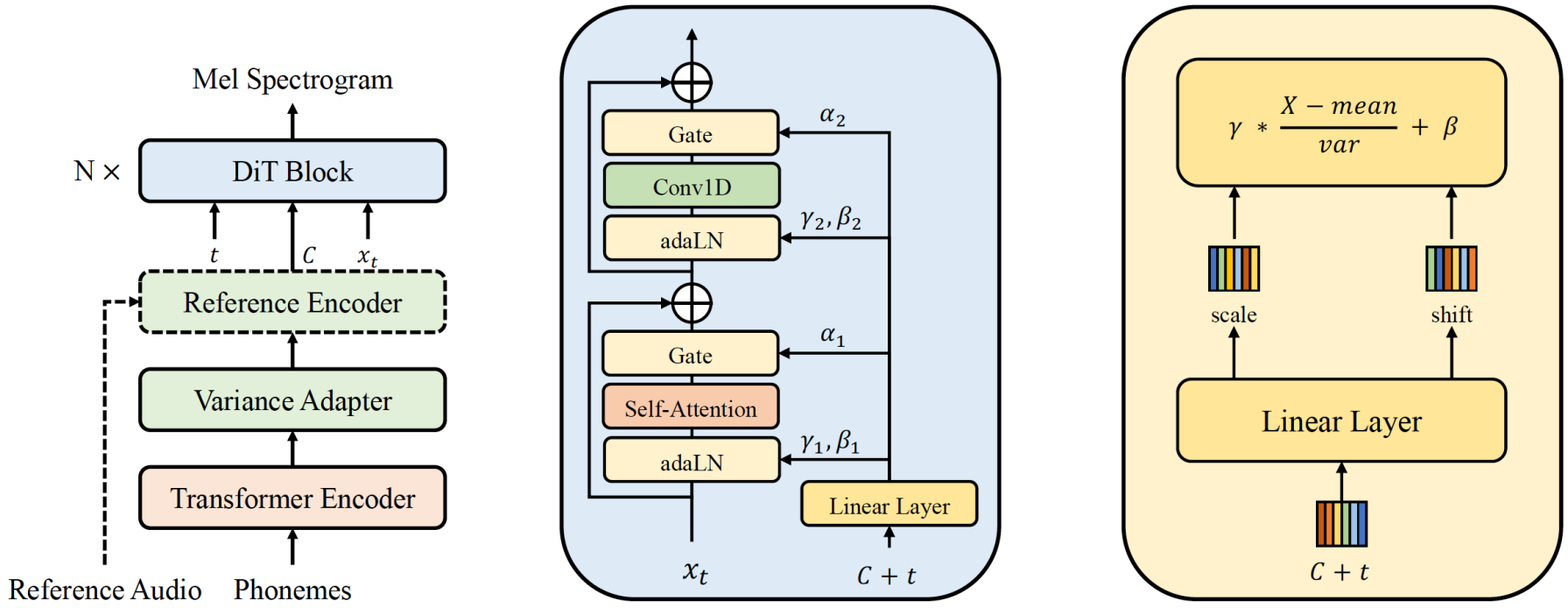
Diffusion Transformer for Adaptive Text-to-Speech

Haolin Chen



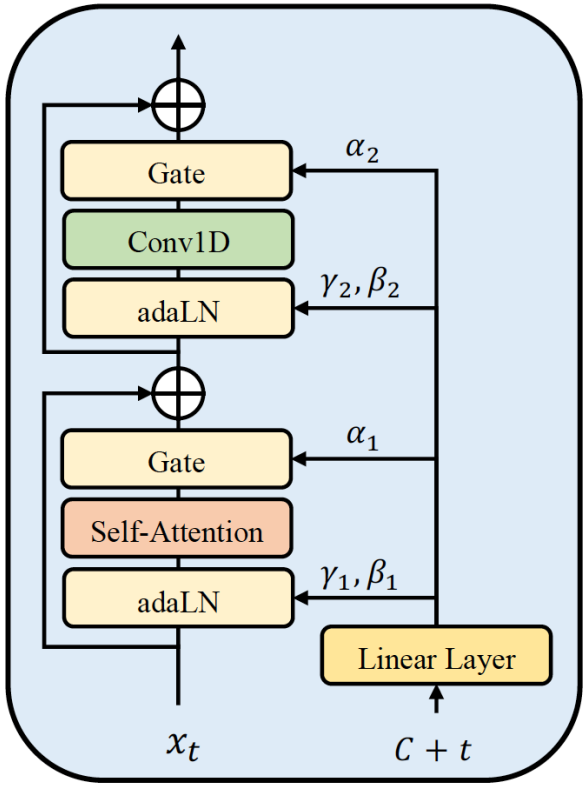
(a) *Overview* (b) *DiT block* (c) *Adaptive layer norm (adaLN)*

Figure 1: *The architecture of the DiT-based acoustic model. The reference encoder only exists in adaptive TTS systems.*

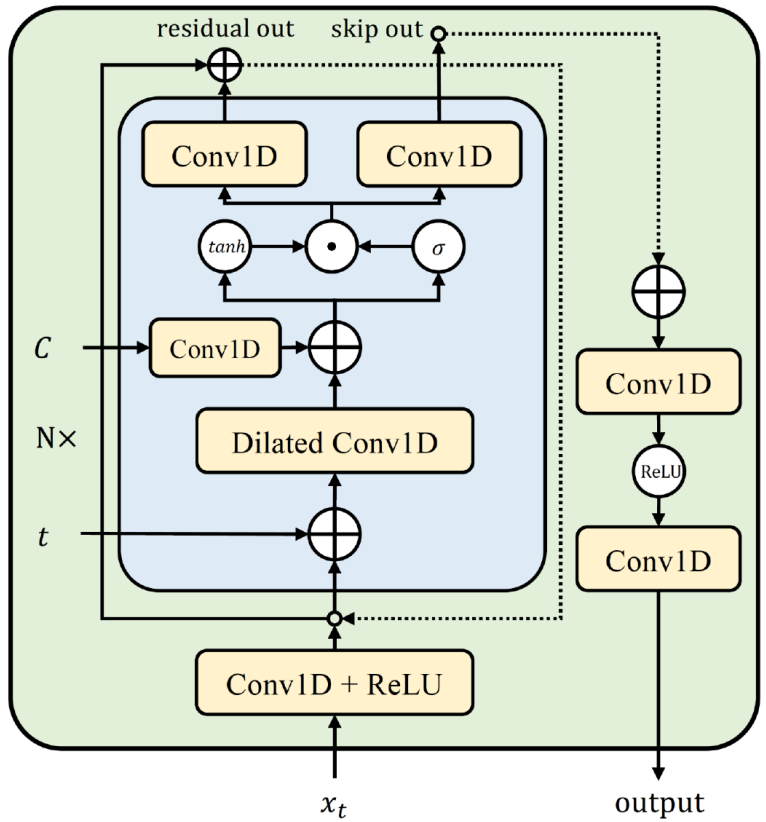
H. Chen and P. N. Garner, 'Diffusion Transformer for Adaptive Text-to-Speech', in Proc. 12th ISCA Speech Synthesis Workshop (SSW2023)

Faster & Higher Adaptability

- 2.4x faster than the non-causal WaveNet
- Adaptable layer norm module
  - 1.7M parameters, ~5%
  - <1 min. of audio required for 1 speaker



DiT Block



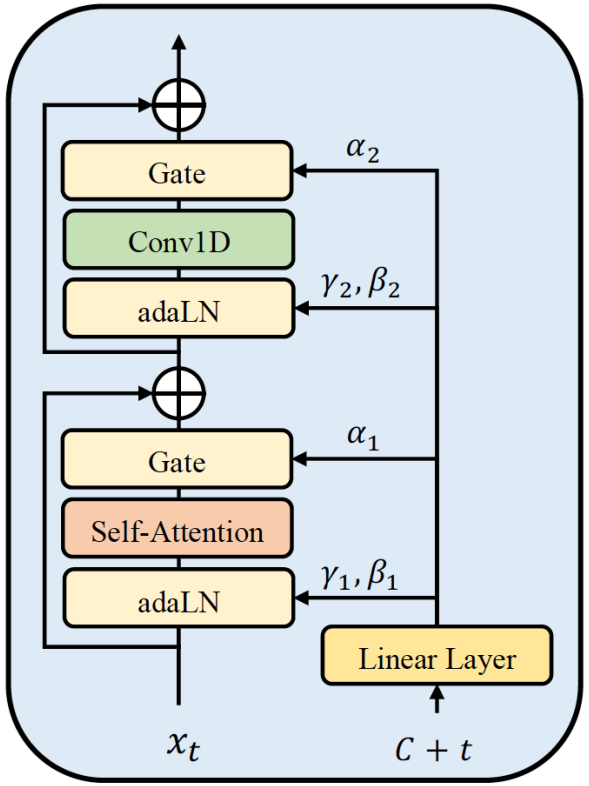
Non-causal WaveNet

H. Chen and P. N. Garner, 'Diffusion Transformer for Adaptive Text-to-Speech', in Proc. 12th ISCA Speech Synthesis Workshop (SSW2023)

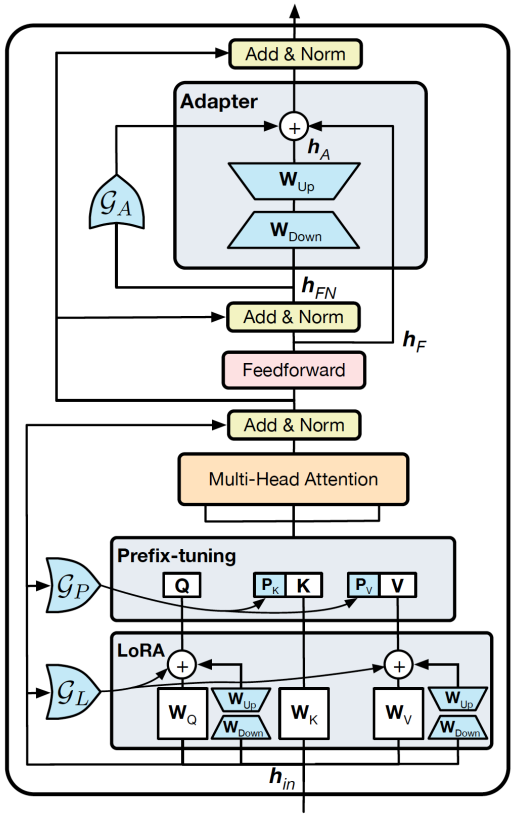
Architecture Unification Enables General Adaptation

Enabling parameter-efficient finetuning (PEFT) techniques designed for Transformer

- bottleneck adapters
- Prefix-Tuning
- LoRA
- (IA)^3
- UniPELT
- …



DiT Block



UniPELT

H. Chen and P. N. Garner, 'Diffusion Transformer for Adaptive Text-to-Speech', in Proc. 12th ISCA Speech Synthesis Workshop (SSW2023)