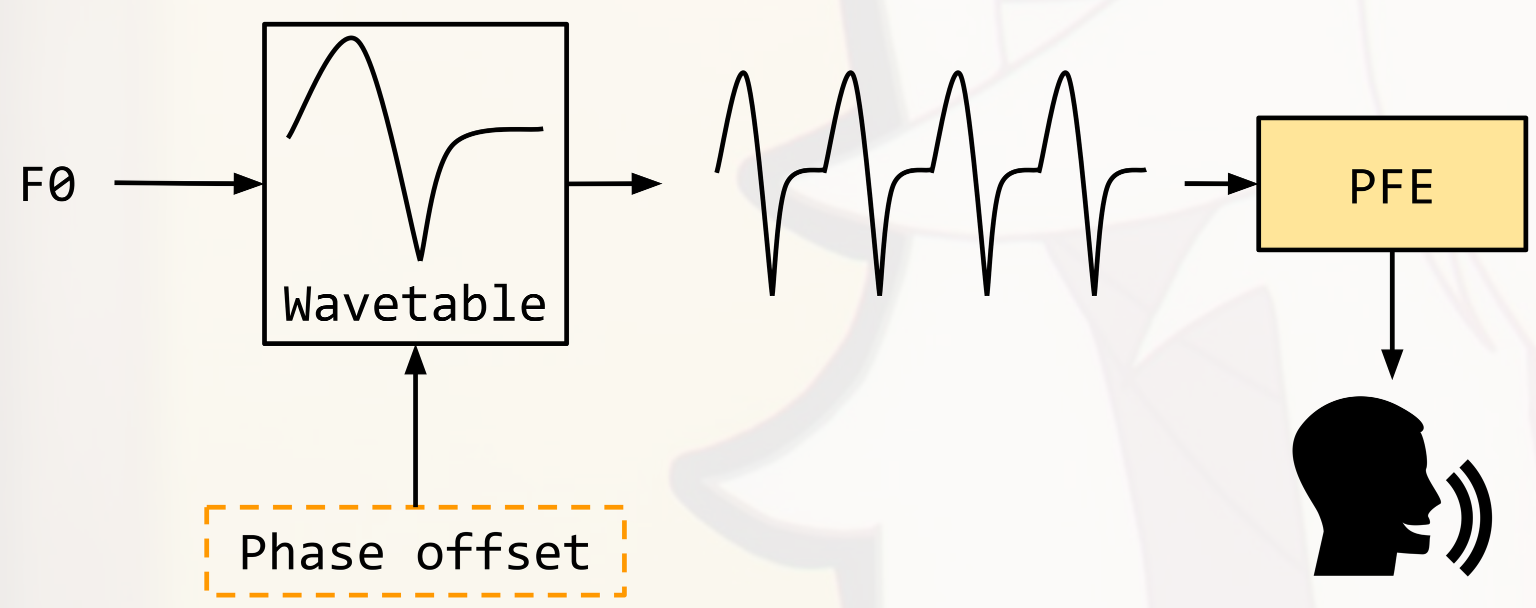
Phase-Accurate End-to-End Analysis-by-Synthesis for Singing Voice

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

Motivations

Neural vocoders are commonly trained using spectral and adversarial losses, which does not consider the **phase information** in voice. Phase is critical for editing and manipulating voices. In addition, **end-to-end learning** enables joint modelling for all the system components at once, ensuring the system behave the same regardless during training or evaluation. Learning phase-accurate synthesis in an end-to-end way enables more robust system.

The Synthesiser (Modified from GOLF [1])



Additive Wavetable

Linear interpolation on fixed-size wavetables introduces aliasing. Solution: **Band-limited additive synthesis.**

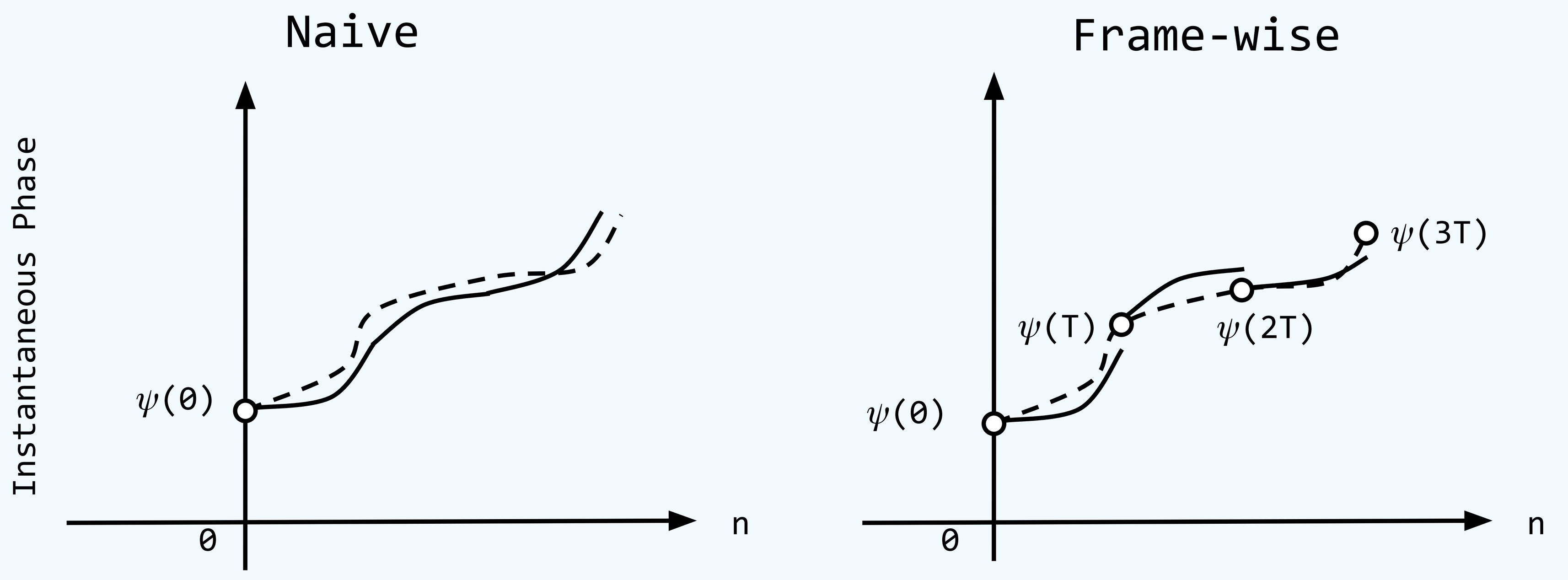
$$x(n) = a_1 \cos(\psi(n)) + \sum_{k=2}^{\left\lfloor \frac{\pi}{f_0(n)} \right\rfloor} a_k \cos(k\psi(n) + \phi_k)$$

- Control parameters
  - $\psi(n)$: Instantaneous phase
- Wavetable parameters
  - $\{a_1, a_2, ..., a_K\}$: Amplitudes
  - $\{\phi_2, \phi_3, ..., \phi_K\}$: Relative phase differences

Instantaneous Phase Calculation

$$\psi(n) = \psi(0) + \sum_{m=1}^{n} f_0(n)$$

- $\psi(0)$: Initial phase, $f_0(n)$: Instantaneous frequency
- Issue: F0 estimation errors accumulate
- Solution: Inspired by quasi-harmonic model [2], predict sub-sampled $\psi(mT)$ where $m \in \{0,1,2,...\}$.
  - Errors are limited within T samples



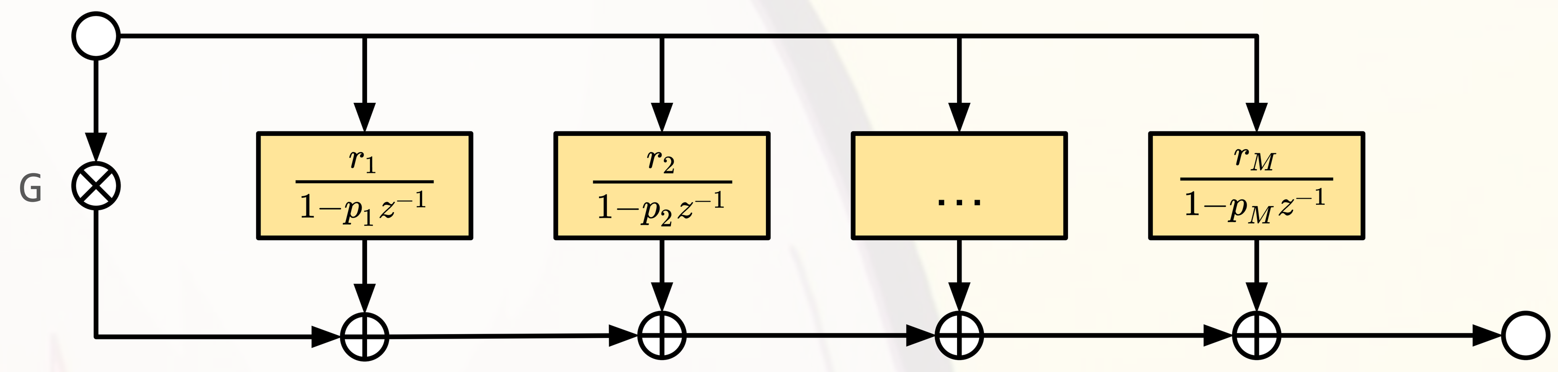
➢ Solid lines: Predicted $\psi(n)$
➢ Dashed lines: Ground truth $\psi(n)$

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.

Partial Fraction Expansion (PFE) Filter

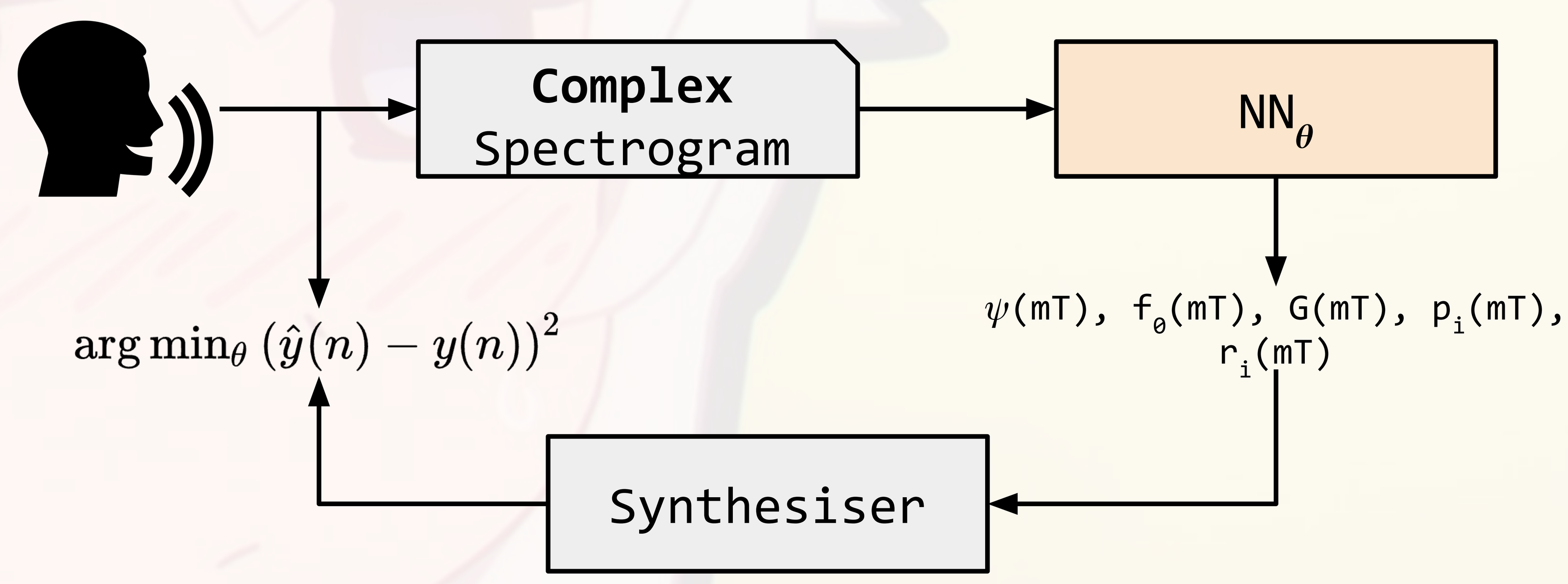
Advantages over linear prediction (all-pole) filter:

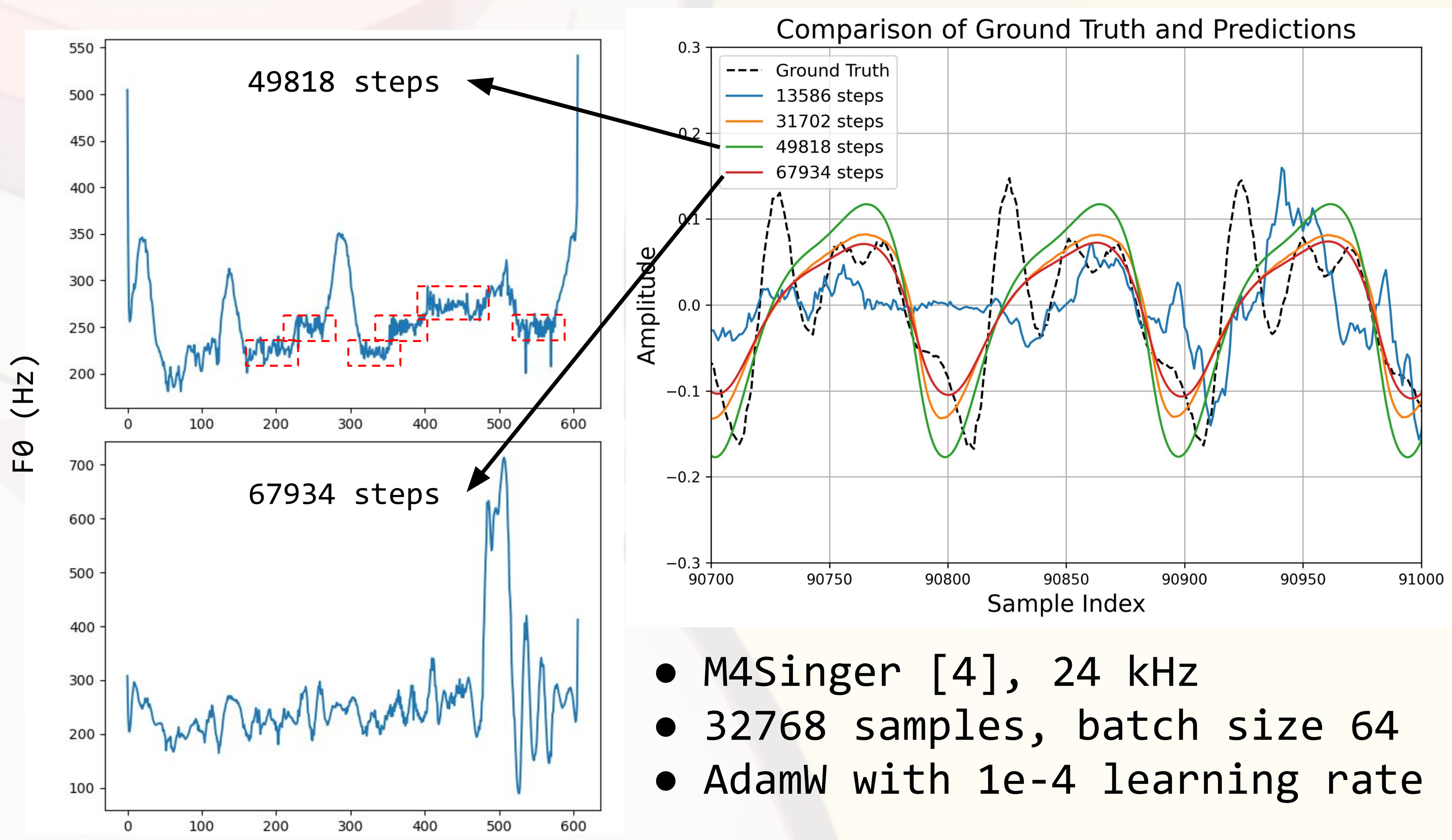
1. Can be **accelerated on the GPU** using parallel scan [3].
2. Guarantee **time-varying stability** when poles are inside the unit circle.



G: filter gain, p_i: pole, r_i: residue, M: filter order

End-to-End Copy Synthesis Experiment





- M4Singer [4], 24 kHz
- 32768 samples, batch size 64
- AdamW with 1e-4 learning rate

Results

- < 50k steps: Successfully captures the instantaneous phase and the glottal pulses are aligned.
- > 50k steps: Complex amplitude and frequency modulations emerge and the learnt $f_\theta(n) \neq$ instantaneous frequency. The encoder tries to fit the **non-determistic components.**

Conclusions and Future Works

- End-to-end phase modelling is feasible with frame-wise phase accumulation and time-domain loss function.
- Insert more zeros to the filter (longer FIRs) to increase capacity.
- Modelling the stochastic components by neural nets with regularisation to avoid simply copying the input.

Reference

[1] Yu, C.-Y., & Fazekas, G. (2024). GOLF: A Singing Voice Synthesiser with Glottal Flow Wavetables and LPC Filters. TISMIR, 7(1), 316–330. doi: 10.5334/tismir.210
[2] Chen, S., & Toda, T. (2024). QHM-GAN: Neural Vocoder based on Quasi-Harmonic Modeling. Proc. Interspeech 2024, 3889-3893. doi: 10.21437/Interspeech.2024-2371
[3] M. Harris, S. Sengupta, and J. D. Owens (2007). Parallel prefix sum (scan) with CUDA. GPU gems, 3(39), 851–876.
[4] Zhang, L., Li, R., Wang, S., Deng, L., Liu, J., Ren, Y., ... & Zhao, Z. (2022). M4singer: A multi-style, multi-singer and musical score provided mandarin singing corpus. Advances in Neural Information Processing Systems, 35, 6914-6926.

UK Research and Innovation

centre for digital music

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