Revisiting the linear prediction analysis-by-synthesis speech coding paradigm using real-time convex optimization

Daniele Giacobello[1*], Manohar Murthi[2],
Tobias Lindstrøm Jensen[3], and Mads Græsbøll Christensen[3]

[1]Sonos Inc., Santa Barbara, CA

[2]University of Miami, Coral Gables, FL

[3]Aalborg University, Aalborg, Denmark

[*]Research carried out at Aalborg University

AALBORG UNIVERSITY
DENMARK

- LPAS at the core of speech coding technology.
- CELP (code-excited linear prediction) probably the most successful embodiment of LPAS:
  - linear prediction (LP) parameters are found in an open-loop configuration,
  - the *excitation* models the prediction residual and is found in a closed-loop configuration,
  - use of perceptually weighted distortion between the original and synthesized speech segment to find the best excitation.
- Since the predictor is quantized *transparently* all the responsibility for the signal approximation falls on the choice of the residual.
- Is the prediction model good enough? Net mismatch in bit allocation between predictor and residual (e.g., AMR-WB 23.85 kbps: 90% vs 10%).

Proposed Solution
Sparse Linear Prediction in LPAS

- Better predictive model $\rightarrow$ more balanced bit allocation
- Use of sparse linear prediction[1] to define a new LPAS framework:
  1. high-order sparse predictor allow for modeling long-term and short-term redundancies;
  2. sparse residual allows for *direct* sparse encoding (no quasi-Gaussian codebook).
- The predictor estimation is included in the distortion minimization of the LPAS scheme.

[1] Daniele Giacobello et al. "Sparse linear prediction and its applications to speech processing". In: *IEEE Trans. Audio, Speech, Lang. Proc.* 20.5 (2012).

Proposed Solution

LPAS as a convex optimization problem

- Weighted minimization of the difference between the original and modeled waveform:

$$\|\mathbf{W}(\mathbf{x} - \hat{\mathbf{x}})\|_2^2.$$

- Sparse constraints applied to a high-order predictor and on the residual used to parametrize the signal:

$$\alpha\|\hat{\mathbf{a}}\|_0 + \beta\|\hat{\mathbf{r}}\|_0 < \delta.$$

- We can write the distortion term as:

$$D(\mathbf{x}, \hat{\mathbf{x}}) = \|\mathbf{W}(\mathbf{x} - \boldsymbol{\Phi}(\hat{\mathbf{a}})\hat{\mathbf{r}})\|_2,$$

  - $\mathbf{H} = \boldsymbol{\Phi}(\mathbf{a})$ is the *synthesis* matrix used in the LPAS equations obtained from the impulse response of $\mathbf{a}$;
  - $\mathbf{W}$ is the matrix that performs the projection in the perceptual domain.
- The distortion is related to the *rate* used for $\hat{\mathbf{a}}$ and $\hat{\mathbf{r}}$:

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

  Considering the cardinality proportional to the rate:

$$R(\hat{\mathbf{x}}) = \alpha\|\hat{\mathbf{a}}\|_0 + \beta\|\hat{\mathbf{r}}\|_0;$$

- $D(\cdot)$, $R(\cdot)$ terms for operational rate-distortion theory[2].

[2] P. Prandoni and M. Vetterli. "R/D optimal linear prediction". In: *IEEE Trans. Speech and Audio Proc.* 8.6 (2000).

Conventional LPAS Formulation

Linear Prediction

- Consider the speech production model where a sample of speech $x(n)$ is a linear combination of P past samples:

$$x(n) = \sum_{p=1}^{P} a_p x(n-p) + r(n),$$

where $\{a_p\}$ are the prediction coefficients (order P) and $r(n)$ is the prediction error.

- The optimization problem to estimate $\{a_p\}$ is

$$\underset{\mathbf{a}}{\text{minimize}} \, \|\mathbf{x} - \mathbf{X}\mathbf{a}\|_q^q + \gamma \|\mathbf{a}\|_k^k,$$

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

- P, N, N_1, N_2, q, k, γ, are chosen according to the problem.

Conventional LPAS Formulation
Codebook Search (1/2)

- The synthesis equations in CELP coders follow the form[3]:

$$\mathbf{x}_k = \mathbf{H}_k \begin{bmatrix} \hat{\mathbf{r}}_{k-1} \\ \mathbf{r}_k \end{bmatrix} = \begin{bmatrix} \mathbf{H}_k^{\mathrm{U}} & \mathbf{H}_k^{\mathrm{L}} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{r}}_{k-1} \\ \mathbf{r}_k \end{bmatrix},$$

where $\mathbf{H}_k$ is the $N \times 2N$ convolution matrix obtained with the truncated impulse response of $\mathbf{a}$. Thus:

$$\mathbf{x}_k = \mathbf{H}_k^{\mathrm{L}} \mathbf{r}_k + \mathbf{H}_k^{\mathrm{U}} \hat{\mathbf{r}}_{k-1},$$

where the term $\mathbf{H}_k^{U} \hat{\mathbf{r}}_{k-1}$ is the *zero input response*. Subtracting it from the signal to quantize $\mathbf{x}_k$, we obtain the *target signal*:

$$\tilde{\mathbf{x}}_k = \mathbf{H}_k^{\mathrm{L}} \mathbf{r}_k.$$

[3] W. B. Kleijn, D. J. Krasinski, and R. H. Ketchum. "Fast methods for the CELP speech coding algorithm". In: *IEEE Trans Acoustics, Speech, Sig. Proc.* 38.8 (1990).

- The *target signal* is reconstructed by adding two excitation vectors:

$$\|\mathbf{W}\left(\tilde{\mathbf{x}}_k - \mathbf{H}_k^{\mathrm{L}}\left(g_{\mathrm{a}}\mathbf{c}_k^{\mathrm{a}} + g_{\mathrm{f}}\mathbf{c}_k^{\mathrm{f}}\right)\right)\|_2^2,$$

  where
  - $g_{\mathrm{a}}\mathbf{c}_k^{\mathrm{a}}$ contribution from the *adaptive* codebook,
  - $g_{\mathrm{f}}\mathbf{c}_k^{\mathrm{f}}$ contribution from the *fixed* codebook,
  - $\mathbf{W}$ is the perceptual weighting matrix.
- Combinatorial problem generally solved one variable at the time.

Alternative Formulation

New synthesis equations for LPAS (1/2)

- If we consider the *conventional IIR formulation*[4] for the LPAS synthesis equations:

$$\mathbf{x}_k = \mathbf{H}_k \begin{bmatrix} \hat{\mathbf{r}}_{k-1} \\ \mathbf{r}_k \end{bmatrix} \rightarrow \mathbf{r}_k = \mathbf{A}_k \begin{bmatrix} \hat{\mathbf{x}}_{k-1} \\ \mathbf{x}_k \end{bmatrix},$$

$$\mathbf{r}_k = \begin{bmatrix} \mathbf{A}_k^{\mathrm{U}} & \mathbf{A}_k^{\mathrm{L}} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{x}}_{k-1} \\ \mathbf{x}_k \end{bmatrix} = \mathbf{A}_k^{\mathrm{L}} \mathbf{x}_k + \mathbf{A}_k^{\mathrm{U}} \hat{\mathbf{x}}_{k-1}.$$

- For high-order filters with $P = N - 1$:

$$\mathbf{A}_k^{\mathrm{L}} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ a_1 & 1 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ a_P & a_{P-1} & \cdots & \cdots & a_1 & 1 \end{bmatrix}, \quad \mathbf{A}_k^{\mathrm{U}} = \begin{bmatrix} 0 & a_P & \cdots & \cdots & a_2 & a_1 \\ & & & & & \vdots \\ 0 & 0 & a_P & \cdots & \cdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 & a_P \\ 0 & 0 & \cdots & \cdots & 0 & 0 \end{bmatrix}.$$

[4] T. Bäckström. “Comparison of windowing in speech and audio coding”. In: *IEEE WASPAA*. 2013.

- Given the structure of $\mathbf{A}_k$, we can rewrite:

$$\mathbf{r}_k = \mathbf{A}_k \begin{bmatrix} \hat{\mathbf{x}}_{k-1} \\ \mathbf{x}_k \end{bmatrix} \quad \rightarrow \quad \bar{\mathbf{X}}_k \mathbf{a}_k = \mathbf{r}_k,$$

where:

$$\bar{\mathbf{X}}_k = [\check{\mathbf{x}}_k | \check{\mathbf{X}}_k] = \left[\begin{array}{c|ccccc} x_{k,0} & \hat{x}_{k-1,N} & \cdots & & \cdots & \hat{x}_{k-1,1} \\ x_{k,1} & \vdots & \ddots & & \ddots & \hat{x}_{k-1,2} \\ & \ddots & \ddots & & \ddots & \vdots \\ \vdots & \ddots & \ddots & & \ddots & \vdots \\ \vdots & \vdots & \ddots & & x_{k,0} & \hat{x}_{k-1,N} \\ x_{k,N} & x_{k,N-1} & \cdots & & x_{k,1} & x_{k,0} \end{array}\right],$$

and $\mathbf{a}_k = [1, a_1, \ldots, a_P]^T$. We can see that $\check{\mathbf{x}}_k = \mathbf{x}_k$.

- The optimization problem becomes (omitting k):

$$
\begin{aligned}
\text{minimize}_{\mathbf{a},\mathbf{r}} \quad & \|\mathbf{r} - \check{\mathbf{x}} - \check{\mathbf{X}}\mathbf{a}\|_2^2 \\
\text{subject to} \quad & \|\mathbf{a}\|_1 \leq \delta \\
& \|\mathbf{r}\|_1 \leq \gamma
\end{aligned}
$$

- 1-norm chosen as a convex relaxation of the 0-norm.
- Easy to control tradeoff between the quality of the reconstruction ($\approx$ distortion) and the sparsity of the representation ($\approx$ rate).
- One step estimation for $\mathbf{a}_k = [a_1, \ldots, a_P]^T$ and $\mathbf{r}_k = [r_1, \ldots, r_N]^T$ $(P = N - 1)$.

- The gradient is given by

$$\nabla f(\mathbf{x}) = \nabla f(\mathbf{a},\mathbf{r}) = \begin{bmatrix} \check{\mathbf{X}} & -I \end{bmatrix}^T \left(\begin{bmatrix} \check{\mathbf{X}} & -I \end{bmatrix} \begin{bmatrix} \mathbf{a} \\ \mathbf{r} \end{bmatrix} + \check{\mathbf{x}} \right).$$

- We solve the SpLPAS problem using a simple variant of the fast gradient projection method[5]:
  - the objective is smooth (quadratic),
  - the gradient and the projection onto the set $\{x \mid \|x\|_1 \leq \rho\}$ can be calculated efficiently,
  - use of a fixed step-size ensures convergence and avoids an iterative line-search algorithm.
- For a problem with $N = 160$ and $P = 159$, the primal problem by an interior-point method (CVX+SeDuMi) takes, on average, 934.0 ms vs 1.7 ms of the fast method.

[5] Yuri Nesterov. *Introductory Lectures on Convex Optimization, A Basic Course*. Kluwer Academic Publishers, 2004.

- We analyzed one hour of clean speech extracted from the TIMIT database:
  - different gender, age, pitch, regional accent;
  - normalized at -26 dBFS;
  - resampled to $f_s = 8$kHz;
  - frame size $N = 160$ (20 ms).
- Order $P = 159$ (cover pitch periods with $[70\text{ Hz}, 500\text{ Hz}]$).
- Values of δ and γ in the table chosen from 50% of the data.

- We define three versions of the SpLPAS algorithm

$$
\begin{aligned}
\text{minimize}_{\mathbf{a},\mathbf{r}} \quad & \|\mathbf{r} - \check{\mathbf{x}} - \check{\mathbf{X}}\mathbf{a}\|_2^2 \\
\text{subject to} \quad & \|\mathbf{a}\|_1 \leq \delta \\
& \|\mathbf{r}\|_1 \leq \gamma
\end{aligned}
$$

with

METHOD	δ	γ
$\text{SpLPAS}_{\text{v1}}$	2	.30
$\text{SpLPAS}_{\text{v2}}$	1.9	.27
$\text{SpLPAS}_{\text{v3}}$	1.7	.21

- No close form bounds for δ and γ, values are found empirically and are related to the desired sparsity.
- If $\delta = 0$ and $\gamma = 0$, trivial solution $\mathbf{a} = \mathbf{0}$ thus $\mathbf{r} = \check{\mathbf{x}} = \mathbf{x}$.

- Once we obtained $\hat{\mathbf{r}}$ and $\hat{\mathbf{a}}$, we quantize them losslessly using a simple variable-rate coding/decoding structure[6].
  - **Lossy compression is controlled uniquely by defining δ and γ;**
  - mix of parametric and nonparametric modeling for quantizing $\mathbf{a}$ and $\mathbf{r}$;
  - binary mask (location of coefficients) coded as a memoryless random process with $\log_2 \binom{N-1}{\|\mathbf{a}\|_0}$ and $\log_2 \binom{N}{\|\mathbf{r}\|_0}$;
  - the encoding/decoding uses $32\times 32 \rightarrow 64$ bit operations;
  - use of the same encoding/decoding scheme for all configurations.
- Average of 9 bits per r and 14 bits per a.
- Stability of the predictor not necessary!

[6] F. Ghido and I. Tabus. “Sparse modeling for lossless audio compression”. In: *IEEE Trans. Audio, Speech, Lang. Proc.* 21.1 (2013).

Experimental Evaluation

Example: SpLPAS$_{v1}$ for Voiced Speech

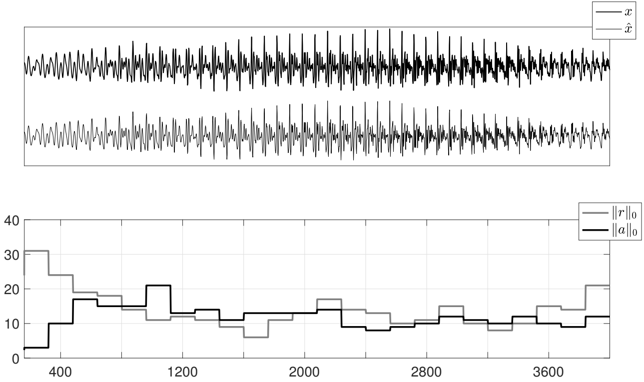

Figure: Top pane shows the original and reconstructed speech signal. Bottom pane shows the cardinality of the estimated $\mathbf{r}$ and $\mathbf{a}$.

Experimental Evaluation

Example: $\text{SpLPAS}_{\text{v1}}$ for Unvoiced Speech

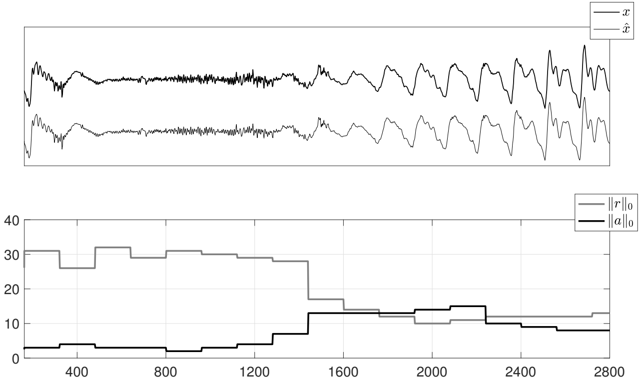

Figure: Top pane shows the original and reconstructed speech signal. Bottom pane shows the cardinality of the estimated **r** and **a**.

We compare SpLPAS with the G.711 waveform coder (μ-law PCM), and the low-complexity CELP[7]. PESQ was used for MOS scores. The rate is expressed in kbps.

METHOD	$\|\mathbf{a}\|_0$	$\|\mathbf{r}\|_0$	$R_{\mathbf{a}}$	$R_{\mathbf{r}}$	R_{tot}	MOS
G711	N/A				64	4.22
CELP	10+1	26	1.8+1.2	12	16	3.97
SpLPAS$_{v1}$	15.4	17.3	18.1	15.3	38.8	4.07
SpLPAS$_{v2}$	13.1	14.0	15.6	12.6	32.9	3.84
SpLPAS$_{v3}$	10.2	11.5	12.3	10.1	23.2	3.29

Code for SpLPAS available at `https://github.com/giacobello/SpLPAS`

[7] Juin-Hwey Chen. "Toll-quality 16 kb/s CELP speech coding with very low complexity". In: *IEEE ICASSP*. 1995.

Conclusion

- New formulation for LPAS allows for
  - one-step estimation of predictor and residual,
  - possibility of choosing the right sparsity and distortion level for each speech frame,
  - better tradeoffs between $\mathbf{a}$ and $\mathbf{r}$ (R/D interpretation).
- Encoding/decoding scheme should be better tailored for the problem at hand (e.g., the predictor can be factorized in short-term and long-term components).
- Defining a proper $\mathbf{W}$ can help reducing the bit rate dramatically!
- High-order predictor are promising for audio as well
  - possibility of using our scheme for joint speech-audio coding (current approaches switch between MDCT-based and ACELP-based depending on the content[8]).

[8] Max Neuendorf et al. "MPEG Unified Speech and Audio Coding - The ISO/MPEG Standard for High-Efficiency Audio Coding of All Content Types". In: *Audio Engineering Society Convention 132*. 2012.