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

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

[1]Dept. of Electronic Systems, Aalborg Universitet, Aalborg, Denmark

[2]Dept. of Electrical Engineering (ESAT-SCD), Katholieke Universiteit Leuven, Leuven, Belgium

1 Introduction

- Linear Prediction suffers from well known problems when the 2-norm error minimization criterion is employed in the analysis and coding of voiced speech.
- 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.
- In this work we define a joint estimation approach based on the observation of the behavior of the short-term and long-term cascade polynomial.
- Imposing sparsity on a high order predictor, we obtain a polynomial that can be easily factorized into long-term and short-term predictors.
- This method incorporated into an ACELP scheme shows to have better performance than traditional cascade methods and other joint estimation methods.

2 Joint Estimator

- In order to remove near-sample redundancies and distant-sample redundancies, a cascade of a short-term linear predictor $F(z)$ and a long-term linear predictor $P(z)$ is employed.
- The cascade of the two predictors corresponds the multiplication in the z-domain of the two transfer functions:

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

- $A(z)$ will therefore be highly sparse. Sparsity is then taken into account in new error minimization criterion:

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

where the 1-norm is employed as a relaxation of the non-convex 0-norm and:

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

- $A(z)$ is now similar to the multiplication between a short-term and a long-term predictor:

$$A(z) \approx A_{LTP}(z)A_{stp}(z).$$

- The first N_{stp} coefficients are used as the estimated coefficients of the short-term predictor $A_{stp}(z)$.
- $A_{LTP}(z)$ is created by taking the quotient of the division between $A(z)$ by $A_{stp}(z)$. The minimum value and its position will correspond to our estimate of the pitch gain and delay (parameters of the predictor $P(z)$):

$$g_p = \min\{a_{LTP}\},$$
$$T_p = \arg\min\{a_{LTP}\}.$$

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

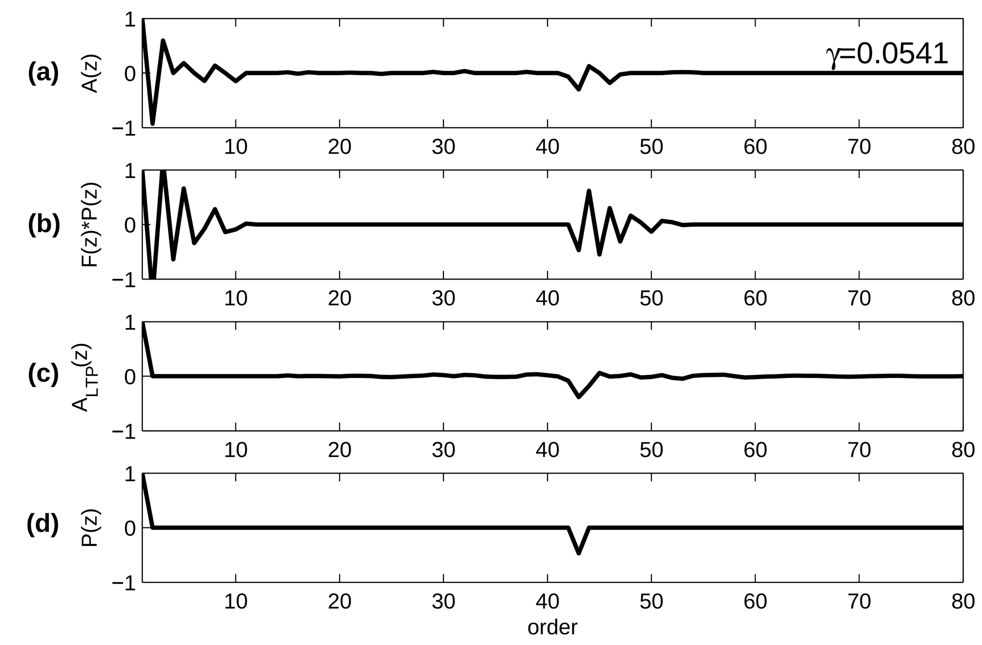

Figure 1: **(a)** and **(b)** show a comparison between the polynomial obtained with regularized minimization $A(z)$ and multiplication of the two predictors $F(z)P(z)$ obtained in cascade; **(c)** and **(d)** a comparison of the two long-term predictors $A_{LTP}(z)$ and $P(z)$.

3 Regularization Parameter

- The regularization parameter γ is intimately related to the *a priori* knowledge that we have on the coefficients vector $\{a_k\}$ (how sparse $\{a_k\}$ is) considering our minimization criterion from a Bayesian point of view.
- The best trade-off between the 2-norm of the residual and the 1-norm of the solution vector is found finding the point of maximum curvature of the curve $(\|\mathbf{x} - \mathbf{X}\mathbf{a}_\gamma\|_2, \|\mathbf{a}_\gamma\|_1)$ (L-curve).
- γ is bounded $(0 < \gamma < \|\mathbf{X}^T\mathbf{x}\|_\infty)$.
- We investigate three approaches for the selection of γ based on the magnitude of the difference between the encoded-decoded signal and the original signal:
  constant $(\mathbf{R_c})$. The value that on average gave the best result.
  adaptive $(\mathbf{R_a})$. The value of γ intimately related to the pitch gain g_p. We update γ using the following approximate relation:
  $$\gamma(n+1) = -0.18g_p^2(n) + 0.2.$$
  optimal $(\mathbf{R_o})$. γ is tuned for every frame analyzed in order to obtain the best result.
- Selection of γ is based on the magnitude of the difference between the encoded-decoded signal and the original signal.

4 Validation

- The joint method is implemented in an ACELP scheme.
- The order of the optimization problem 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 $N_{LTP} = 1$.
- Using $K = 110$ we can cover pitch pitch frequencies in the interval $[82 \text{ Hz}, 571 \text{ Hz}]$.
- Residual vector is encoded using 40 non-zero samples constrained with ± 1 values and a gain (Algebraic Codebook).
- For each method $(\mathbf{R_c}, \mathbf{R_a}, \mathbf{R_o}, \mathbf{A_j})$, the signals coming out of the encoding-decoding scheme are compared to the original speech and the traditional ACELP $\mathbf{A_c}$, PESQ evaluation is then performed.

METHOD	ΔDIST	ΔMOS
$\mathbf{R_o}$	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

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

5 Discussion

- The increase in accuracy is given by the more precise search of the algebraic codeword (spectrally white residual) and improved pitch tracking.
- Number of taps is highly customizable and can be chosen using an Analysis-by-Synthesis scheme or a Model Order Selection criterion.
- Lower emphasis on peaks is achieved by intrinsically taking into consideration the possible outliers due to the pitch excitation in the minimization process. This reflects in a lower sensitivity of the short-term predictor to quantization than traditional LP.
- The cascade $A_{stp}(z)P(z)$ has a very low instability rate (less than 0.01%).
- The optimization problem can be posed as a quadratic programming problem and solved efficiently using an interior-point algorithm.

6 Conclusion

- A new formulation for the minimization process involved in the linear prediction has been presented.
- We have obtained a better statistical fitting for the model of speech that makes analysis and coding more straightforward and accurate.
- Higher accuracy than with traditional LP have been obtained due to whiter residual, improved pitch tracking and predictors that are less sensitive to quantization.

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', *Proc. IEEE*, 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*, vvol. 2, pp. 157–160, 2003.

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

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

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

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