Singing Voice Synthesis Using Differentiable LPC and Glottal-Flow-Inspired Wavetables

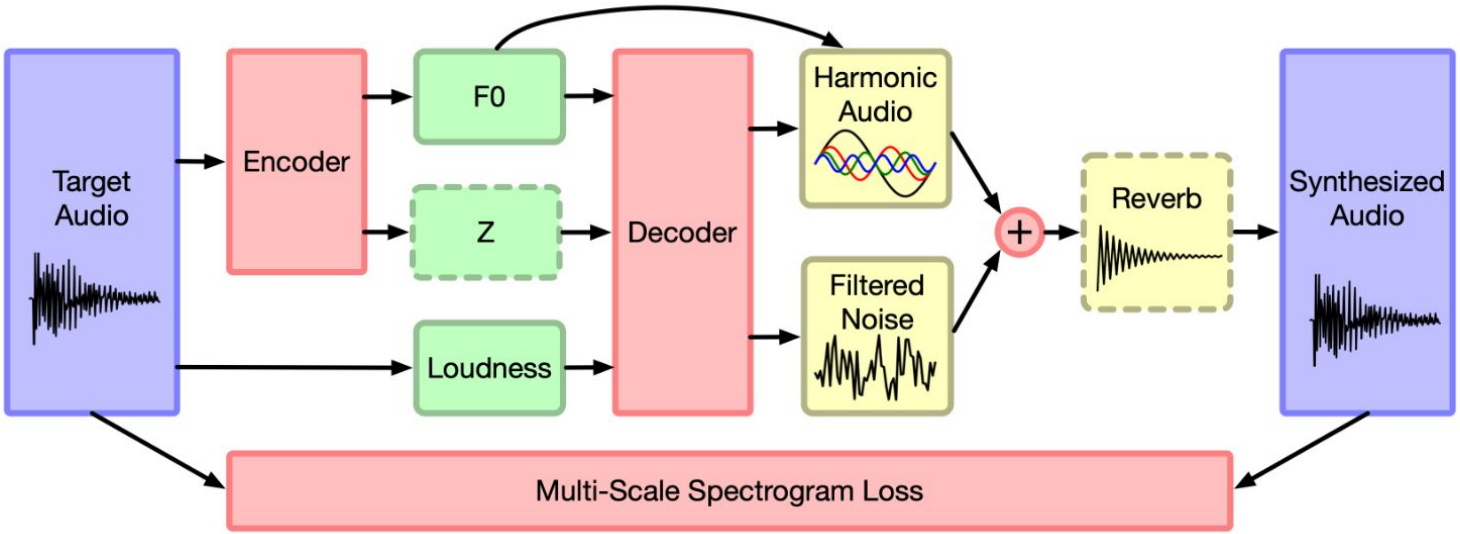
Chin-Yun Yu and György Fazekas
Centre for Digital Music, Queen Mary University of London

ISMIR 2023

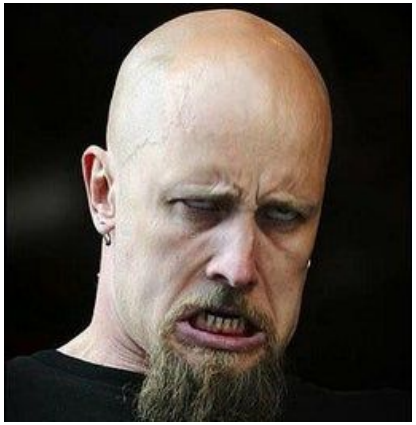
AI + MUSIC

Queen Mary University of London

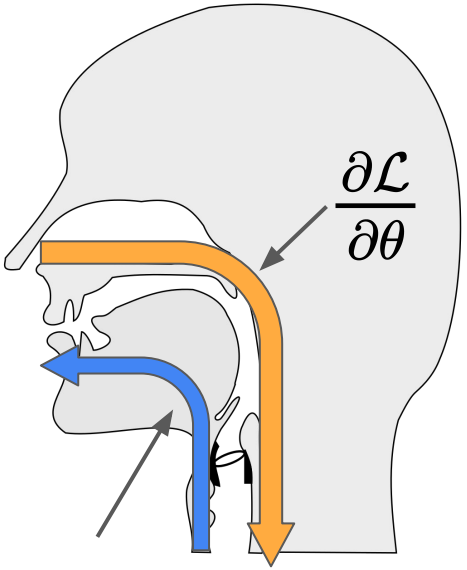
UK Research and Innovation



Engel, Jesse, Chenjie Gu, and Adam Roberts. "DDSP: Differentiable Digital Signal Processing." International Conference on Learning Representations. 2019.



$\mathcal{L}$



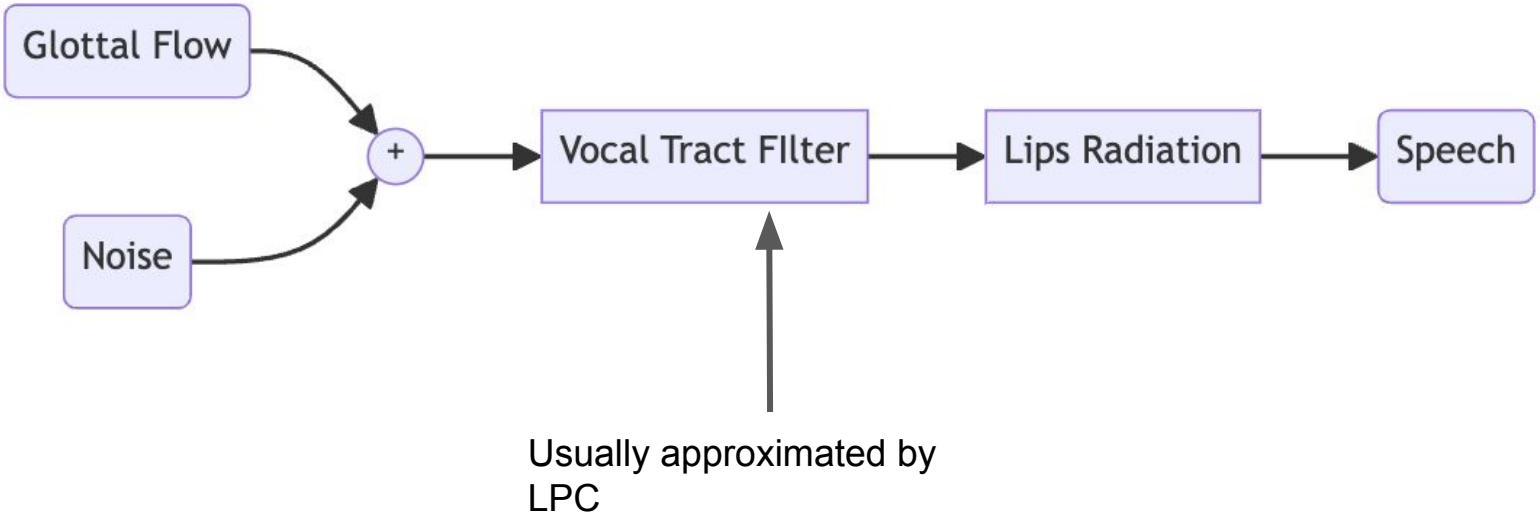
$\boldsymbol{\theta}$:= ?

We propose...

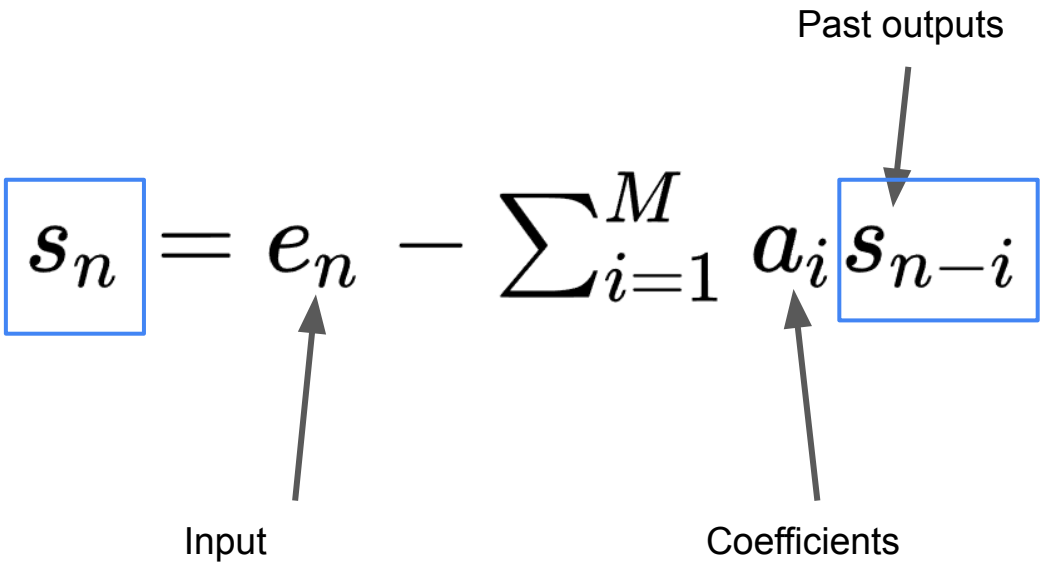
GlOttal-flow LPC Filter (GOLF)

1. Glottal flows as the source signal
2. Differentiable, efficient LPC synthesis simulating the vocal tract

Glottal Source-Filter Model (simplified)

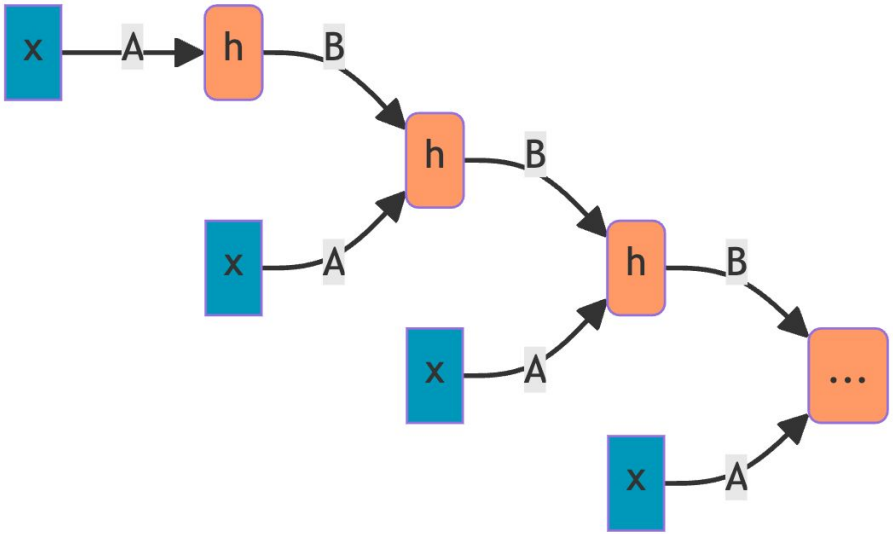


Linear Predictive Coding (LPC)



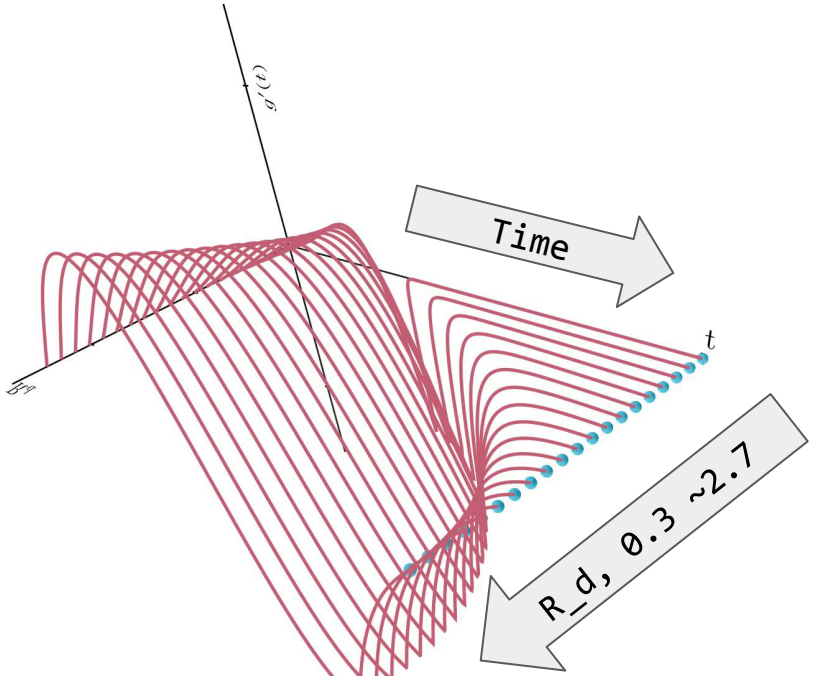
Recursion is slow...

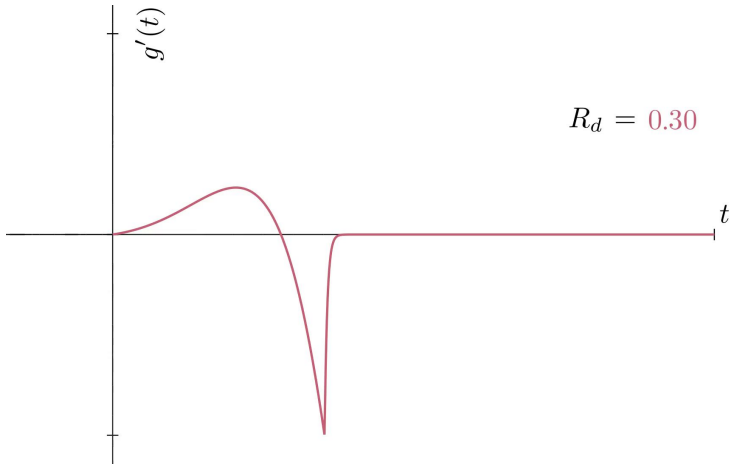
- Just like RNN
- Long sequence == very deep computational graph

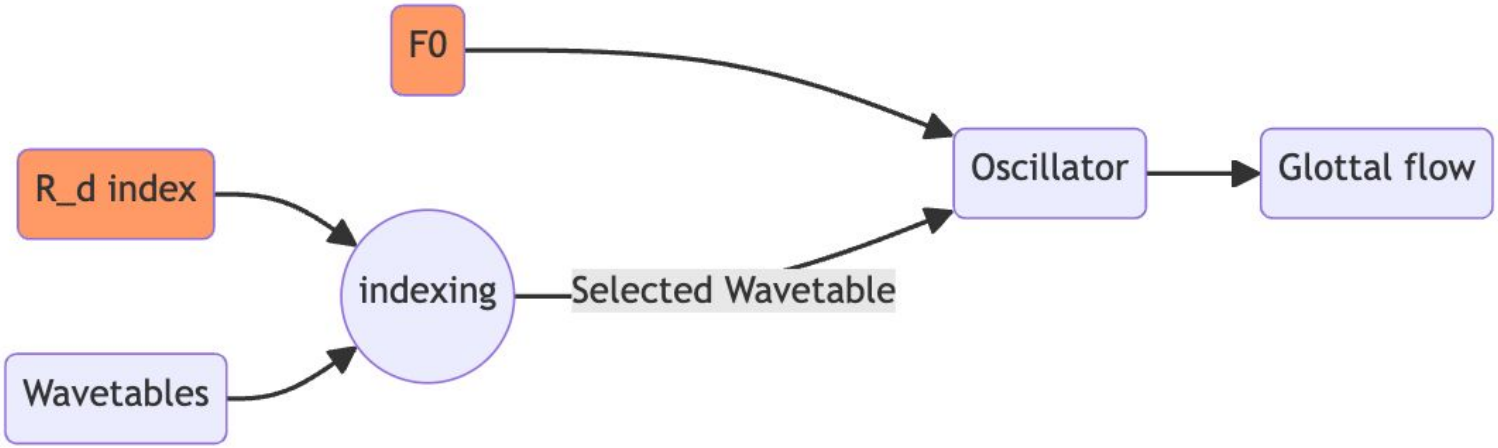


Proposed Methodology

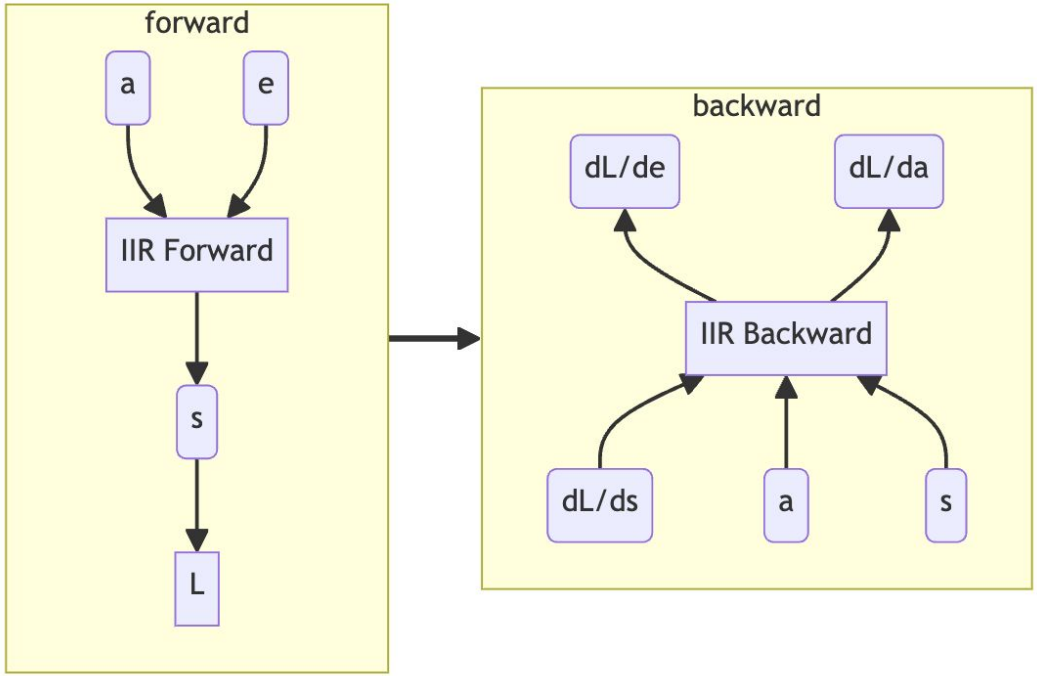
Glottal-flow Wavetables

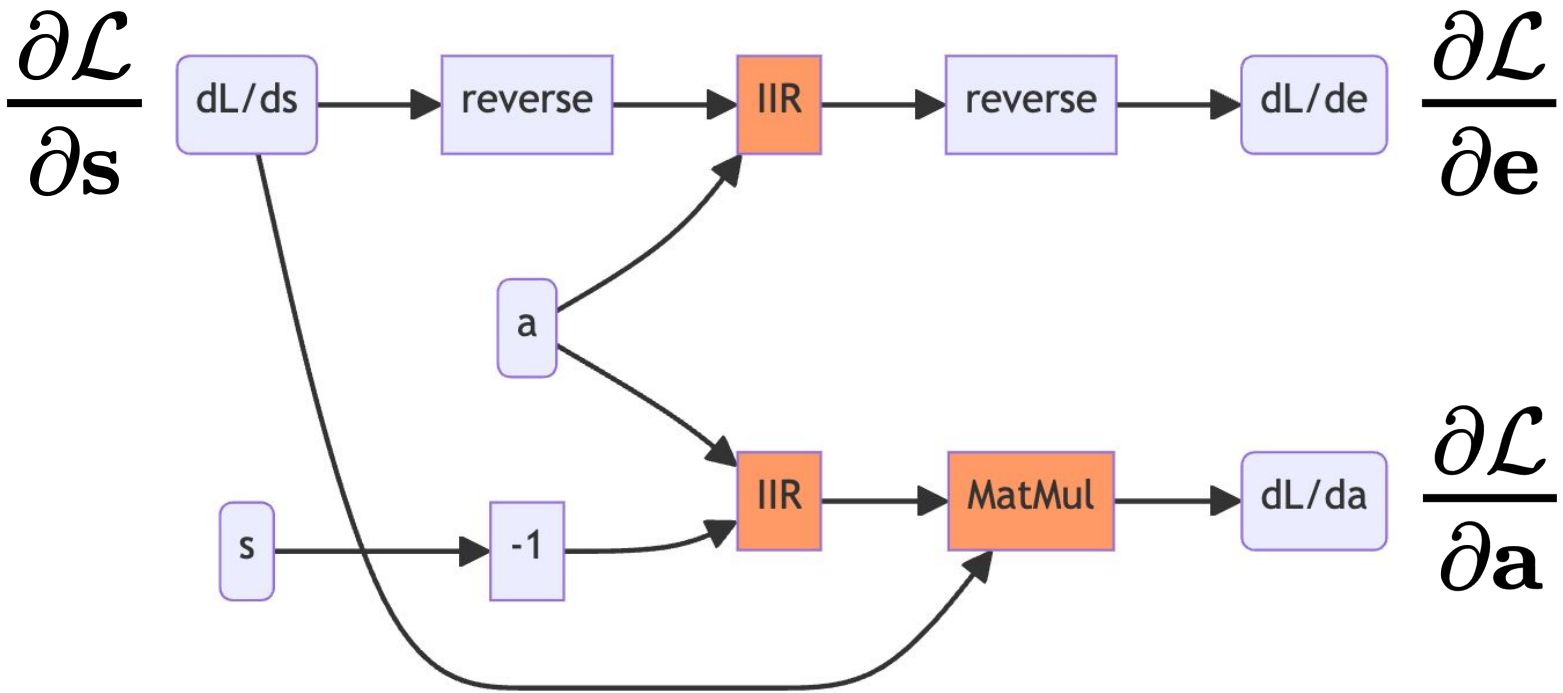





F0
R_d index
Oscillator
Glottal flow
indexing
Selected Wavetable
Wavetables

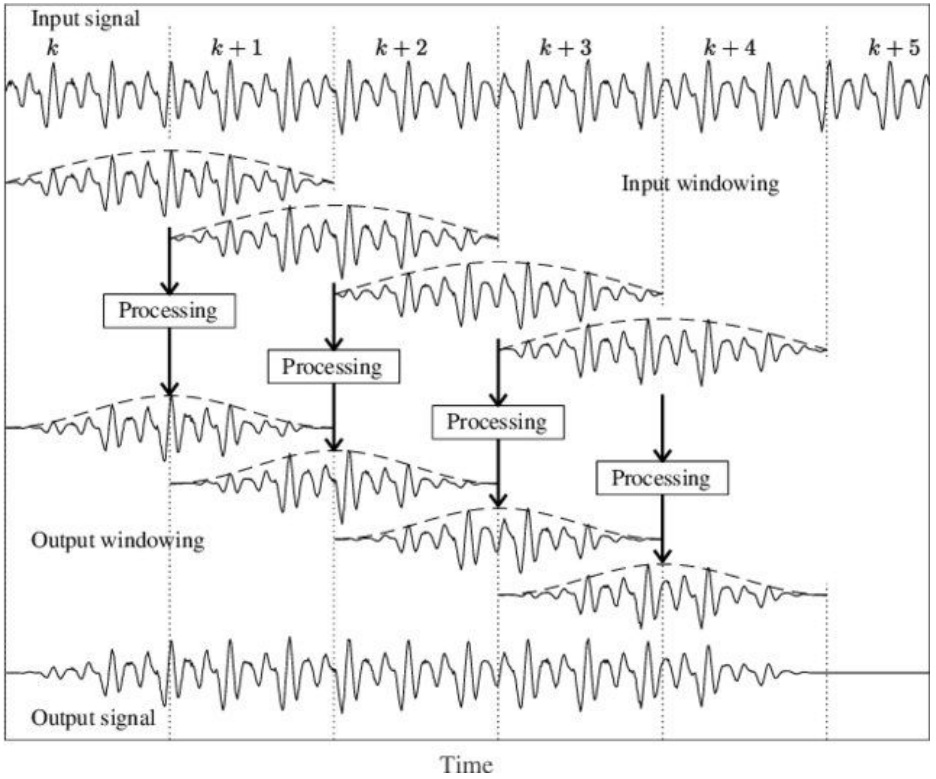
Efficient and Differentiable LPC




$\frac{\partial \mathcal{L}}{\partial \mathbf{s}}$
dL/ds
reverse
IIR
reverse
dL/de
$\frac{\partial \mathcal{L}}{\partial \mathbf{e}}$
a
IIR
MatMul
dL/da
$\frac{\partial \mathcal{L}}{\partial \mathbf{a}}$
s
-1

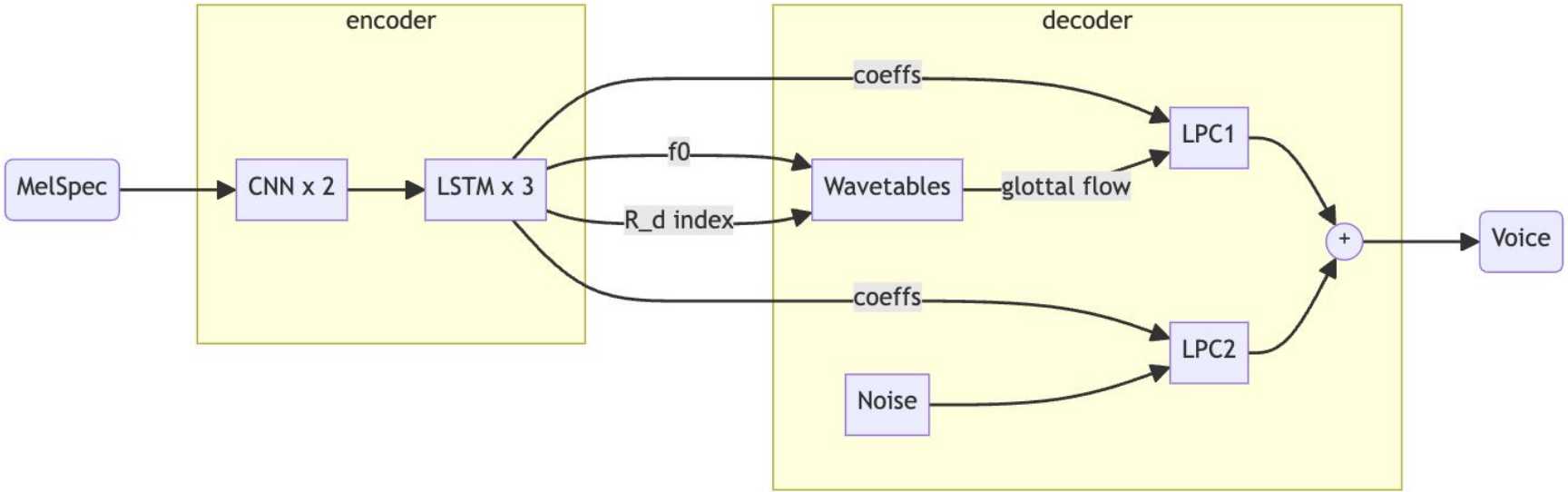
Simulating Time-Varying LPC

- Parallelisable



Bäckström, Tom. "Overlap-add windows with maximum energy concentration for speech and audio processing." ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019.

GOLF Vocoder



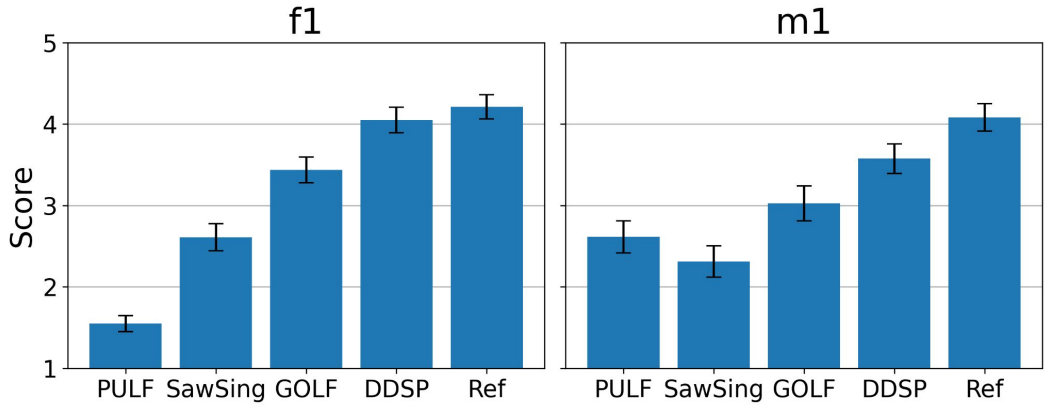
The Vocoder Experiment

- Dataset: MPop600[1]
  - f1/m1 singers
  - exclude 3 songs for evaluation
- Input features: 80 mel-frequencies
- Criterion
  - multi-resolution STFT
  - F0
  - voiced/unvoiced
- 800k training steps

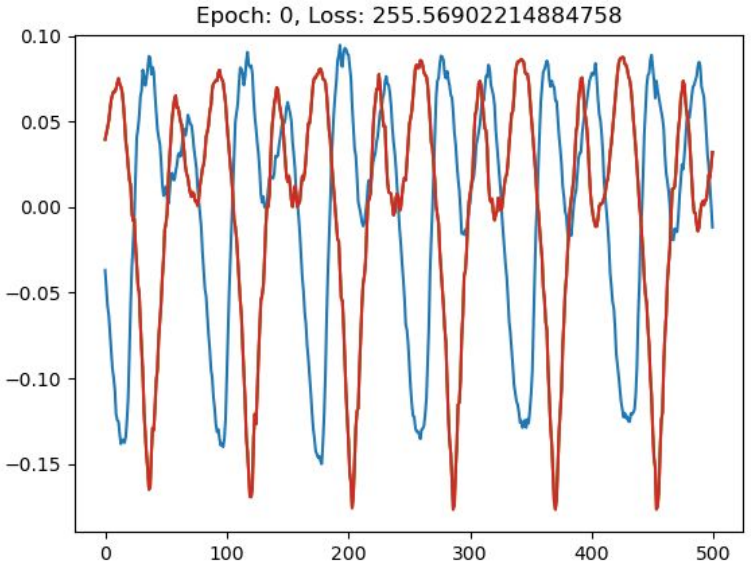
- Decoder
  - DDSP
  - SawSing[2]
  - GOLF
  - PULF (replace glottal flow with pulse train)
- Metrics
  - multi-resolution STFT
  - Mean Absolute Error on F0
  - FAD
  - MOS on singing quality

[1] Chu, Chan-Chuan, et al. "MPop600: A Mandarin popular song database with aligned audio, lyrics, and musical scores for singing voice synthesis." 2020 APSIPA ASC. IEEE, 2020.
[2] Wu, Da-Yi, et al. "DDSP-based singing vocoders: A new subtractive-based synthesizer and a comprehensive evaluation." 2022 Proc. International Society for Music Information Retrieval.

Singers	Models	MSSTFT	MAE-f0 (cent)	FAD
f1	DDSP	**3.09**	**74.47**±1.19	0.50±0.02
	SawSing	3.12	78.91±1.18	**0.38**±0.02
	GOLF	3.21	77.06±0.88	0.62±0.02
	PULF	3.27	76.90±1.11	0.75±0.04
m1	DDSP	**3.12**	**52.95**±1.03	0.57±0.02
	SawSing	3.13	56.46±1.04	**0.48**±0.02
	GOLF	3.26	54.09±0.30	0.67±0.01
	PULF	3.35	54.60±0.73	1.11±0.04



Models	Memory	RTF		Waveform L2	
		GPU	CPU	Min	Max
DDSP	7.3	0.015	0.237	71.83	88.77
SawSing	7.3	0.015	0.240	75.72	93.16
GOLF	**2.6**	**0.009**	**0.023**	**21.98**	**64.82**
PULF	7.5	0.015	0.248	44.08	70.59



In Summary, GOLF...

...is more computationally efficient.

...has better ability to model human voice phase response.

Differentiable IIR: **torchaudio.functional.lfilter**

Source Code: https://github.com/yoyololicon/golf/

Audio Samples: https://yoyololicon.github.io/golf-demo/