An example of the cost function α_k for a segment of voiced speech. The values used for the order selection $k_{GMIN} = 6$, $k_{LMAX1} = 23$ and $k_{LMIN2} = 32$ are shown.*

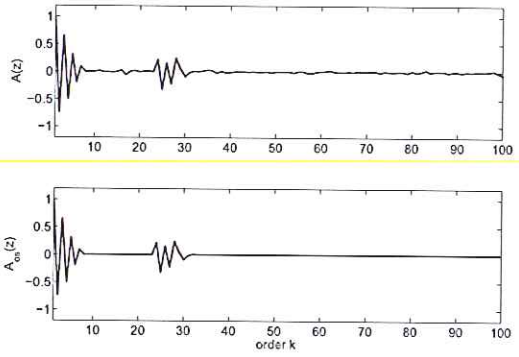


Figure 3: *An example of the high order predictor coming out of the minimization process $A(z)$ and its "clean" version $A_{os}(z)$*

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 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 3. 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 $A_{os}(z)$:

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

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

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

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:

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

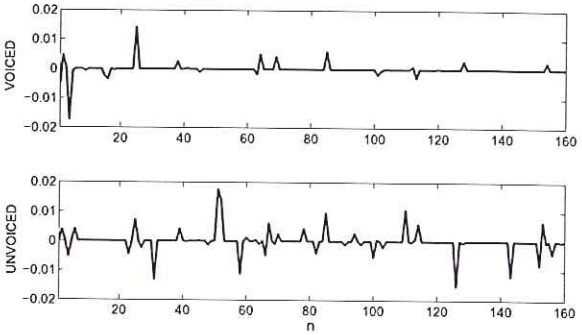


Figure 4: *An example of the sparse residual vector for a segment of voiced (above) and unvoiced speech (below).*

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 $P(z)$.

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 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 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 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 (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

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

Table 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.

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 [70Hz, 500Hz].

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 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 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 reconstructed 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 (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 (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 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 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 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 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.

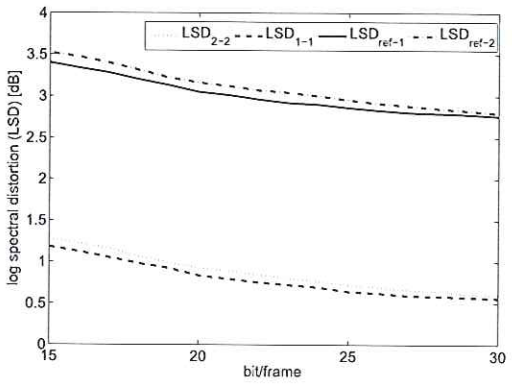


Figure 5: *Values of average log spectral distortion (LSD) for voiced speech at different bits per LSFs frame. The figure shows the LSD values between the two AR models (obtained with our scheme ($N_{stp} = 8$) and with 2-norm minimization ($N_{stp} = 10$)) and their quantized version (LSD_{1-1q} vs. 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.*

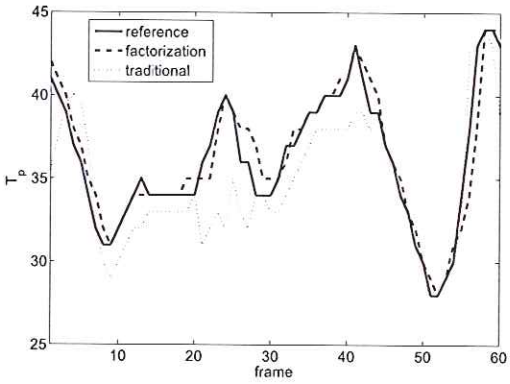


Figure 6: *Integer pitch lag (T_p) tracking performances for our method based on the factorization of short-term and long-term predictor compared with traditional close-loop method based on autocorrelation and a reference value based on subspace pitch estimation [12].*

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.