INTERSPEECH 2024

2 Sept. 2024

Mobile PresenTra:

NICT Fast Neural Text-To-Speech System on Smartphones With Incremental Inference of MS-FC-HiFi-GAN for Low-Latency Synthesis

Takuma Okamoto, Yamato Ohtani, and Hisashi Kawai

National Institute of Information and Communications Technology, Japan

Press release

- 21-language, fast and high-fidelity neural text-to-speech
- Synthesizing one second of speech at high speed in only 0.1 seconds using a CPU
- Fast synthesis with a latency of 0.5 seconds on a smartphone without network connection

https://www.nict.go.jp/en/press/2024/07/26-1.html

Hi-Fi-CAPTAIN corpus

High-fidelity and high-capacity conversational speech synthesis corpus developed by NICT
1 female and 1 male (English): 14K utts (parallel: 13K)
1 female and 1 male (Japanese): 19K utts (parallel: 18.5K)
ESPnet2-TTS recipe for JETS-based E2E-TTS

https://ast-astrec.nict.go.jp/en/release/hi-fi-captain/

1. Introduction

- **Background**
  - Fast and High-fidelity neural text-to-speech (TTS) can be realized
  - NICT developed 21-language neural TTS models and implemented them on VoiceTra

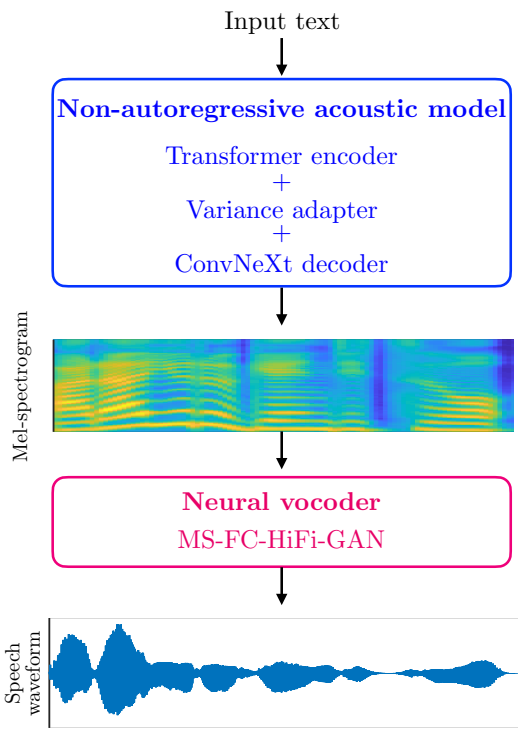
VoiceTra®

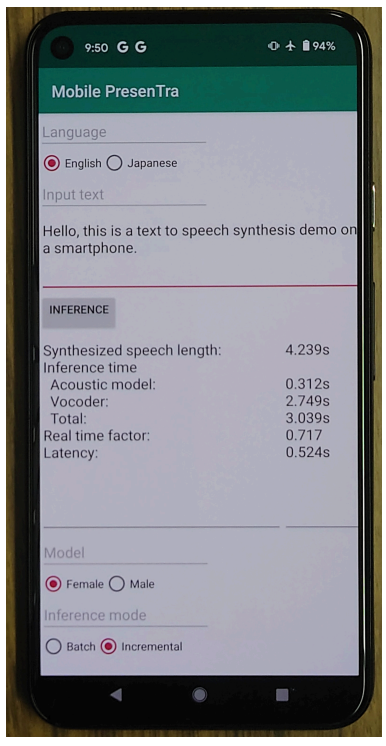
https://voicetra.nict.go.jp/en/index.html

  - Typical neural TTS systems are run on servers and require network connectivity
- **Purpose**
  - Fast and High-fidelity neural TTS models working on edge devices are required

2. Prototyped Mobile PresenTra

- **Fast and high-fidelity neural TTS model on smartphones @ 24 kHz**
  - Trained on PyTorch and implemented to smartphones using LibTorch and C++
  - Inference time, real-time factor and latency can be displayed

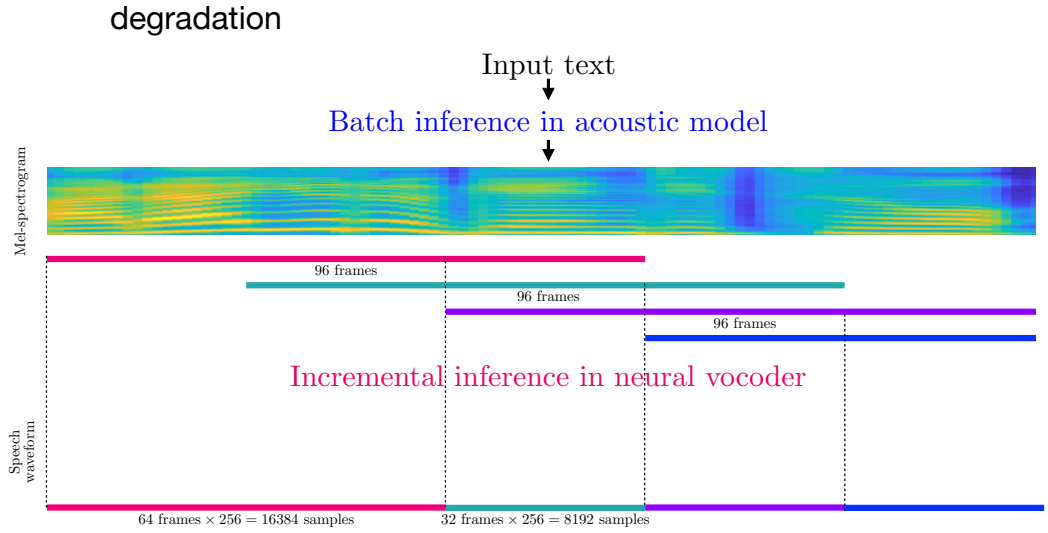




- **ConvNext-TTS & ConvNext-VC (Okamoto+ ICASSP 2024)**
  - E2E-S2S-TTS and VC models with ConvNeXt-based encoder and decoder instead of Transformer-based ones
    - ✳ Improving inference speed while keeping synthesis quality
- **MS-FC-HiFi-GAN (Yamashita+ IEEE Access 2024)**
  - HiFi-GAN-based neural vocoder with linear layer-based upsampling
    - ✳ Improving inference speed while keeping synthesis quality

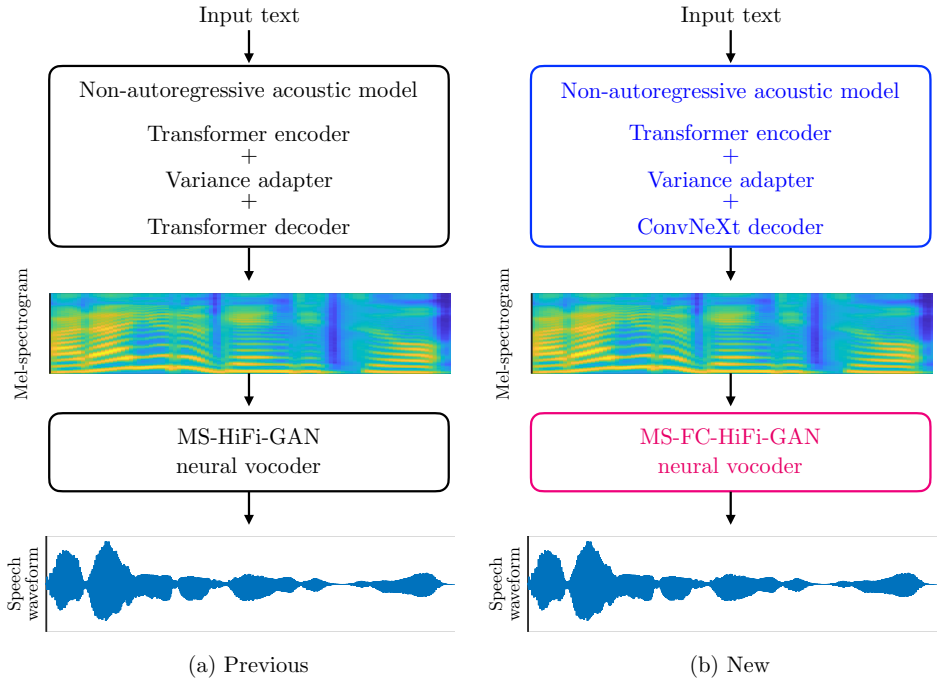
3. Incremental inference for low-latency synthesis

- **Incremental inference applied only to neural vocoder**
  - Acoustic model: Incremental inference increases synthesis error -> batch inference
  - Neural vocoder: Incremental inference without performance degradation



4. Results of evaluations

- **Previous systems VS Mobile PresenTra**
  - Dataset: Hi-Fi-CAPTAIN (English)



(a) Previous (b) New

	Previous system Acoustic model: Transformer encoder Transformer decoder Neural vocoder: MS-HiFi-GAN	Mobile PresenTra Acoustic model: Transformer encoder ConvNeXt decoder Neural vocoder: MS-FC-HiFi-GAN	Original
UTMOS	4.39	**4.43**	4.45
RTF for batch inference on server	0.2	**0.08**	
RTF for batch inference on smartphone	0.85	**0.30**	
Latency for incremental inference on server	0.35 s	**0.17 s**	
Latency for incremental inference on smartphone	1.13 s	**0.47 s**	