Sparsity in Linear Predictive Coding of Speech

Daniele Giacobello

Aalborg Universitet, Denmark

September 17th, 2010

- LP is, arguably, the most successful tools for the analysis and coding of speech signals.
- Analysis: correspondence with modeling the speech production process.
- Coding: interesting attributes like low delay, scalability and, in general, low complexity.
- Fundamental part of many coding architectures since the early works on speech coding to the most recent proposals for unified speech and audio coders.

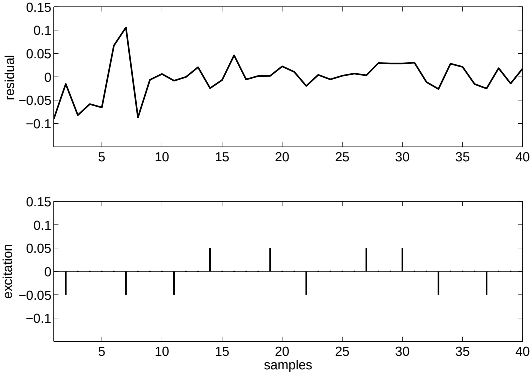
- 2-norm minimization is widely used in inverse problems.
- Amenable of producing an optimization problem that is attractive both theoretically and computationally.
- Consistent with producing a representation with minimal energy.

Motivation
Why sparsity?

- In many signal processing applications it is more beneficial to find solutions with the fewest nonzero coefficients as possible, a maximal *sparse* solution.
- Grown significantly in the recent years do to the increasing use of transform domain representations.
- Large number of mathematical tools available for sparse approximation.
- A concise signal representation in a given domain is generally required for efficient coding.

Motivation
Why sparsity in Linear Prediction?

- Main Idea: reduce the mismatch between a “white noise”-like prediction residual and a *sparse* approximation.



Motivation
Why sparsity in LP analysis and coding?

- Voiced speech generally modeled as an impulse train, i.e., a *sparse* sequence.
- A concise signal representation arguably related to more efficient coding.
- Why not! ...new formulations for LP-based approaches may be of general interest (e.g., Audio, ECG, etc.)

Outline

Introduction

Background

LPAS Coding

Sparsity in Signal Processing

Contributions

Conclusions

Outline

- The theory behind the widespread use of LP all-pole modeling of speech, arises from the source-filter model of speech production.
- Emitted speech sound is a combination of the excitation process (the air flow) and the filtering process (vocal tract effect).

- Two different ways in which speech sounds are produced.
- Voiced speech:
  - the airflow is periodically inhibited for short intervals by the vocal folds,
  - this periodicity f_0 contributes to the perceived *pitch*,
  - strong periodic components rich in harmonics.
- Unvoiced speech:
  - the airflow is constricted or completely stopped for a short interval,
  - source has noise-like or impulsive-like characteristics without harmonic structure.

Speech Production
Mathematical Model

- First attempts to provide a mathematical model for speech production in acoustics rather than signal processing.
- Clear relation between the physics of speech production and the theory of sound wave propagation in acoustic cavities.
- These early works suffered quite consistently from high requirements on specific a priori knowledge of the voice.

Bishnu Atal's Work

Simplify the Speech Production Theory

B. S. Atal, "Determination of the Vocal-Tract Shape Directly from the Speech Wave," *J. of the Acoustical Society of America*, **1970**.

- Approximated the vocal tract with a lossless tube made by cylindrical sections of equal length but different diameter.
- Exploited the relations of the lossless tube model with digital filters.
- Vocal tract $\rightarrow$ K poles transfer function (number of sections of the lossless tube).

B. S. Atal ans S. L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," *J. of the Acoustical Society of America*, **1971**.

- Introduced the discrete speech production model.
- Speech signal is analyzed and synthesized as the output of a discrete linear all-pole time-varying filter.
- The excitation is either a periodic pulse train or a white noise sequence.

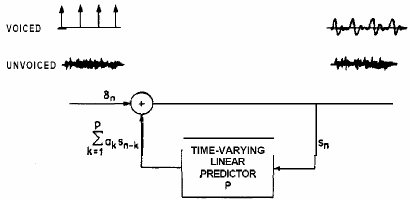


FIG. 1. Block diagram of a functional model of speech production based on the linear prediction representation of the speech wave.

B. S. Atal ans M. R. Schroeder, "Predictive Coding of Speech," *Proc. Conference on Communication*, **1967**.

- Concept of *predictive coding* to *decorrelate* a speech segment by applying a order K prediction filter.
- The idea of predictive coding of speech came before the relations with the mathematical model of speech production!

- Atal linked these two theories: the prediction filter is theoretically consistent with the speech production model.
- The corresponding order K all-pole model carries the information of the tube model of the vocal tract.
- Summarizing:
  Vocal tract $\rightarrow$ K poles transfer function $\rightarrow$ K order predictor.

LP Based Speech Analysis
The original sin...

- In Atal's work, the all-pole coefficients are identified by minimizing the mean-squared (2-norm) error of the difference between the observed signal and the predicted signal:

$$\frac{\partial E\left[|e(n)|^2\right]}{\partial a_k} = 0,$$

where:

$$e(n) = x(n) - \sum_{k=1}^{K} a_k x(n-k), \quad 0 < n \leq N.$$

- This forms the *Yule-Walker* equations for *autoregressive* (AR) model fitting (solver: Levinson recursion).

LP Based Speech Analysis
Unvoiced Speech

- The inverse of K-order LP *analysis* filter represents the vocal tract transfer function.
- The prediction error (the *residual* signal) represents the source.
- Unvoiced speech lends itself readily to the principles of the 2-norm error criterion (white noise excitation).
- Itakura also provided the statistical interpretation of the 2-norm minimization with the fitting of the error in a Gaussian i.i.d. distribution.

LP Based Speech Analysis
Voiced Speech

- In voiced speech, this approach is questionable and, theoretically not well founded:
  - The all-pole spectrum does not provide a good spectral envelope,
  - does not provide a good approximate of the harmonics amplitude.
- The LP tries to cancel the input voiced speech harmonics:
  - the LP spectrum tends to overestimate the spectral powers at the formants,
  - sharper contour than the original vocal tract response.

- Atal was able to reduce the entropy of a 5 ms speech segment sampled at 6.67 kHz from 3.3 b/sample to 1.3 b/sample by applying a $10th$ order predictor.
- LP is used to decorrelate the input leaving a residual that is ideally white, and easier to quantize.
- Minimizing the 2-norm of the residual consistent with the fundamental theorem of predictive quantization.
- Achieving minimal variance of the residual $\longrightarrow$ efficient coding.

LP Based Speech Coding
Traditional Usage

- Modeling only the the spectral envelope, capturing the short-term redundancies.
- The excitation is usually estimated with some constrained structure on it.
- *Sparse* techniques are usually employed to model the excitation for efficient coding.
- Examples since early works on speech coding (MPE) to the currently deployed sparse algebraic codewords (ACELP).

Why 2-norm based LP still so popular?

- To the author's best knowledge, the 2-norm is the only criterion in LP used in commercial speech codecs.
- In 40 years, no one has noted the 2-norm based LP shortcomings?

Why 2-norm based LP still so popular?

- **1967**: B. S. Atal and M. R. Schroeder, "**Predictive Coding of Speech**."

...nothing happened...

- **2008**: D. Giacobello, M. G. Christensen, J. Dahl, S. H. Jensen, and M. Moonen, "**Sparse Linear Predictors for Speech Processing**."

...really?

Why 2-norm based LP still so popular?

...no, not really!

- A rich literature exists addressing the deficiencies of 2-norm based LP in speech analysis and coding.
- Several explanations for the 2-norm popularity, going around the same concept: simplicity.

Why 2-norm based LP still so popular?
Mathematical Tractability

- The minimization of the 2-norm of the prediction error results in the Yule-Walker equations and can be efficiently solved via the Levinson recursion.
- The 2-norm cost function is strongly convex allowing for a unique solution.
- The roots of the corresponding all-pole filter are guaranteed to be inside the unit circle (stability is intrinsically guaranteed by the construction of the problem).

Why 2-norm based LP still so popular?

Statistical Interpretation

- Correspondence to the *Maximum Likelihood* (ML) approach when the error signal is considered to be a set of i.i.d. Gaussian variables.
- The Gaussian p.d.f. is arguably the most used and well know distribution for tractable mathematics.
- The Yule-Walker equations can be derived from the maximum likelihood approach.

Why 2-norm based LP still so popular?

Frequency-Domain Interpretation

- Minimizing the 2-norm of the error in the time-domain is equivalent to minimizing the error ratio between the true and estimated spectra (Parseval's Theorem).
- Minimizing the squared error in the time domain and in the frequency domain leads in both cases to the Yule-Walker equations.

Outline

- Coding paradigm that has set the standard for speech coding for the past 30 years.
- Three main stages of the LPAS coding paradigm:
  - Linear predictive analysis,
  - Modeling of the excitation,
  - Modeling of the pitch periodicity.

- 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. Assuming that $x(n) = 0$ for $n < 1$ and $n > N$:

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

with $N_1 = 1$ and $N_2 = N + K$:

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

Linear Predictive Analysis
Minimization Problem

- The coefficients are found by minimizing the prediction error:

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_p^p,$$

where $\|\cdot\|_p$ is the p-norm.

- When $p = 2$ ($N_1 = 1$ and $N_2 = N + K$) *autocorrelation method*:

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_2^2 = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{x}.$$

- Corresponds to solving the Yule-Walker equations.

The Excitation Model

General Formulation

- Key step of the analysis-by-synthesis procedure.
- LP coefficients $\widehat{\mathbf{a}}$ calculated in a open-loop configuration.
- The choice of the excitation $\widehat{\mathbf{r}}$ is done in a close-loop configuration (so the name analysis-by-synthesis):

$$\widehat{\mathbf{r}} = \arg\min_{\mathbf{r}} \|\mathbf{W}(\mathbf{x} - \mathbf{H}\mathbf{r})\|_2^2, \quad \text{s.t.} \quad \mathbf{struct}(\mathbf{r}).$$

- $\mathbf{H}$ is a convolution matrix, called the *synthesis matrix* (obtained from $\widehat{\mathbf{a}}$).
- $\mathbf{W}$ is the perceptual weighting matrix (obtained from $\widehat{\mathbf{a}}$).
- $\mathbf{struct}(\cdot)$ represents the structural constraints imposed on the excitation (determine the encoding strategy).

- In multipulse encoding (MPE) coders, the excitation consists of K freely located pulses in each segment of length N.
- MPE provides an approximation to the optimal approach, when all possible combinations of K positions in the approximated residual of length N are analyzed, i.e.:

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

- Significant amount of bits to be spent on describing their location on the excitation sequence.
- In regular-pulse encoding (RPE) the pulses are constrained on a grid with spacing S.

- The RPE can be considered as the first idea to include a predetermined structure on the excitation.
- This idea has also been developed, around the same time, in code-excited LP (CELP).
- Sequence selected by a predetermined codebook populated by "random white noise" sequences:

$$\hat{\mathbf{r}} = \arg\min_{\mathbf{c}} \|\mathbf{W}(\mathbf{x} - \mathbf{Hc})\|_2^2, \quad \text{s.t.} \quad \mathbf{c} \in C.$$

- The general idea, is also to have the sequences pre-quantized, thus truly selecting the optimal sequence to be sent to the encoder.
- Computationally heavy! Algebraic codebooks (ACELP) are used instead (simple algebra rather than look-up tables).

Modeling the Pitch Periodicity

Accounting for Pitch Periodicity

- Voiced speech segments exhibits strong long-term correlation components due to the presence of a pitch excitation.
- The strategies presented to model the excitation do not exploit this (quasi-) periodicity.
- To account for these correlations, two strategies can be implemented:
  - find a long-term linear predictor,
  - account for the periodicity directly in the excitation model.

- First attempt implemented to account for long-term correlations.
- This interpretation is similar to modeling the short-term correlations.
- The common choice is:

$$P(z) = 1 - g_p z^{-T_p}.$$

- g_p and T_p determined by minimizing the residual error signal after the LP predictor.
- T_p is approximately the inverse of f_0 (fundamental frequency).
- The frequency response of $P(z)$ is a comb-like structure, thus resembling a line spectrum, consistent with the harmonic structure of the voiced speech sounds.

Modeling the Pitch Periodicity
Adaptive Codebook

- Account for the periodicity in the modeled excitation.
- The excitation can be seen as a linear combination between a pseudo-random component $\mathbf{c}_f$, and a periodic component given by the pitch excitation $\mathbf{c}_a$:

$$\widehat{\mathbf{r}} = g_f \mathbf{c}_f + g_a \mathbf{c}_a,$$

  where $\mathbf{c}_f$ is the *fixed* codeword ($\mathbf{c}_f \in C_f$) and $\mathbf{c}_a$ is the *adaptive* codeword ($\mathbf{c}_a \in C_a$), g_f and g_a are the gains.
- Begin with the search for the adaptive codebook (open-loop estimate of T_p).
- The adaptive codeword is built up based on the "refined" pitch period estimate T_p and its gain.

Outline

- Many natural signals have a concise representation when expressed in the proper basis or a dictionary of elementary building blocks.
- When this sparse representation is truncated in a suitable way, high precision approximations are obtained even when very few terms are retained.
- First works where sparsity was successfully applied was indeed speech coding: speech of any desired quality can be obtained providing a sufficient number of pulses at the input of the synthesis filter.
- Sparse approximation approaches have enjoyed considerable popularity in recent signal processing applications.

- The canonical form of the problem:

$$\min_{\mathbf{x}} \|\mathbf{x}\|_0, \quad \text{s.t.} \quad \mathbf{Ax} = \mathbf{b},$$

  where $\mathbf{A} \in \mathbb{R}^{N \times M}$ represent an overcomplete basis.
- If $\mathbf{x}$ is K-sparse ($K \ll M$) only K entries in $\mathbf{x}$ are sufficient to reconstruct $\mathbf{b}$ without distortion.
- Accounting for modeling errors or measurement:

$$\min_{\mathbf{x}} \|\mathbf{x}\|_0, \quad \text{s.t.} \quad \|\mathbf{Ax} - \mathbf{b}\|_2^2 \leq \epsilon.$$

- Combinatorial problems! Search for the optimal K-sparse representation would require solving up to $\binom{M}{K}$ linear systems.
- Need for approximate solutions.

Algorithms for Approximate Sparse Solutions
Greedy Methods

- Greedy methods "break" the optimization problem in a sequence of smaller problems in which a optimal solution can be easily found.
- Matching pursuit type algorithm (MP and OMP) iteratively solve the sparse approximation problem applying a sequence of locally optimal choices in an effort to determine a globally optimal solution.
- The procedure usually terminates when the given sparsity level K is achieved.
- Limitations: optimizing over a series of K sub-problems, generally not converge to a globally optimal solution.

Algorithms for Approximate Sparse Solutions

Minimization of Diversity Measures

- Replace the combinatorial problem with a related convex program (relaxation).
- Differently from greedy algorithms, it is based on global optimization, thus, in general, finds improved sparse solutions.
- The 1-norm is chosen as closest convex approximation of the 0-norm:

$$\min_{\mathbf{x}} \|\mathbf{x}\|_1, \quad \text{s.t.} \quad \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 \leq \epsilon.$$

- Recent algorithms based on the 1-norm find more focal solutions (reweighted schemes).
- This category is the one that has received the most interest lately (significant improvements in convex optimization algorithms).

Outline

Sparse Linear Prediction
Fundamentals 1/2

- Consider the speech production model in matrix form:

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

- 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* **a** is (prior knowledge of **a**).
- 1-norm used as a convex relaxation of the 0-norm:

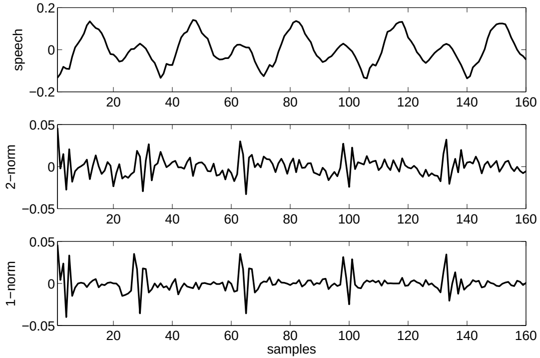
$$\min_{\mathbf{a}} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_1 + \gamma\|\mathbf{a}\|_1.$$

Sparse Linear Prediction
Finding a Sparse Residual

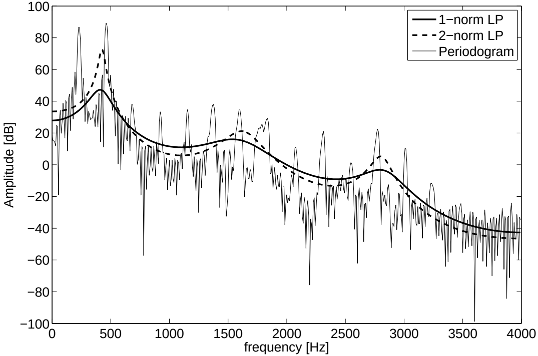
- Consider now the case of a short-term predictor that engenders a sparse residual ($\gamma = 0$):

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

- Better statistical fitting: ML approach when the error sequence is considered to be a set of i.i.d. Laplacian random variables.
- Helpful against over-emphasis on peaks in the envelope estimation: outperforms the 2-norm in finding a more proper linear predictive representation in voiced speech.
- Sparser residual beneficial also in unvoiced speech.



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



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.

Aalborg Universitet, Denmark

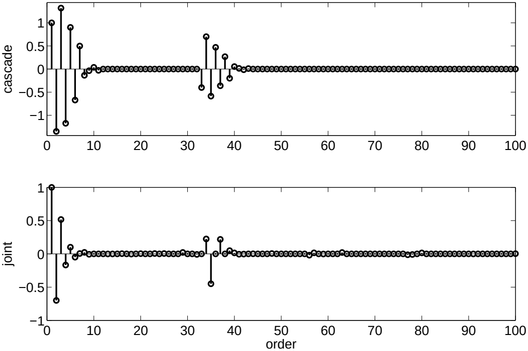
- Consider the cascade of a short-term linear predictor $F(z)$ and a long-term linear predictor $P(z)$ to remove respectively near-sample redundancies:

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

- 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:

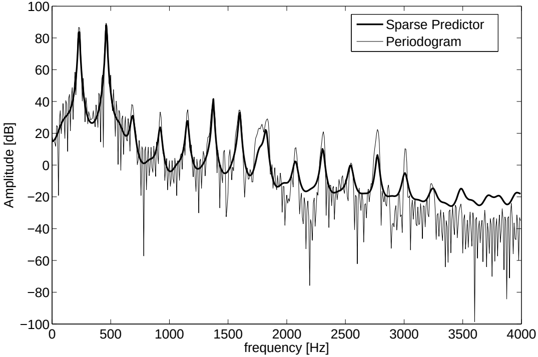
$$\min_{\mathbf{a}} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_p^p + \gamma\|\mathbf{a}\|_1.$$

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



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

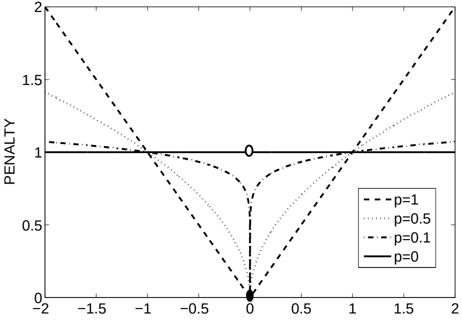
Sparse Linear Prediction
Finding a High-Order Sparse Predictor - Example



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.
- Different approaches to select γ:
  - fixed,
  - adaptive (behavior of γ strictly related to how *voiced* the segment is),
  - "optimal" $\approx L-$curve
- If γ is chosen appropriately, we can obtained again $F(z)$ and $P(z)$ thorough approximate factorization.
- Intrinsic model order selection!

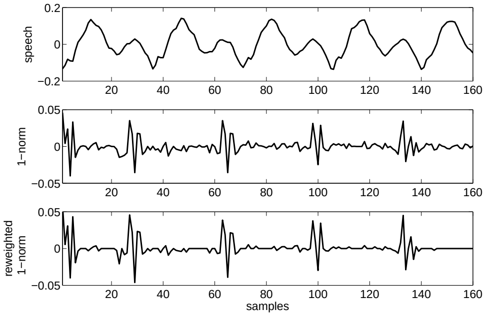
Sparse Linear Prediction
Reducing the 1-norm 0-norm mismatch



- 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).

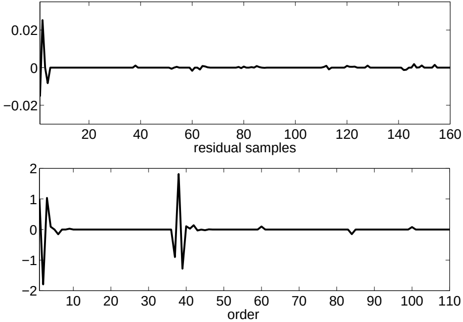
Sparse Linear Prediction

Reweighted 1-norm - Example



Five iterations of reweighted 1-norm help enhancing sparsity.

Sparse Linear Prediction
Reweighted 1-norm - Example



Residual and High-Order Predictor after five iterations of the reweighting algorithm.

- 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.

- Retrieving the sparse residual with a *known* predictor:

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

or

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

- The problem is projected onto a lower dimensional space by the random basis Φ of dimension $M \times N$ ($M \propto T$).

- 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 → computationally faster.

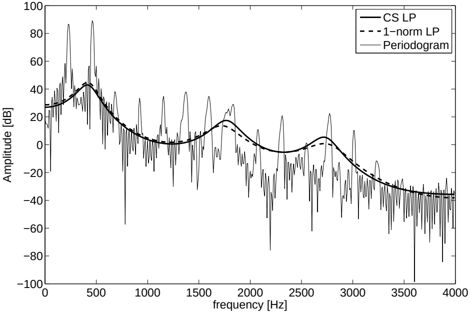
- To adapt CS principles to the estimation of the predictor as well, consider the relation between the synthesis matrix $\mathbf{H}$ and the analysis matrix $\mathbf{A}$ ($\mathbf{A} = \mathbf{H}^{+}$):

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

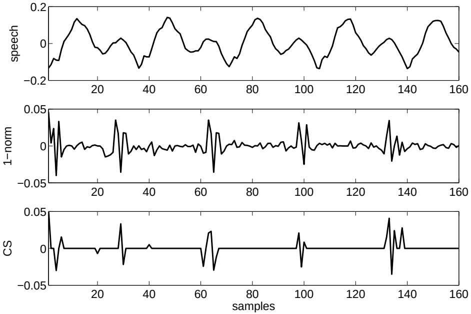
- Equivalent to our original formulation projected onto a lower-dimensional space.
- When looking for a high order sparse predictor, similarly:

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

- Both formulation can also involve a reweighting procedure.



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



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

- Consequences of the efficient decoupling between the source and the filter.
- Lower Spectral Distortion for the spectral envelope estimation.
- Invariant to small shift of the analysis window.
- Pitch Independence.

- Synergistic multi-stage coding.
- Possibility variable rate coding (model order selection and intrinsic V/UV classification).
- More “fair” distribution between bit allocations on **a** and **r**.
- Less parameters necessary.

Sparse Linear Prediction
Drawbacks

- Stability not guaranteed. Methods to tackle this problem.
  - Existing ones: 1-norm Burg Method, Reweighted 2-norm.
  - 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.
  - 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.

Re-estimation of LP parameters
Motivation

- The linear prediction parameters are first found in a open-loop configuration and then quantized transparently.
- The search for the best excitation (given certain constraints) is then done in a closed-loop configuration...
- ...all the responsibility for the distortion is basically on the residual!
- Sparse Linear Prediction already introduced to reduce the burden on the excitation...
- ...but why not finding the predictor also in a "pseudo" closed-loop configuration?

Re-estimation of LP parameters
Estimation of the impulse response

- The minimization problem to find the residual, can be rewritten as:

$$\widehat{\mathbf{H}} = \arg\min_{\mathbf{H}} \|(\mathbf{x} - \mathbf{H}\widehat{\mathbf{r}})\|_2^2 \rightarrow \widehat{\mathbf{h}} = \arg\min_{\mathbf{h}} \|(\mathbf{x} - \widehat{\mathbf{R}}\mathbf{h})\|_2^2.$$

- This means that given the residual $\widehat{\mathbf{r}}$, we can find the truncated impulse response that generates the speech segment:

$$\mathbf{x} = \widehat{\mathbf{R}}\mathbf{h}.$$

- It is clear that the "optimal" sparse linear predictor $A(z)$ is the one that has $\mathbf{h}$ as truncated impulse response.

Re-estimation of LP parameters
Least squares approximation of the impulse response

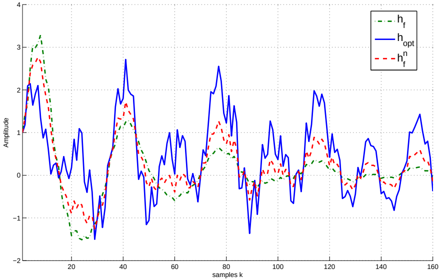
- Assuming $\mathbf{h}_f$ the impulse response of the short-term predictor $1/F(z)$ and $\mathbf{h}_p$ the impulse response of the long-term predictor $1/P(z)$, we can rewrite the problem as:

$$\widehat{\mathbf{H}}_f, \widehat{\mathbf{H}}_p = \arg\min_{\mathbf{H}_f, \mathbf{H}_p} \|(\mathbf{x} - \mathbf{H}_f \mathbf{H}_p \widehat{\mathbf{r}})\|_2.$$

- We can then proceed with the re-estimation of the impulse response of the short-term predictor by solving the problem:

$$\widehat{\mathbf{h}}_f = \arg\min_{\mathbf{h}_f} \|(\mathbf{x} - (\mathbf{H}_p \widehat{\mathbf{R}}) \mathbf{h}_f)\|_2.$$

- Finally, find the IIR filter predictor that approximates $\widehat{\mathbf{h}}_f$ through least squares (Y-W eq.). This guarantees stability and simplicity of the solution.



The different impulse responses: original, the re-estimated adapted to the quantized residual, and the approximated impulse response of the new short-term predictor.

Re-estimation of LP parameters
Properties

- A more "crude" approach to speech coding:

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

- More fair distribution between complexity (and bit allocated) of the two descriptions.
- Improvement in the general performances of the Sparse Linear Prediction framework.

Frame Dependent/Independent Coding
Motivation

- 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.
- Overcoming this mismatch by 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.
- Achieving this by exploiting the flexibility of Sparse LP and the re-estimation procedure!

- A sparse linear predictive framework is employed to achieve a more compact description of all the features extracted from a speech frame:

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

- The sparse structure of the high order predictor allows a joint estimation of a short-term and a long-term predictors $A(z) \approx \hat{F}(z)\hat{P}(z)$, the sparse residual allows for efficient sparse encoding.

- We rethink the analysis-by-synthesis (AbS):

$$\widehat{\mathbf{r}} = \arg\min_{\mathbf{r}} \|\mathbf{W}(\mathbf{x} - \widehat{\mathbf{H}} \left[\widehat{\mathbf{r}}_{-}^{T}, \mathbf{r}^{T}\right]^{T})\|_2$$
$$\text{s.t.} \quad \mathbf{struct}(\mathbf{r}).$$

- The residual term $\left[\widehat{\mathbf{r}}_{-}^{T}, \mathbf{r}^{T}\right]^{T}$ is composed of the K previous residual samples $\widehat{\mathbf{r}}_{-}$ (the filter memory, already quantized) and the current $N \times 1$ residual vector $\mathbf{r}$ that has to be estimated.
- Find two residual estimates $\widehat{\mathbf{r}}^{FD}$ (using $\widehat{\mathbf{r}}_{-}$) and $\widehat{\mathbf{r}}^{FI}$ (not using $\widehat{\mathbf{r}}_{-}$).

Frame Dependent/Independent Coding

System Architecture: Re-estimation of LP coefficients

- With $\widehat{\mathbf{r}}^{FI}$ and $\widehat{\mathbf{r}}^{FD}$, we calculate the truncated impulse response that generates them:

$$\widehat{\mathbf{h}} = \arg\min_{\mathbf{h}} \|(\mathbf{x} - \widehat{\mathbf{R}}\mathbf{h})\|_2.$$

- We can split the two contribution as:

$$\hat{A}(z) = \hat{F}(z)\hat{P}(z) \rightarrow \widehat{\mathbf{H}} = \widehat{\mathbf{H}}_f \widehat{\mathbf{H}}_p,$$

  and re-estimate only the short-term impulse response.
- We can then obtain two estimates of the impulse responses, a frame dependent one $\widehat{\mathbf{h}}_f^{FD}$ and a frame independent one $\widehat{\mathbf{h}}_f^{FI}$.
- Autoregressive modeling of $\widehat{\mathbf{h}}^{FD}$ and $\widehat{\mathbf{h}}^{FI}$ and obtain two new short-term predictive filters $\hat{F}^{FI}(z)$ and $\hat{F}^{FD}(z)$.

- The reconstructed speech for the frame independent case:

$$\hat{\mathbf{x}}^{FI} = \hat{\mathbf{H}}_p \hat{\mathbf{H}}_f^{FI} \hat{\mathbf{r}}^{FI},$$

- For the frame dependent case:

$$\hat{\mathbf{x}}^{FD} = \hat{\mathbf{H}}_p \hat{\mathbf{H}}_f^{FD} \left[(\hat{\mathbf{r}}_{-}^{FD})^T, (\hat{\mathbf{r}}^{FD})^T \right]^T.$$

- We transmit the frame independent parameters ($\hat{\mathbf{r}}^{FI}$, $\hat{A}^{FI}(z) = \hat{P}(z)\hat{F}^{FI}(z)$) and a side stream with the differences between the two short-term predictors $\hat{F}^{\Delta}(z)$ and the differences between the two residuals $\hat{\mathbf{r}}^{\Delta}(z)$.

Frame Dependent/Independent Coding

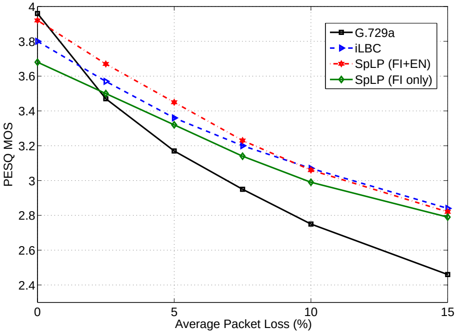
Working mode

- If there is no loss of speech packets:

$$\widehat{\mathbf{x}} = \widehat{\mathbf{H}}_p(\widehat{\mathbf{H}}_f^{FI} + \widehat{\mathbf{H}}_f^{EN}) \left[(\widehat{\mathbf{r}}_-^{FI} + \widehat{\mathbf{r}}_-^{EN})^T, (\widehat{\mathbf{r}}^{FI} + \widehat{\mathbf{r}}^{EN})^T\right]^T$$

  where $\widehat{\mathbf{H}}^{EN}$, $\widehat{\mathbf{r}}_-^{EN}$ and $\widehat{\mathbf{r}}^{EN}$ are functions of the parameters used to define the enhancement layer $\hat{F}^{\Delta}(z)$ and $\widehat{\mathbf{r}}^{\Delta}(z)$.
- When a $k-$th frame is missing, the $k+1-$th frame is self-constructed only from the frame independent parameters.
- The $k+2-$th frame will be reconstructed using the frame dependent information (convert the part of the residual of the $k+1-$th frame $\widehat{\mathbf{r}}_-^{FI}$, that will appear in the reconstruction equation, into the frame dependent one $(\widehat{\mathbf{r}}_-^{FI} + \widehat{\mathbf{r}}_-^{EN})$).

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).

- Scheme representative of a more general problem:

$$\begin{array}{rl} \min. & w_{p_L} D(\mathbf{x}, \widehat{\mathbf{x}}^{FI}) + (1 - w_{p_L}) D(\mathbf{x}, \widehat{\mathbf{x}}^{FI} + \widehat{\mathbf{x}}^{EN}) \\ \text{s.t.} & R(\widehat{\mathbf{x}}^{FI}) + R(\widehat{\mathbf{x}}^{EN}) \leq R^*. \end{array}$$

- The expected distortion will be proportional to the different bit allocations.
- $w_{p_L} \propto p_L$ $(0 \leq w_{p_L} < 1)$.
- The bit allocated for the enhancement layer can be also used to bring information for the packet loss concealment on how to reconstruct the missing frames when the loss rate is high.

Outline

Conclusions

Sparsity in the LP optimization framework

- Analysis: a more efficient decoupling between the pitch harmonics and the spectral envelope.
- Coding: a synergistic new 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

- A computationally efficient near-optimum multipulse approach using CS.
- New method for the re-estimation of the prediction 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.
- Possibility of estimating predictors and residuals that create a independently decodable frame of speech.

Outlook

Provide a Common Coding Framework for Speech and Audio Coding

- Current approach: switch between LP and MDCT coding.
- LP filter is generally a quite adequate tool to model the spectral peaks which play a dominant role in perception.
- Low delay, scalability and low complexity make the extension of LP to audio coding also appealing.
- High-order sparse linear predictors for audio and speech processing:
  - attractive in modeling the harmonic behavior of audio and speech signals,
  - concise parametric representation by exploiting harmonicity,
  - accurate spectral modeling consistent with high-order LP.
- Exploiting CS to reduce the computational complexity in audio, wideband and super-wideband speech.

Outlook
Redefine the LPAS Coding Scheme

- In the AMR-WB coder (23.85 kbit/s configuration) 80% of the bits are allocated for the excitation and only 10% for the predictor.
- The re-estimation procedure for the predictor was proposed to find a tradeoff between the complexity of the excitation and the complexity of the predictor.
- Tradeoff of the sparse representation of the excitation and the sparse representation of the high-order sparse predictor can also be considered.
- Arguably, a clear relation between sparsity and rate:

$$R(\widehat{\mathbf{x}}) \approx \alpha\|\widehat{\mathbf{a}}\|_0 + \beta\|\widehat{\mathbf{r}}\|_0.$$

Outlook

Provide Frame Independent Coding

- PLC strategies have achieved a certain degree of maturity.
- It is still important to reduce, if not eliminate, inter-frame dependencies making each frame independently decodable.
- The coding algorithm we have presented divides the representation of the speech segment is divided between a frame independent core and a frame dependent enhancement layer.
- The distortion term can be made dependent on the loss rate and therefore adjusting the bit allocation on the frame dependent and frame independent parts.

Acknowledgements

- Prof. Søren Holdt Jensen
- Prof. Marc Moonen
- Prof. Mads Græsbøll Christensen
- Prof. Manohar N. Murthi
- Dr. Joachim Dahl
- Dr. Toon van Waterschoot
- The whole MISP Section and the SIGNAL people!