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]Department of Electronic Systems, Aalborg Universitet, Aalborg, Denmark
[2]Department of Electrical and Computer Engineering, University of Miami, USA
[3]Department of Electrical Engineering, Katholieke Universiteit Leuven, Leuven, Belgium

1 Introduction

- A new speech coding concept is created by introducing sparsity constraints in a linear prediction scheme both on the residual and on the high order 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 provided by exploiting the sparse characteristics of the high order predictor.
- We show that better statistical modeling in the context of speech analysis creates an output that offers better coding properties.

2 Sparse Linear Prediction

- The class of problems considered as those covered by the optimization problem associated with finding the prediction coefficient vector $\mathbf{a}$ from a set of observed real samples $x(n)$ for $n = 1, \ldots, N$ so that the 1-norm of the error is minimized:

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

where the 1-norm is employed as a relaxation of the non-convex 0-norm. $\mathbf{x}$ is the observed vector and $\mathbf{X}$ is the matrix containing previous values.

3 Coding Structure

3.1 Selection of the 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 1-norm of the residual and the 1-norm of the solution vector is found as the point of maximum curvature of the curve $(\|\mathbf{x}-\mathbf{X}\mathbf{a}_\gamma\|_1, \|\mathbf{a}_\gamma\|_1)$ (modified L-curve).

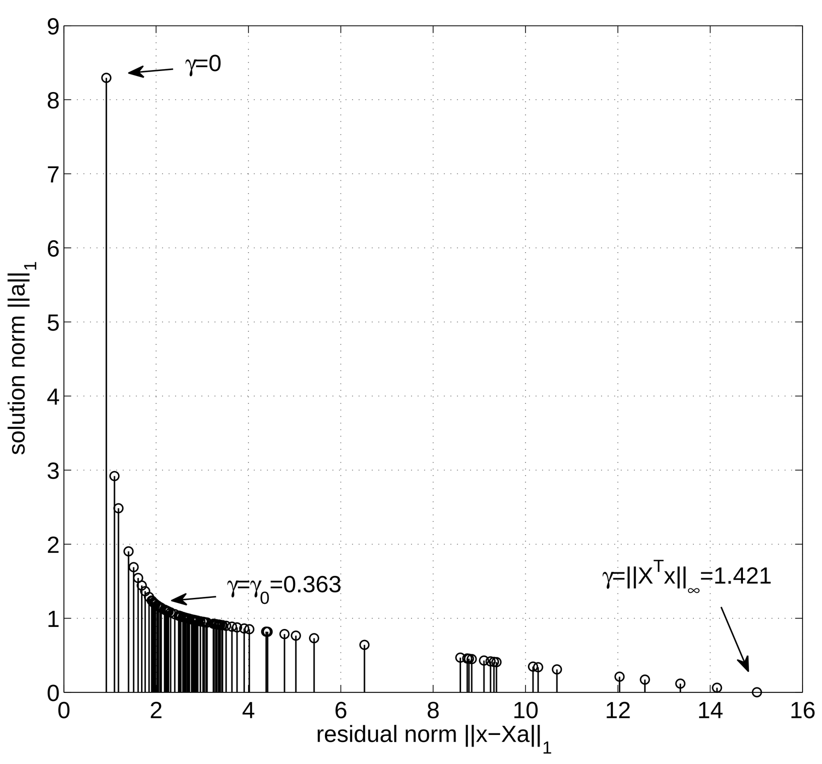

Figure 1: An example of the L-curve $(\|\mathbf{x}-\mathbf{X}\mathbf{a}_\gamma\|_1, \|\mathbf{a}_\gamma\|_1)$ obtained for a segment of 160 samples of speech (20 ms at 8 kHz); the order is $K = 110$. The lower and upper bounds of γ and their respective solution norm and residual norm are also shown.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

3.2 Factorization of the high order predictor

- 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.
- Use of order selection criteria for autoregressive (AR) spectral estimation generalized to the minimization of the sum of absolute values:

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

- 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.

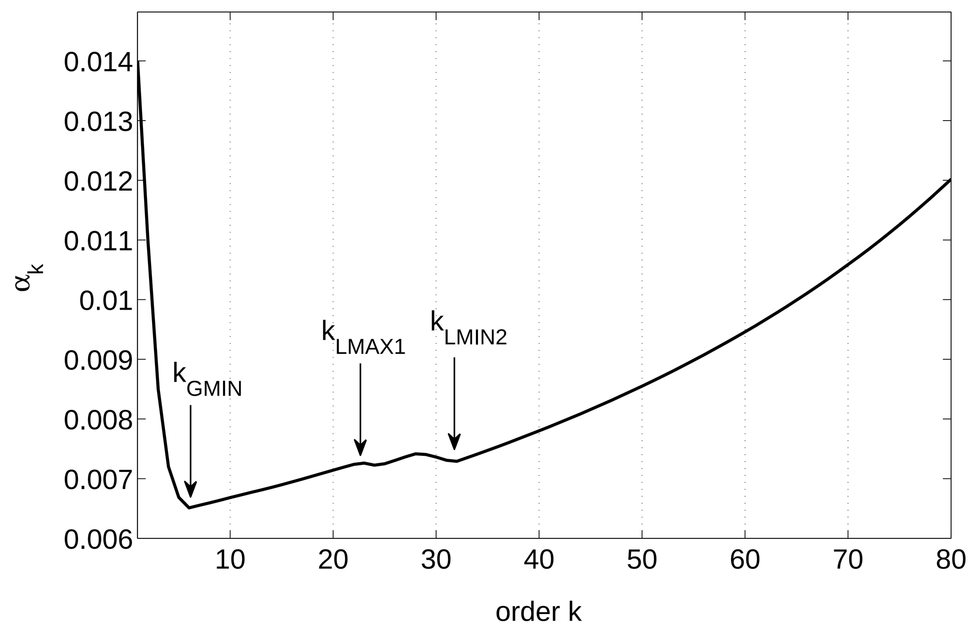

Figure 2: An example of the cost function α_k for a segment of voiced speech. The values used for the order selection $k_{GMIN} = 6$, $k_{LMAX1} = 23$ and $k_{LMIN2} = 32$ are shown.

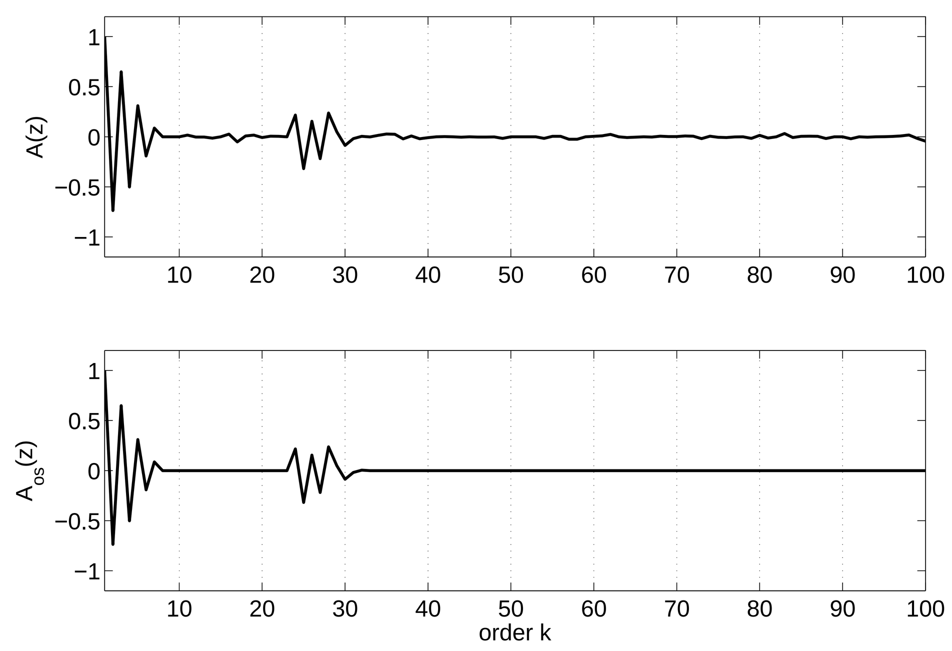

Figure 3: An example of the high order predictor coming out of the minimization process $A(z)$ and its "clean" version $A_{os}(z)$.

3.3 Encoding of the residual

- Use of multipulse encoding (MPE) techniques efficient with the characteristics of the residual.

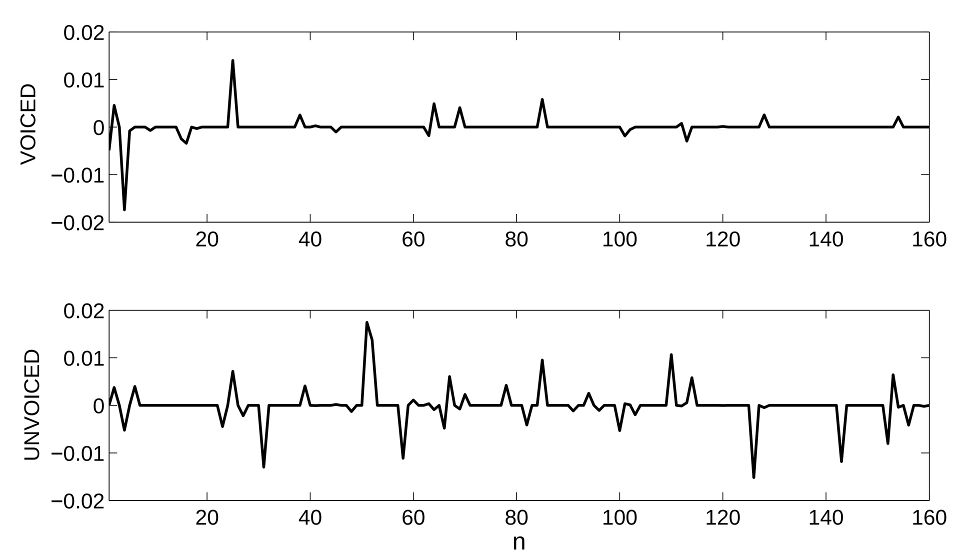

Figure 4: An example of the sparse residual vector for a segment of voiced (above) and unvoiced speech (below).

4 Validation

- Variable rate coding thanks to the model order selection criterion employed.
- Intrinsic classification between voiced an unvoiced speech performed in the factorization procedure of the high-order polynomial.
- Voiced speech: 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 $N_p = 1$ and $N_p = 3$.
- Unvoiced speech: the order is usually between $N_{stp} = 8$ and $N_{stp} = 11$, without long-term information.

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

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. A 95% confidence intervals is given for each value.

5 Discussion

- 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.
- The short-term envelopes are 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.
- 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 for the presented method.

References

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

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

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

[4] 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. ICASSP*, 2009.