Differentiable Time-Varying Linear Prediction in the Context of End-to-End Analysis-by-Synthesis

Chin-Yun Yu and György Fazekas
Centre for Digital Music, Queen Mary University of London

Motivations

The popular frame-based approximation of time-varying linear prediction (LP) filter is fast for end-to-end training, but it introduces

- mismatch between training and real-time inference condition
- high resonance filters due to windowing
- incontinuous filter representations between adjacent frames due to overlap-add

Time-Varying Linear Prediction

$$s(t) = \mathrm{LP}_{\tilde{\mathbf{a}}(t)}(e(t)) = e(t) - \sum_{i=1}^{M} \tilde{a}_i(t) s(t-i)$$

Output
Input
Time-varying coefficients
M = filter order
t = time index
i = coefficient index

Problem Definition

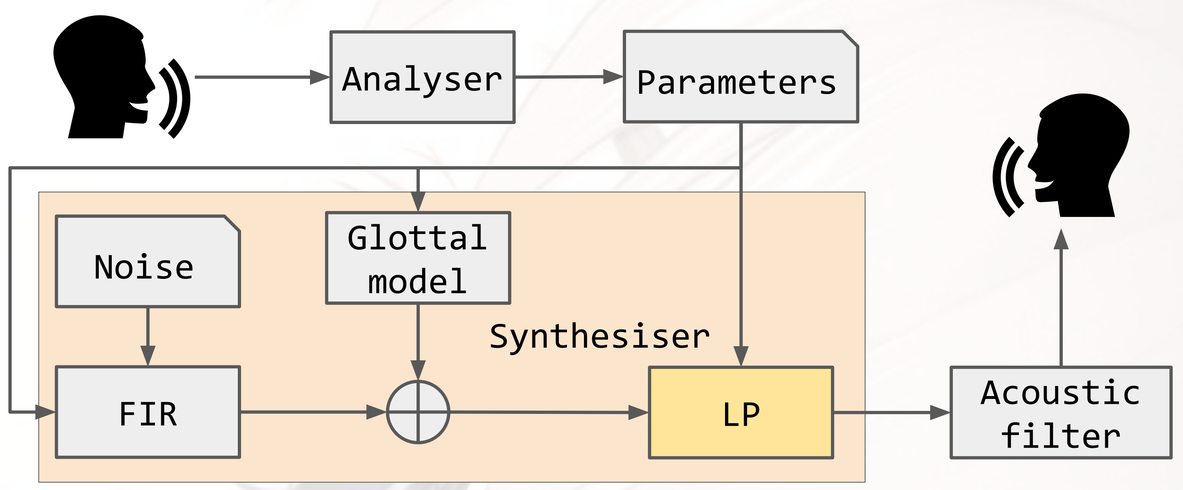
Find an efficient and accurate way to get

- Gradients w.r.t input e(t)
- Gradients w.r.t coefficients \tilde{**a**}(t)

Contributions

- An efficient gradient backpropagation algorithm for time-domain training of LP filter
- Our differentiable LP filter written in PyTorch is open source and available on PyPI
- Improving previous GOLF vocoder[1] using source-filter formulation and differentiable sample-wise LP for exact time-varying modelling

End-to-End Synthesis Experiment



Efficient Gradient Backpropagation Algorithm

`pip install torchlpc`

LP as time-varying IIRs

$$\mathrm{LP}_{\tilde{\mathbf{a}}(t)}(e(t)) = e(t) + \sum_{d=1}^{t} b_d(t) e(t-d)$$

IIR coefficients

$$b_d(t) = \sum_{\mathbf{q}\in\mathcal{G}_d} (-1)^{|\mathbf{q}|} \prod_{j=1}^{|\mathbf{q}|} \tilde{a}_{q_j}\left(t - \sum_{k=1}^{j} Q(\mathbf{q})_k\right)$$

$$\mathcal{G}_d = \bigcup_{i=1}^{\min(d,M)} \{[i;\mathbf{q}] : \mathbf{q} \in \mathcal{G}_{d-i}\}$$

- Backpropagation = LP in backward + multiplication
- The backpropagation is expressed as LP so we can reuse our efficient Numba implementation

Backpropagation with IIRs

$$\frac{\partial \mathcal{L}}{\partial e(t)} = \frac{\partial \mathcal{L}}{\partial s(t)}\frac{\partial s(t)}{\partial e(t)} + \sum_{d=1}^{T-t} \frac{\partial \mathcal{L}}{\partial s(t+d)}\frac{\partial s(t+d)}{\partial e(t)} = \frac{\partial \mathcal{L}}{\partial s(t)} + \sum_{d=1}^{T-t} b_d(t+d)\frac{\partial \mathcal{L}}{\partial s(t+d)},$$

-/+ → causal/non-causal

$$b_d(t+d) = \sum_{\mathbf{q}\in\mathcal{G}_d} (-1)^{|\mathbf{q}|} \prod_{j=1}^{|\mathbf{q}|} \hat{a}_{q_j}\left(t + \sum_{k=1}^{j} Q(\mathbf{q})_k\right)$$

$$\hat{a}_i(t) = \tilde{a}_i(t+i)$$

T-t: Time reversal indexing

1 Backpropagation with LP

$$\frac{\partial \mathcal{L}}{\partial e(T-t)} = \mathrm{LP}_{\hat{\mathbf{a}}(T-t)}\left(\frac{\partial \mathcal{L}}{\partial s(T-t)}\right)$$

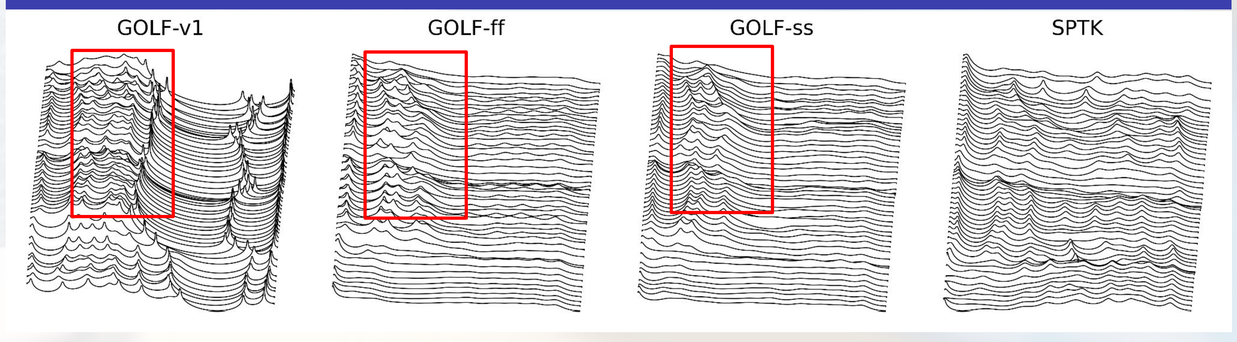
2 Gradients w.r.t coefficients

$$z_i(t) = -\tilde{a}_i(t) s(t-i)$$

$$\rightarrow \frac{\partial \mathcal{L}}{\partial e(t)} = \frac{\partial \mathcal{L}}{\partial z_i(t)}$$

$$\frac{\partial \mathcal{L}}{\partial \tilde{a}_i(t)} = \frac{\partial \mathcal{L}}{\partial z_i(t)}\frac{\partial z_i(t)}{\partial \tilde{a}_i(t)} = -\frac{\partial \mathcal{L}}{\partial e(t)} s(t-i)$$

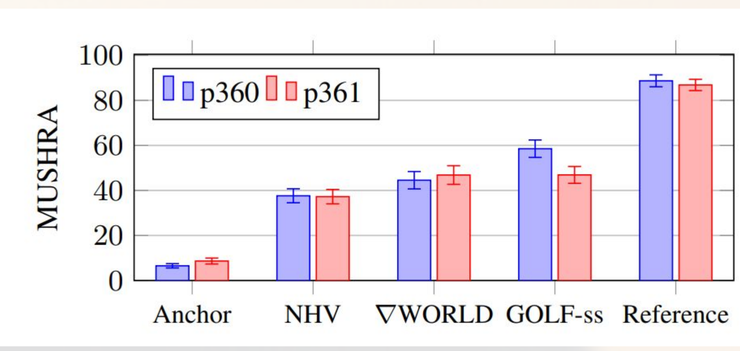
Results



Form.	Model	MSS⇩	MCD⇩	PESQ⇧	FAD⇩
HpN	DDSP	2.965	3.42	2.42	32.7±7.7
	NHV	**2.914**	3.32	2.58	31.8±7.4
	GOLF-v1	3.026	3.54	2.36	39.6±9.4
SF	WORLD	3.515	6.07	1.77	270.6±56.1
	MLSA	3.006	3.35	2.48	40.1±10.0
	∇WORLD	2.918	**3.26**	**2.66**	**22.4±5.6**
	GOLF-ss	3.005	3.43	2.49	38.4±9.2
	GOLF-ff	3.011	3.46	2.39	34.0±7.7
	GOLF-fs	3.074	3.70	2.16	44.1±10.1

- Dataset: VCTK 0.92
- **v1**: the original GOLF, **ff**: frame-wise LP, **ss**: sample-wise LP (proposed), **HpN**: harmonic-plus-noise, **SF**: source-filter
- **Training with sample-wise LP performs the best among all GOLF variants**

- GOLF-ss outperforms other baselines on the test speaker p360 (male)
- The overall scores of speaker p361 are lower due to poor performance of the Dio pitch estimator on that speaker



Conclusions and Future Works

- Source-filter form helps learning more reasonable filter response
- The proposed differentiable LP not only outperform the frame-wise method, its learnt filter representation is also the smoothest, which is a desired characteristic
- Exploring forward-mode automatic differentiation, second-order gradients, more analysis on the learnt representations, etc.

The first author is a research student at the UKRI CDT in AI and Music, supported jointly by UK Research and Innovation [grant number EP/S022694/1] and Queen Mary University of London.

Reference

[1] Yu, Chin-Yun, and György Fazekas. "SINGING VOICE SYNTHESIS USING DIFFERENTIABLE LPC AND GLOTTAL-FLOW-INSPIRED WAVETABLES." in Proceedings of ISMIR, 2023.
[2] Yu, Chin-Yun, and György Fazekas. "Differentiable Time-Varying Linear Prediction in the Context of End-to-End Analysis-by-Synthesis." in Proceedings of INTERSPEECH, 2024.

torchlpc

Audio Samples

Code

Contact
Email: chin-yun.yu@qmul.ac.uk
GitHub/X: @yoyololicon