Interspeech 2024
1-5 September 2024, Kos, Greece

Mobile PresenTra: NICT Fast Neural Text-To-Speech System on Smartphones with Incremental Inference of MS-FC-HiFi-GAN for Law-Latency Synthesis

Takuma Okamoto, Yamato Ohtani, Hisashi Kawai

National Institute of Information and Communications Technology, Japan
okamoto@nict.go.jp, yamato.ohtani@nict.go.jp, hisashi.kawai@nict.go.jp

Abstract

For achieving fast and high-fidelity neural text-to-speech on edge smartphone devices without network connection, we NICT prototyped Mobile PresenTra by introducing non-autoregressive acoustic model with Transformer encoder and ConvNeXt decoder, and MS-FC-HiFi-GAN neural vocoder. Additionally, the incremental inference is applied only to neural vocoder for low-latency synthesis without performance degradation. Compared with a previous NICT system with Transformer encoder, Transforme decoder and MS-HiFi-GAN neural vocoder, the proposed Mobile PresenTra can realize high-fidelity and fast synthesis on a middle-range smartphone with a real-time factor of about 0.3 for batch inference, and a latency of less than 0.5 s for incremental inference. In the Show & Tell, attendees can freely experience the demonstration of Mobile PresenTra systems implemented on actual smartphones for English, Japanese and Chinese with arbitrary text input.

Index Terms: Edge computing, incremental inference, low-latency, Mobile PresenTra, neural text-to-speech

1. Introduction

With the recent development of neural network technologies, fast and high-fidelity text-to-speech (TTS) can be realized. Additionally, end-to-end neural TTS models, which can directly synthesize speech waveforms from input text with a single neural network [1–5]. We NICT have been developed 21-language[1] neural TTS models and implemented them in VoiceTra[2] which is a multilingual speech translation application for smartphones. In typical neural TTS systems including VoiceTra, the TTS engines are run on servers and require network connectivity. To reduce network communication costs, the development of neural TTS systems working on edge devices is quite important. To realize these systems, it is necessary to introduce neural TTS models with low computational requirements and high quality.

For this purpose, we NICT prototyped Mobile PresenTra, a fast and high-fidelity neural text-to-speech model working on edge smartphone devices without network connection. Additionally, the incremental inference is applied only to neural vocoder for low-latency synthesis without performance degradation. The results of evaluations indicate that the proposed Mobile PresenTra can realize high-fidelity and fast synthesis on a middle-range smartphone with a real-time factor of about 0.3 for batch inference, and a latency of less than 0.5 s for incremental inference, compared with a previous NICT system.

[1]Japanese, English, Chinese, Korean, Thai, French, Indonesian, Vietnamese, Spanish, Burmese, Filipino, Brazilian Portuguese, Khmer, Nepali, Mongolian, Arabic, Italian, Ukrainian, German, Hindi and Russian

[2]https://voicetra.nict.go.jp/en/

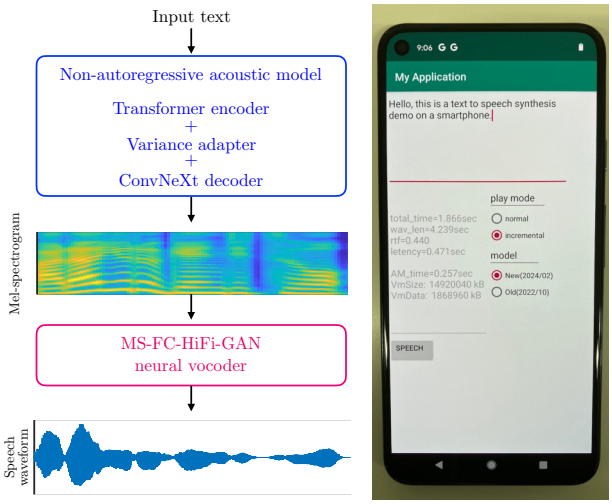

Figure 1: *(Left) Network archtecture of Mobile PresenTra. (Right) Mobile PresenTra with incremental inference of MS-FC-HiFi-GAN implemented on a smartphone (Google Pixel 5).*

2. Prototyped system: Mobile PresenTra

2.1. Fast and High-Fidelity Neural Text-To-Speech Models

Although a previous NICT neural TTS system introduced non-autoregressive Transformer encoder, non-autoregressive Transformer decoder [6], and Multi-stream HiFi-GAN neural vocoder [7], and realized fast and high-fidelity synthesis on servers, the inference speed on smartphone was not fast. To further accelerate the inference speed while keeping the synthesis quality, the prototyped Mobile PresenTra introduces non-autoregressive ConvNeXt decoder [5] and MS-FC-HiFi-GAN neural vocoder [8] (Fig. 1(left)).[3] Monotonic alignment search [9] is introduced to automatically train the alignment between input phonemes and output mel-spectrograms. The sampling frequency is 24 kHz.

[3]Although WaveNeXt [4] is faster than MS-FC-HiFi-GAN, WaveNeXt is not introduced because MS-FC-HiFi-GAN can realize higher quality synthesis [4]. Addtionally, the acoustic models and neural vocoders are separately trained and jointly finetuned instead of end-to-end training without mel-spectrogram representation because end-to-end models are sometimes unstable especially for very short text input not included in the training data.

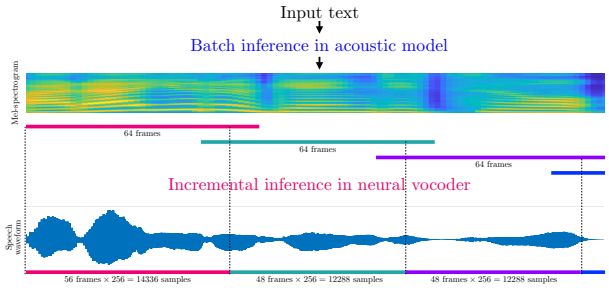

Figure 2: *Batch inference in acoustic model and incremental inference in neural vocoder.*

2.2. Incremental Inference Applied Only to Neural Vocoder for Law-Latency Synthesis without Performance Degradation

Although Mobile PresenTra can accelerate the inference speed by introducing fast and high-fidelity acoustic model and neural vocoder, the longer the input text, the longer it takes to play the synthesized speech waveform. To reduce the latency to play the synthesized speech waveform as much as possible, the incremental inference is applied only to the neural vocoder (Fig. 2). If the incremental inference is applied to the acoustic model, the synthesis error will increase because the acoustic model is non-autoregressive and is trained in a sequence-to-sequence manner. Conversely, since MS-FC-HiFi-GAN is not trained in a sequence-to-sequence manner, the incremental inference can be applied without performance degradation. In the implementation, the length of the subframes and the shift length of the subframes were 64 and 48, respectively.

2.3. Implemeted on Servers and Smartphones

All the neural network models were trained using PyTorch.[4] The trained acoustic model and MS-FC-HiFi-GAN were then jointly finetuned as [1]. The trained models were finally implemented on a server with Intel(R) Xeon(R) Gold 6152 CPU 2.10 GHz and a muddle-range smartphone (Google Pixel 5) using LibTorch[5] and C++. The application implemented on the smartphone can realize real-time neural TTS inference, and display the inference time, real-time factor, latency and memory consumption (Fig. 1(right)).

3. Evaluations

To evaluate the prototyped Mobile PresenTra on the server and middle-range smartphone, the inference speed was measured[6] and the synthesized speech samples were objectively evaluated using UTMOS [10]. In the evaluations, both the female and male English voices in Hi-Fi-CAPTAIN corpus [11] were used for training and synthesis. The results of the evaluations are shown in Table 1. The results of the evaluations indicated that the proposed Mobile PresenTra can realize high-fidelity and fast synthesis on a middle-range smartphone with a real-time factor of about 0.3 for batch inference, and a latency of less than 0.5 s for incremental inference, compared with the previous system.

[4] https://pytorch.org

[5] https://pytorch.org/cppdocs/installing.html

[6] Only one core of CPU was used for the server system evaluation.

Table 1: *Results of UTMOS for subjective evaluations, real-time factor (RTF) and latency for inference time measurements.*

	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**	

4. Show & Tell Demonstration

In the Show & Tell, attendees can freely experience the demonstration of Mobile PresenTra systems implemented on a middle-range smartphone (Google Pixel 5) and high-range smartphones (Google Pixel 8) for English, Japanese and Chinese with arbitrary text input.

5. Acknowledgement

We thank Masayuki Nishikawa for his help in implementing the neural TTS models on the server and smartphone.

6. References

[1] T. Hayashi, R. Yamamoto, T. Yoshimura, P. Wu, J. Shi, T. Saeki, Y. Ju, Y. Yasuda, S. Takamichi, and S. Watanabe, "ESPnet2-TTS: Extending the edge of TTS research," *arXiv:2110.07840*, 2021.

[2] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *Proc. ICML*, July 2021, pp. 5530–5540.

[3] D. Lim, S. Jung, and E. Kim, "JETS: Jointly training FastSpeech2 and HiFi-GAN for end to end text to speech," in *Proc. Interspeech*, Sept. 2022, pp. 21–25.

[4] T. Okamoto, H. Yamashita, Y. Ohtani, T. Toda, and H. Kawai, "WaveNeXt: ConvNeXt-based fast neural vocoder without iSTFT layer," in *Proc. ASRU*, Dec. 2023.

[5] T. Okamoto, Y. Ohtani, T. Toda, and H. Kawai, "ConvNeXt-TTS and ConvNeXt-VC: ConvNeXt-based fast end-to-end sequence-to-sequence text-to-speech and voice conversion," in *Proc. ICASSP*, Apr. 2024, pp. 12 456–12 460.

[6] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech 2: Fast and high-quality end-to-end text to speech," in *Proc. ICLR*, May 2021.

[7] T. Okamoto, T. Toda, and H. Kawai, "Multi-stream HiFi-GAN with data-driven waveform decomposition," in *Proc. ASRU*, Dec. 2021, pp. 610–617.

[8] H. Yamashita, T. Okamoto, R. Takashima, Y. Ohtani, T. Takiguchi, T. Toda, and H. Kawai, "Fast neural speech waveform generative models with fully-connected layer-based upsampling," *IEEE Access*, vol. 12, pp. 31 409–31 421, 2024.

[9] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-TTS: A generative flow for text-to-speech via monotonic alignment search," in *Proc. NeurIPS*, Dec. 2020, pp. 8067–8077.

[10] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, "UTMOS: UTokyo-SaruLab system for VoiceMOS Challenge 2022," in *Proc. Interspeech*, Sept. 2022, pp. 4521–4525.

[11] T. Okamoto, Y. Shiga, and H. Kawai, "Hi-Fi-CAPTAIN: High-fidelity and high-capacity conversational speech synthesis corpus developed by NICT," https://ast-astrec.nict.go.jp/en/release/hi-fi-captain/, 2023.