Diffusion Transformer for Adaptive Text-to-Speech

Haolin Chen

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

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

Enabling parameter-efficient finetuning (PEFT) techniques designed for Transformer
- bottleneck adapters
- Prefix-Tuning
- LoRA
- (IA)^3
- UniPELT
- ...