Sparsity in Linear Predictive Coding of Speech

Daniele Giacobello

January 31th, 2013

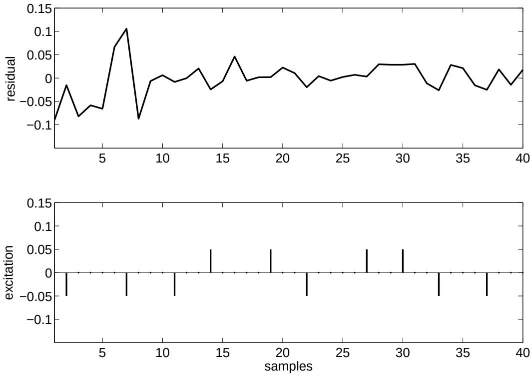
Motivation
In one sentence

- Revisiting early concepts in speech and audio analysis in light of the new development in sparse representation.

Motivation

Why sparsity in Linear Prediction?

- Initial idea: reduce the mismatch between a "white noise"-like prediction residual and a *sparse* approximation.



Motivation

Why sparsity in Linear Predictive Analysis-by-Synthesis coding?

- One of the earliest problems in speech coding!

$$\widehat{\mathbf{r}} = \arg\min_{\mathbf{r}} \|\mathbf{W}(\mathbf{x} - \mathbf{H}\mathbf{r})\|_2 + \gamma\|\mathbf{a}\|_0$$

- Solution is impractical due to the combinatorial nature of the problem.
- Suboptimal algorithm was proposed to find one pulse at the time: Multi-Pulse Encoding (MPE) (= Matching Pursuit).

Outline

Introduction

Sparse Linear Prediction

Sparsity in LPAS Coders

Conclusions

Outline

Sparse Linear Prediction
Fundamentals 1/2

- A speech sample $x(n)$ is approximated as a linear combination of past samples:

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

where $\{a_k\}$ are the prediction coefficients, $e(n)$ is prediction error. In matrix form becomes:

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

- We can consider a generalized optimization framework to find $\mathbf{a}$:

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

- How to choose p, k and γ depends on the kind of applications we want to implement.

Sparse Linear Prediction
Fundamentals 2/2

- If we want to introduce sparsity in the LP optimization framework, we can set $p = 0$ and $k = 0$:

$$\min_{\mathbf{a}} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_0 + \gamma \|\mathbf{a}\|_0,$$

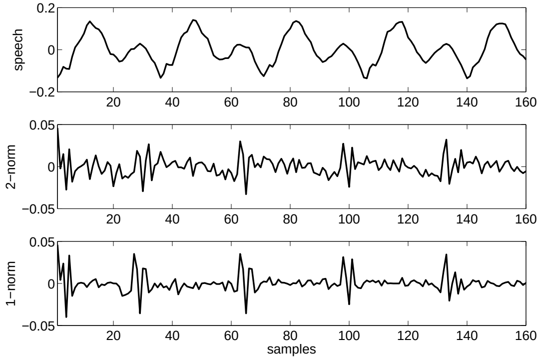
- γ relates to *how sparse* $\mathbf{a}$ is (prior knowledge of $\mathbf{a}$).
- 1-norm used as a convex relaxation of the 0-norm:

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

- Consider now the case of a short-term predictor that engenders a sparse residual ($\gamma = 0$):

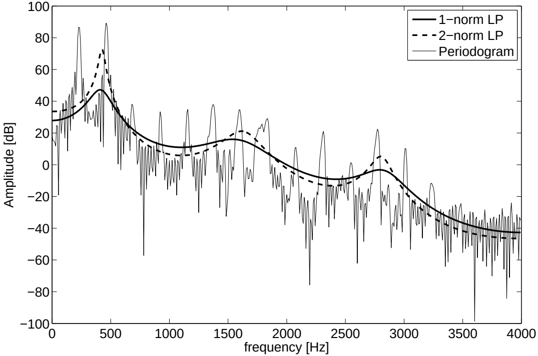
$$\min_{\mathbf{a}} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_1 .$$

- ML approach when the error sequence is considered to be a set of i.i.d. Laplacian random variables.
- Sparser residual beneficial for both analysis and coding purposes.



The spiky train characteristic of voiced speech is retrieved more accurately when we look for a sparse residual.

Sparse Linear Prediction
Finding a Sparse Residual - Example



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.

- Consider the cascade of a short-term linear predictor $F(z)$ and a long-term linear predictor $P(z)$ to remove respectively near-sample redundancies:

$$A(z) = \left(1 - \sum_{k=1}^{N_f} f_k z^{-k}\right)\left(1 - \sum_{k=1}^{N_p} g_k z^{-(T_p+k-1)}\right).$$

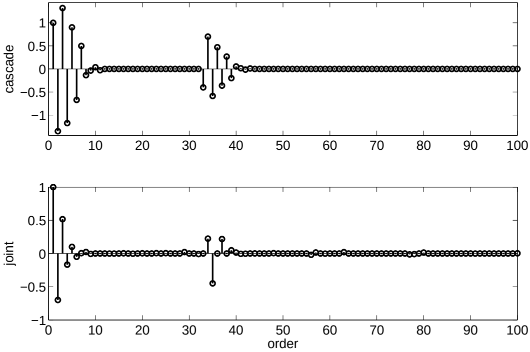
- The resulting prediction coefficient vector $\mathbf{a} = \{a_k\}$ of the high order polynomial $A(z)$ will therefore be highly sparse.
- We can impose sparsity on a high-order predictor:

$$\min_{\mathbf{a}} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_p^p + \gamma\|\mathbf{a}\|_1.$$

- When $p = 2$ minimum variance approach, $p = 1$ encourages sparsity also on the residual.

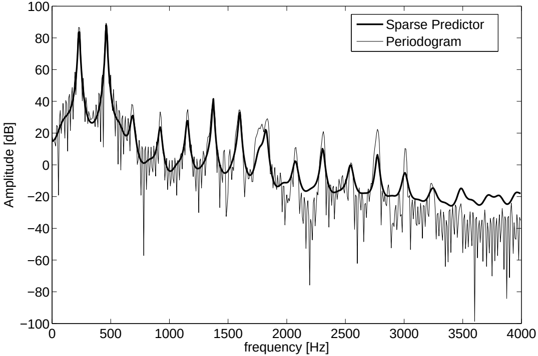
Sparse Linear Prediction

Finding a High-Order Sparse Predictor - Example



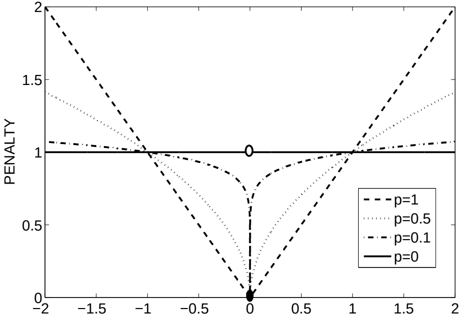
The prediction coefficients vector is similar to the multiplication of the short-term prediction filter and long-term prediction filter usually obtained in cascade.

Sparse Linear Prediction
Finding a High-Order Sparse Predictor - Example



Spectral modeling properties of a high order sparse predictor with only nine nonzero coefficients.

- The purpose of the high order sparse predictor is to model the *whole* spectrum, i.e., the spectral envelope and the spectral harmonics.
- γ controls the sparsity of the prediction coefficient vector. If γ is chosen appropriately, we can obtained again $F(z)$ and $P(z)$ through approximate factorization.
- Intrinsic model order selection!



- Reweighted 1-norm minimization balances the dependence on the magnitude of the 1-norm.
- Changing the cost function and moving the problem towards the 0-norm minimization with convex tools (convergence to the log-sum penalty function).

- If sparse, our solution lies in a subspace of reduced dimensionality where the Euclidean distance between all points in the signal model is preserved.
- Two ingredients needed for CS: a domain where the signal is sparse and the sparsity level T.
- Exploiting knowledge of T a limited number of $M \propto T$ random projections are sufficient to recover our predictors and sparse residual with high accuracy.
- The *shrinkage* of the minimization problem in a lower dimensional space will have a clear impact on the complexity.

- Original MPE problem (known predictor):

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

- CS Formulation:

$$\hat{\mathbf{r}} = \arg\min_{\mathbf{r}} \|\mathbf{r}\|_1 + \gamma \|\Phi\mathbf{x} - \Phi\mathbf{H}\mathbf{r}\|_2^2.$$

- 1-norm global optimization as convex relaxation of the 0-norm: near-optimal selection of sparse excitation.
- Sparsity-knowledge-based *shrinkage*: reduction of constraints $\rightarrow$ computationally faster.

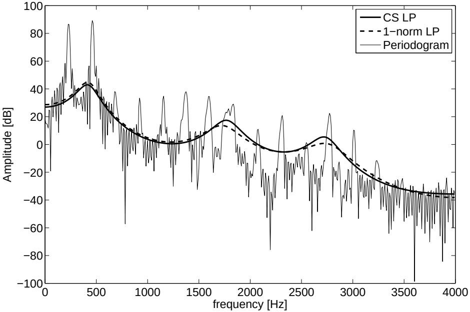
- To adapt CS principles to the estimation of the predictor as well, consider the relation between the synthesis matrix $\mathbf{H}$ and the analysis matrix $\mathbf{A}$ ($\mathbf{A} = \mathbf{H}^{+}$):

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

- Equivalent to our original formulation projected onto a lower-dimensional space.
- When looking for a high order sparse predictor, similarly:

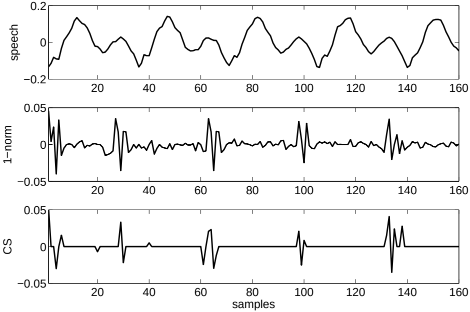
$$\min_{\mathbf{a},\mathbf{r}} \|\mathbf{r}\|_1 + \gamma\|\mathbf{a}\|_1 \quad \text{s.t.} \quad \Phi\mathbf{r} = \Phi(\mathbf{x} - \mathbf{X}\mathbf{a}).$$

- Both formulation can also involve a reweighting procedure.



1-norm solution with and without CS shrinkage (170 equations vs. 80 equations).

Sparse Linear Prediction
Compressed Sensing in Sparse LP - Example



CS recovery of the sparse residual. The imposed sparsity level is $T = 20$, corresponding to the size $M = 80$ for the sensing matrix.

Sparse Linear Prediction
Analysis

- Main advantage is to overcome some of 2-norm LP known issues
  - Lower Spectral Distortion for the spectral envelope estimation.
  - Invariant to small shift of the analysis window.
  - Pitch Independence.

- Synergistic multi-stage coding (sparse predictor $\rightarrow$ sparse encoding).
- Possibility variable rate coding through high-order predictor modeling ($\rightarrow$ model order selection and intrinsic V/UV classification).
- More “fair” distribution between bit allocations on $\mathbf{a}$ and $\mathbf{r}$.
- Less parameters necessary.

Sparse Linear Prediction
Drawbacks

- Stability not guaranteed. Defined new methods to tackle this problem:
  - Reducing the numerical range of the shift operator.
  - Constrained 1-norm based on the alternative Cauchy bound.
- Computational Complexity:
  - Compressed Sensing reduces the number of constraints.
  - Efficient convex optimization algorithms tailor made for LP.
  - Much of the total computational cost in a speech coder is saved by the "one-step" procedure.
- Non-Uniqueness (still optimal!).
- Lack of a Frequency-Domain interpretation.

Sparse Linear Prediction

References

- "Sparse Linear Prediction and Its Applications to Speech Processing," *IEEE T-ASL*, 2012.
- "Retrieving Sparse Patterns Using a Compressed Sensing Framework: Applications to Speech Coding," *IEEE SPL*, 2010.
- "High-Order Sparse Linear Predictors for Audio Processing," *Proc. EUSIPCO*, 2010.
- "Stable 1-norm Error Minimization Based Linear Predictors for Speech Modeling," *IEEE T-ASL*, 2012.
- "Real-time Implementations of Sparse Linear Prediction for Speech Processing," *Proc. ICASSP*, 2013.

Outline

- Minimize distortion between the original signal $\mathbf{x}$ and its synthesized version $\widehat{\mathbf{x}}$ subject to some constraints regarding the rate:

$$\begin{array}{ll} \text{minimize} & D(\mathbf{x}, \widehat{\mathbf{x}}), \\ \text{subject to} & R(\widehat{\mathbf{x}}) \leq R^*. \end{array}$$

where $D(\mathbf{x}, \widehat{\mathbf{x}})$ is the distortion measure, $R(\widehat{\mathbf{x}})$ is the rate used to describe $\widehat{\mathbf{x}}$, and R^* is the maximum possible rate.

- In LPAS coders The distortion $D(\mathbf{x}, \widehat{\mathbf{x}})$ is directly associated with the choice of the predictor $\widehat{\mathbf{a}}$ and the prediction residual $\widehat{\mathbf{r}}$:

$$D(\mathbf{x}, \widehat{\mathbf{x}}) = \|\mathbf{W}(\mathbf{x} - \Upsilon(\widehat{\mathbf{a}})\widehat{\mathbf{r}})\|_2,$$

  where $\mathbf{H} = \Upsilon(\mathbf{a})$ is the synthesis matrix used in the AbS equations (nonlinear transformation of $\mathbf{a}$) and $\Upsilon(\cdot)$ being the nonlinear operator that maps $\mathbf{a}$ into $\mathbf{H}$.
- $\mathbf{W}$ is the matrix that performs the projection in the perceptual domain.

- Distortion is related to the selection of $\widehat{\mathbf{a}}$ and $\widehat{\mathbf{r}}$, we can split the rate accordingly:

$$R(\widehat{\mathbf{x}}) = R(\widehat{\mathbf{a}}) + R(\widehat{\mathbf{r}}).$$

- If we consider the cardinality of the two vectors as a coarse approximation of the rate:

$$R(\widehat{\mathbf{x}}) \cong \alpha\|\widehat{\mathbf{a}}\|_{\mathbf{0}} + \beta\|\widehat{\mathbf{r}}\|_{\mathbf{0}},$$

- we can reformulate the problem as (using Lagrange Multipliers):

$$\widehat{\mathbf{a}}, \widehat{\mathbf{r}} = \arg\min_{\mathbf{a},\mathbf{r}} \|\mathbf{W}(\mathbf{x} - \Upsilon(\mathbf{a})\mathbf{r})\|_2^2 + \gamma(\alpha\|\mathbf{a}\|_0 + \beta\|\mathbf{r}\|_0).$$

- The problem is nonconvex, and nonlinear! Use of convex relaxation and alternate minimization to solve it.

Alternate Minimization Procedure
Estimating Sparse High-Order LP and Residual

- Initial $\mathbf{a}$ estimate via Sparse LP, determine $\mathbf{H}$.
- Exploiting prior knowledge on the sparsity:

$$\widehat{\mathbf{r}} = \arg\min_{\mathbf{r}} \|\mathbf{r}\|_1 \quad \text{s.t.} \quad \Phi\mathbf{x} = \Phi\mathbf{H}\mathbf{r}.$$

- Known $\mathbf{r}$, we estimate $\mathbf{a}$:

$$\widehat{\mathbf{a}} = \arg\min_{\mathbf{a}} \|\mathbf{x} - \Upsilon(\mathbf{a})\mathbf{r}\|_2^2 + \chi\|\mathbf{a}\|_0,$$

- Calculate a minimum variance approximation of the impulse response:

$$\widehat{\mathbf{H}} = \min_{\mathbf{H}} \|\mathbf{x} - \mathbf{H}\mathbf{r}\|_2^2 \quad \leftrightarrow \quad \widehat{\mathbf{h}} = \min_{\mathbf{h}} \|\mathbf{x} - \mathbf{R}\mathbf{h}\|_2^2$$

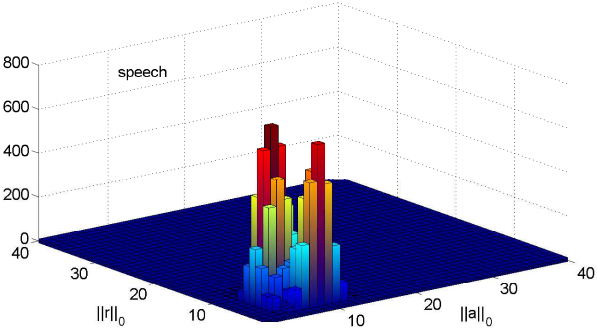
- Ultimately, recalculate the predictor:

$$\widehat{\mathbf{a}} = \arg\min_{\mathbf{a}} \|\mathbf{a}\|_1 \quad \text{s.t.} \quad \Psi\mathbf{t}'' + \Psi\mathbf{T}''\mathbf{a} = \mathbf{0}.$$

where $\mathbf{AT} = \mathbf{I}$, $\mathbf{t} = \widehat{\mathbf{h}} = \mathbf{R}^{-1}\mathbf{x}$ is the minimum norm solution $\mathbf{x} - \mathbf{Tr} = \mathbf{0}$.

Alternate Minimization Procedure
Observations

- Needs a priori knowledge of the sparsity of **a** and **r** to determine the size of the random matrices for the CS formulations.
- Creates extremely sparse solutions.



Application of sparsity in LPAS

Frame Dependent/Independent Coding

- We apply the rate/distortion sparse approach to the problem of speech coding robust to packet loss.
- An approach to cope high packet loss in VoIP is to create speech coders that are totally frame independent.
- In the case of telephony with dedicated circuits, high quality is achieved by the exploitation of inter-frame dependencies.

Application of sparsity in the LPAS

Frame Dependent/Independent Coding

- Frame independent are created cope with packet loss
- Frame dependent coders achieve high quality by the exploitation of inter-frame dependencies.
- Splitting the information present in each speech packet into two components:
  - one to independently decode the given speech frame,
  - one to enhance it by exploiting inter-frame dependencies.
- We formulate the problem as:

$$
\begin{aligned}
\min. \quad & D(\mathbf{x}, \widehat{\mathbf{x}}^{FI} + \widehat{\mathbf{x}}^{EN}) \\
\text{s.t.} \quad & R(\widehat{\mathbf{x}}^{FI}) + R(\widehat{\mathbf{x}}^{EN}) \leq R^*.
\end{aligned}
$$

- This problem in a LPAS framework can then be formulated as:

$$\arg\min_{\mathbf{a},\mathbf{r}} \|\mathbf{x} - \Upsilon(\mathbf{a}^{FI} + \mathbf{a}^{EN})(\mathbf{r}^{FI} + \mathbf{r}^{EN})\|_2 + \chi\|\mathbf{a}^{FI} + \mathbf{a}^{EN}\|_0 + \delta\|\mathbf{r}^{FI} + \mathbf{r}^{EN}\|_0.$$

- The frame independent parameters are calculated without state memory:

$$\hat{\mathbf{a}}^{\mathbf{FI}}, \hat{\mathbf{r}}^{\mathbf{FI}} = \arg\min_{\mathbf{a},\mathbf{r}} \|\mathbf{x} - \Upsilon(\mathbf{a}^{FI})\mathbf{r}^{FI}\|_2 + \chi\|\mathbf{a}^{FI}\|_0 + \delta\|\mathbf{r}^{FI}\|_0.$$

- The frame dependent parameters are calculated with state memory ($\mathbf{r}^{FD} = \left[\hat{\mathbf{r}}_-^T, \mathbf{r}^T\right]^T$):

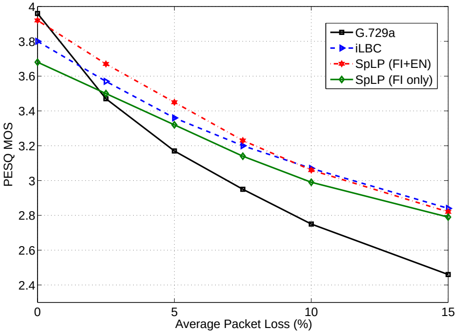
$$\hat{\mathbf{a}}^{\mathbf{FD}}, \hat{\mathbf{r}}^{\mathbf{FD}} = \arg\min_{\mathbf{a},\mathbf{r}} \|\mathbf{x} - \Upsilon(\mathbf{a}^{FD})(\mathbf{r}^{FD})\|_2 + \chi\|\mathbf{a}^{FD}\|_0 + \delta\|\mathbf{r}^{FD}\|_0.$$

Frame Dependent/Independent Coding
General Behavior

- The reconstructed speech for the frame independent case:
$$\widehat{\mathbf{x}}^{FI} = \Upsilon(\mathbf{a}^{FI})\widehat{\mathbf{r}}^{FI},$$
- For the frame dependent case:
$$\widehat{\mathbf{x}}^{FD} = \Upsilon(\mathbf{a}^{FD}) \left[(\widehat{\mathbf{r}}^{FD}_{-1})^T, (\widehat{\mathbf{r}}^{FD})^T \right]^T.$$
- We transmit the frame independent parameters ($\widehat{\mathbf{r}}^{FI}$, $\hat{A}^{FI}(z)$) and a side stream with the differences between the two predictors $\hat{A}^{EN}(z)$ and the differences between the two residuals $\widehat{\mathbf{r}}^{EN}(z)$.
- Multipulse encoding.

Frame Dependent/Independent Coding

Results



Performances of the compared methods: G.729a (8 kbps), iLBC (13.33 kbps), and our method (SpLP) with (FI+EN) and without (FI) the frame dependent enhancement layer (respectively 10.9 and 7.65 kbps).

Sparse Linear Prediction
References

- "Re-estimation of Linear Predictive Parameters in Sparse Linear Prediction" *Asilomar SSC*, 2010.
- "Estimation of Frame Independent and Enhancement Components for Speech Communication over Packet Networks," *ICASSP*, 2010.
- "A Sparse Representation in Linear Predictive Analysis-by-Synthesis for Frame Independent Speech and Audio Coding," *KU Leuven Internal Report*, 2010.

Outline

Conclusions
Sparsity in the LP optimization framework

- Analysis: a more efficient decoupling between the pitch harmonics and the spectral envelope.
- Coding: a more straightforward approach to encode a speech segment.
- Sparse LP applied successfully also in audio processing, transcending some of the limitation of traditional LP.

Conclusions
Sparsity in the LPAS framework

- New method for the estimation of the parameters in speech coding, creating a new meaning for the LP parameters.
- Providing tradeoffs between the complexity, and thus the bit-rate, of the two descriptions.
- Highly flexible: possibility of estimating predictors and residuals that create a independently decodable frames of audio and speech.

Acknowledgements

- Prof. Søren Holdt Jensen (adviser)
- Prof. Marc Moonen (adviser)
- Prof. Mads Græsbøll Christensen (estimation theory and sparse representation)
- Prof. Manohar N. Murthi (speech modeling and coding)
- Dr. Joachim Dahl (convex optimization)
- Dr. Toon van Waterschoot (extension to audio processing and acoustics, e.g., RTF)