Re-estimation of Linear Predictive Parameters in Sparse Linear Prediction

Daniele Giacobello[1], Manohar N. Murthi[2],
Mads Græsbøll Christensen[3], Søren Holdt Jensen[1],
Marc Moonen[4]

[1]Dept. of Electronic Systems, Aalborg Universitet, Denmark
[2]Dept. of Electrical and Computer Engineering, University of Miami, USA
[3]Dept. of Media Technology, Aalborg Universitet, Denmark
[4]Dept. of Electrical Engineering, Katholieke Universiteit Leuven, Belgium

November 4, 2009

Motivation 1/2

- A cascaded structure of short-term and long-term predictors is used to decorrelate the speech signal leaving a residual that consists (ideally) of Gaussian i.i.d. variables...
- ...but sparse encoding techniques are employed for efficient coding of the residual (e.g., MPE, RPE, ACELP).
- This conceptual difference between a quasi-white LP residual and its approximated version creates a mismatch that can raise the distortion significantly.

Motivation 2/2

- 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!

Proposed solutions

- Define synergistic new predictive framework for speech analysis and coding:
  1. to jointly estimate long-term and short-term predictors.
  2. to find a sparse residual for sparse encoding.
- Redefining the *Analysis-by-Synthesis* (AbS) coding procedure.
  1. The predictor should also be included in the distortion minimization scheme.
  2. Why not finding the predictor also in a closed loop configuration?

Outline

1. Introduction
2. Sparse Linear Prediction
3. Redefining AbS coding
4. Experimental Analysis
5. Conclusions

Fundamentals

- A synergistic new predictive framework that jointly finds a sparse prediction residual **r** as well as a sparse high order linear predictor **a** for a given speech frame **x**:

$$\hat{\mathbf{a}}, \hat{\mathbf{r}} = \arg\min_{\mathbf{a}} \|\mathbf{r}\|_1 + \gamma \|\mathbf{a}\|_1, \quad \text{subject to} \quad \mathbf{r} = \mathbf{x} - \mathbf{X}\mathbf{a}; \qquad (1)$$

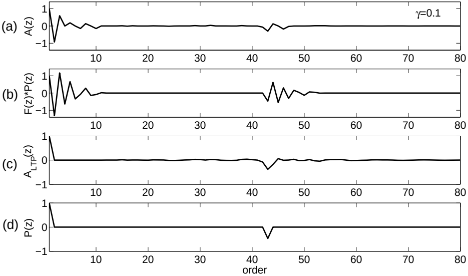
where

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

- $\| \cdot \|_1$ is the 1-norm defined as convex relaxation of the non-convex cardinality measure (the so-called 0-norm).
- The start and end points N_1 and N_2 can be chosen in various ways assuming that $x(n) = 0$ for $n < 1$ and $n > N$.

Joint Estimation of short-term and long-term predictors

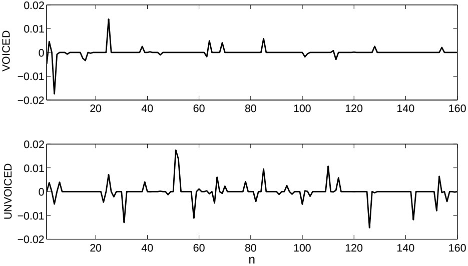
- Keeping the prediction order reasonably high ($K > 100$), we are able to find prediction coefficients the resemble the ones obtained through the cascaded approach.



(**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* $P(z)$*.*

Sparse Residual

- Adapting the residual for sparse encoding



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

Encoding

- frame size $N = 160$, order of the minimization problem $K = 110$.
- High order predictor is factorized into short-term and long-term components $A(z) = F(z)P(z)$ ($N_f = 10$, $N_p = 1$).
- The optimal residual is found in AbS fashion imposing the RPE structure on the residual (20 nonzero samples equally spaced):

$$\tilde{\mathbf{r}} = \arg\min_{\mathbf{r} \in \mathbf{RPE}} \|\mathbf{W}(\mathbf{x} - \tilde{\mathbf{H}}\mathbf{r})\|_2, \tag{2}$$

where $\tilde{\mathbf{H}}$ is the synthesis matrix obtained from the quantized predictor $\tilde{A}(z) = \tilde{F}(z)\tilde{P}(z)$.

Estimation of the impulse response

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

$$\hat{\mathbf{H}} = \arg\min_{\mathbf{H}} \|(\mathbf{x} - \mathbf{H}\tilde{\mathbf{r}})\|_2 \rightarrow \hat{\mathbf{h}} = \arg\min_{\mathbf{h}} \|(\mathbf{x} - \tilde{\mathbf{R}}\mathbf{h})\|_2 \qquad (3)$$

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

$$\|\mathbf{x} - \tilde{\mathbf{R}}\hat{\mathbf{h}}\|_2 = 0. \qquad (4)$$

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

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:

$$\hat{\mathbf{H}}_f, \hat{\mathbf{H}}_p = \arg\min_{\mathbf{H}_f, \mathbf{H}_p} \|(\mathbf{x} - \mathbf{H}_f \mathbf{H}_p \tilde{\mathbf{r}})\|_2. \tag{5}$$

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

$$\hat{\mathbf{h}}_f = \arg\min_{\mathbf{h}_f} \|(\mathbf{x} - (\mathbf{H}_p \tilde{\mathbf{R}})\mathbf{h}_f)\|_2, \tag{6}$$

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

Procedure

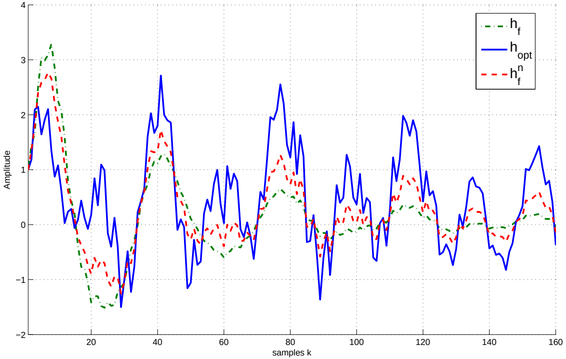
1. Determine $\tilde{A}(z) = F(z)P(z)$ using sparse linear prediction.
2. Calculate $\tilde{\mathbf{r}}$ with RPE encoding.
3. Re-estimate the truncated impulse response $\mathbf{h}_f$.
4. Least-squares IIR approximation of the $\mathbf{h}_f$ using order $N_f = 8, 10, 12$.
5. Optimize the amplitudes of the sparse RPE residual $\tilde{\mathbf{r}}$ using the new synthesis filter $\hat{\mathbf{h}}_f$ (positions and shift stay the same).

Results

METHOD	ΔDIST	ΔMOS
N_f=8	+0.12±0.02 dB	+0.01±0.00
N_f=10	+0.35±0.03 dB	+0.05±0.00
N_f=12	+0.65±0.02 dB	+0.04±0.00
N_f=8 + REST	+0.17±0.01 dB	+0.03±0.00
N_f=10 + REST	+0.41±0.02 dB	+0.06±0.00
N_f=12 + REST	+0.71±0.04 dB	+0.07±0.00

Improvements over conventional SPARSE LP in the decoded speech signal in terms of reduction of log magnitude segmental distortion (ΔDIST) and Mean Opinion Score (ΔMOS) using PESQ evaluation. A 95% confidence intervals is given for each value.

Example



An example of the different impulse response used in the work. The impulse response $\mathbf{h}_f$ *of the original short-term predictor* $F(z)$*, the optimal re-estimated impulse response adapted to the quantized residual* $\mathbf{h}_{opt}$ *and the approximated impulse response* $\mathbf{h}_f^n$ *of the new short-term predictor* $\hat{F}(z)$*. The order is* $N_f = 10$*.*

Conclusions

- We have presented a new method for the enhancing performances in speech coders.
- Sparse linear prediction provides a tighter coupling between the multiple stages of time-domain speech coders.
- Redefining the AbS scheme, the AR modeling can be seen as an IIR approximation of the optimal FIR filter, adapted to the quantized approximated residual, used in the synthesis of the speech segment.
- Improvement in the general performances of the sparse linear prediction framework, but it can be applied also to common methods based on minimum variance linear prediction (e.g. ACELP).