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

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

Published in:
I E E E International Conference on Acoustics, Speech and Signal Processing. Proceedings

DOI (link to publication from Publisher):
10.1109/ICASSP.2009.4960532

Publication date:
2009

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

JOINT ESTIMATION OF SHORT-TERM AND LONG-TERM PREDICTORS IN SPEECH CODERS

Daniele Giacobello[1,2], *Mads Græsbøll Christensen*[1], *Joachim Dahl*[1],
Søren Holdt Jensen[1], *Marc Moonen*[2]

[1]Dept. of Electronic Systems (ES-MISP), Aalborg University, Aalborg, Denmark
[2]Dept. of Electrical Engineering (ESAT-SCD), Katholieke Universiteit Leuven, Leuven, Belgium
`{dg,mgc,joachim,shj}@es.aau.dk`, `marc.moonen@esat.kuleuven.be`

ABSTRACT

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.

Index Terms— Speech analysis, linear predictive coding.

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 2 and Section 3 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 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 general form of the short-term linear predictor is:

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

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

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

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:

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

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 (1) and (3) corresponds the multiplication in the z-domain of the two transfer functions:

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

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 (2) by adding a regularization term that imposes sparsity on the coefficient vector:

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

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]:

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

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 (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 (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 (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:

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

which can be considered to be true observing the coefficients of the polynomial in (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 (6) as an approximation of (5). The problem of finding γ that offers the best fitting of the model in (6) will be addressed in the next section.

Once the solution of (6) has been found, corresponding to the estimated version of the coefficients of $A(z)$ in (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:

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

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:

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

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

One of the main drawbacks is that even though the polynomial corresponding to the solution of (6) is intrinsically stable, by selecting the first N_{stp} coefficients we can risk having the roots of the

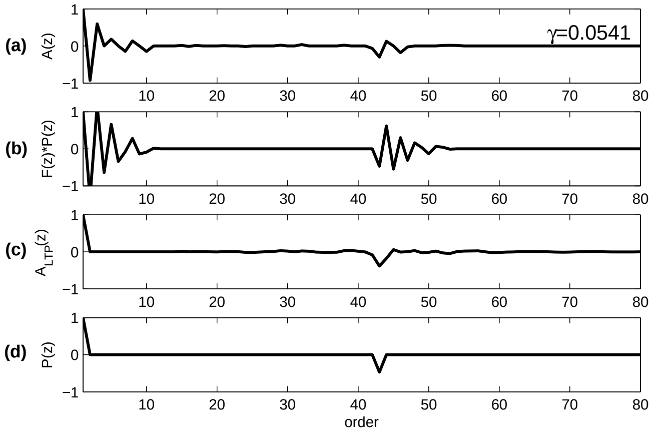

Fig. 1. **(a)** *and* **(b)** *show a comparison between the polynomial obtained with regularized minimization $A(z)$ and multiplication of the two predictors $F(z)P(z)$ obtained in cascade;* **(c)** *and* **(d)** *a comparison of the two long-term predictors $A_{LTP}(z)$ and $P(z)$.*

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 (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 (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 (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:
  $$\gamma(n) = -0.18 g_p^2(n) + 0.2. \tag{10}$$
  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 (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

In order to validate our results, we have analyzed about 100.000 frames of clean speech coming from several different speakers taken from the TIMIT database [11], re-sampled at 8 kHz. The used set of speakers is different from the one used in the analysis and training phase. The three regularized methods with constant, adaptive and optimal γ ($\mathbf{R_c}$, $\mathbf{R_a}$, $\mathbf{R_o}$) are compared with the classical ACELP ($\mathbf{A_c}$) and the ACELP scheme with joint optimization of long-term and short-term predictors ($\mathbf{A_j}$) according to [3].

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 (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 $[N_{stp} + 1, K - N_{stp} - 1]$

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

Table 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.

or equivalently 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:

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

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 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 (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 (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 (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.

7. 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.