Comparison of real-time multi-speaker neural vocoders on CPUs

Keisuke Matsubara^{1,2,*}, Takuma Okamoto², Ryoichi Takashima¹, Tetsuya Takiguchi¹, Tomoki Toda^{3,2} and Hisashi Kawai²

¹Graduate School of System Informatics, Kobe University, Rokkodaicho 1-1, Nada-ku, Kobe, 657-8501 Japan

²National Institute of Information and Communications Technology,
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289 Japan

³Information Technology Center, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, 464-8601 Japan

(Received 3 September 2021, Accepted for publication 16 November 2021)

Keywords: Speech synthesis, Neural vocoder, HiFi-GAN, MWDLP, LPCNet

1. Introduction

Text-to-speech (TTS) and voice conversion are important technologies of speech processing that have long been actively researched. In recent years, deep neural networks have realized great developments in the performance of these technologies, and the naturalness of synthetic speech has become very close to that of natural speech [1,2]. Notably, neural vocoders, which reconstruct speech waveforms from acoustic features using neural networks, can synthesize higher-quality speech waveforms than conventional source-filter vocoders. Starting with the introduction of the WaveNet vocoder [3], various neural vocoders have been proposed to date, and they have greatly contributed to the development of neural speech synthesis.

Because the WaveNet vocoder has an autoregressive architecture that consists of large convolutional layers, its inference speed is very slow and it is not suitable for applications that required low latency. To solve this problem, many neural vocoders have been proposed that can synthesize high-quality speech waveforms in real time. In addition, to reduce the deterioration in speech that occurs when neural vocoders try to synthesize the speech of unseen speakers, many approaches have been proposed that train neural vocoders using mass multi-speaker datasets to acquire the ability to synthesize such speech [4]. These approaches have recently become popular because they can be combined with multi-speaker TTS or many-to-many voice conversion, and some lightweight, multi-speaker neural vocoders can now realize real-time synthesis on CPUs [5,6].

In this paper, we investigate the performances of real-time and multi-speaker neural vocoders: HiFi-GAN [5], Multiband WaveRNN with data-driven linear prediction (MWDLP) [6], and LPCNet [7]. Although official implementations of these vocoder models have been released by their authors, a unified performance comparison has not been conducted. Hence, this paper reveals the performance of these models with respect to both single-speaker and multi-speaker training. Regarding the synthesis speed, we compare the changes in the real-time factor (RTF) when using single-core and multi-core CPUs. In particular, we believe this investigation is important because

the synthesis speed of non-autoregressive models such as HiFi-GAN tends to improve remarkably when the number of CPU cores is increased [8]. Additionally, our previous research revealed that the input acoustic features proposed in LPCNet were robust enough that the TTS model did not perceive any degradation in the output [9]. Hence, we also use these acoustic features in HiFi-GAN and MWDLP. Experimental results show that, for both single-speaker and multi-speaker synthesis, HiFi-GAN is superior in quality and synthesis speed.

2. Real-time multi-speaker neural vocoder

We briefly introduce LPCNet, HiFi-GAN, and MWDLP, which can execute multi-speaker speech synthesis in real time on CPUs. Although the original LPCNet study evaluated only single-speaker speech synthesis, additional work [10] reported that LPCNet can perform multi-speaker speech synthesis. Although low-delay, real-time synthesis is possible for autoregressive models such as MWDLP and LPCNet, this study focuses on real-time synthesis assuming batch processing.

2.1. LPCNet

LPCNet is a neural vocoder based on WaveRNN [11], which has a lightweight architecture that uses recurrent neural networks. Unlike WaveRNN, which infers 16-bit raw audio samples, LPCNet infers the 8-bit residual signals between natural speech and the predicted speech computed by linear prediction coding (LPC). LPCNet consists of two neural network blocks called the frame rate network and sample rate network. The input features consist of 18-dimensional Bark-Frequency Cepstrum Coefficients (BFCCs) and two pitch parameters (period and correlation) for a sampling frequency of 16 kHz.

LPCNet is another autoregressive model that consists of gated recurrent units (GRUs), and therefore, like WaveNet, its inference is inherently slow. However, sparse coding that forces the lowest value of a weight matrix to zero is applied to accelerate inference. LPCNet can then synthesize speech waveforms in real time, even in a restricted environment with a mobile CPU [7].

2.2. MWDLP

MWDLP is also a WaveRNN-based neural vocoder. The original WaveRNN mainly consists of two simple RNNs to

*Work performed as an intern at NICT.

†e-mail: kmatsubara@stu.kobe-u.ac.jp

[doi:10.1250/ast.43.121]

effectively predict 16-bit audio samples, which reduces the number of parameters so that synthesis is fast. However, WaveRNN requires a GPU for real-time inference. Several modifications were introduced in MWDLP to improve inference speed while maintaining synthesis quality, such as reducing the number of audio bits, multi-band processing, and LPC. Owing to these modifications, MWDLP realizes robust synthesis that can handle unseen speakers and real-time synthesis on a single CPU core.

2.3. HiFi-GAN

HiFi-GAN is a generative adversarial network (GAN)-based neural vocoder consisting of a lightweight generator using transposed convolution and two sophisticated discriminators. The generator mainly consists of convolution and transposed convolution layers. When acoustic features are input, the generator upsamples them through transposed convolutions until the length of the output sequence matches the temporal resolution of the audio waveforms. After transposed convolution, multi-receptive field fusion (MRF) is performed. In MRF, the outputs of the convolution layers with various receptive fields are summed so that the various components of frequency in speech waveforms are correctly captured. The discriminator consists of two sub-discriminators, a multi-scale discriminator (MSD) and multi-period discriminator (MPD). The architecture of MSD is a mixture of several sub-discriminators operating on different sampling frequencies. In MPD, input audio samples of length T are sampled for each period p and reshaped into a 2-dimensional feature of shape $(T/p) \times p$. Multiple values of p are used, and sub-discriminators are prepared and trained for each one. MSD and MPD also properly capture the various components of frequency. Using these methods, HiFi-GAN has achieved high-quality synthesis on a lightweight generator and real-time synthesis on a single CPU.

3. HiFi-GAN with LPCNet features

We introduce the use of the LPCNet features for HiFi-GAN. Traditionally, HiFi-GAN uses 80-dimensional mel-spectrograms for input features. When the sampling frequency is 24 kHz, the LPCNet features are 32-dimensional features that consist of 30-dimensional BFCCs, pitch period, and pitch correlation [9,10]. Because the LPCNet feature comprises separate vocal tract and vocal cord information, it is relatively easy for an acoustic model to control. In [9], it was reported that there is little deterioration in quality when combined with a TTS model. Moreover, the LPCNet features are robust against the deterioration due to acoustic models.

We also conducted a preliminary experiment on MWDLP with LPCNet features, but found it could not synthesize properly. We believe the parameters of MWDLP were adjusted for mel-spectrogram inversion, and further investigation is a topic of future work.

4. Experiments

4.1. Experimental conditions

We conducted subjective experiments to evaluate the perceptual qualities of LPCNet, MWDLP, and HiFi-GAN for a sampling frequency of 24 kHz. All the models were implemented and modified using the official implementations

of the authors.^{*†‡} We used 7,697 utterances by a Japanese female speaker in the JSUT corpus [12] for single-speaker training, and about 13,000 utterances by 100 Japanese speakers in the JVS corpus [12] for multi-speaker training. In the single-speaker training, 7,497 utterances were used to train the neural vocoders, and the remaining 200 utterances were used for evaluation and validation. In multi-speaker training, 12,347 utterances spoken by 95 speakers (JVS006 to JVS100) were used for training, and 60 utterances by 2 unseen female and male speakers (JVS001 and JVS004) were used for evaluation. In the TTS condition, we used 4,800 sentences (Basic5000-0201 to Basic5000-5000) of JSUT for which HTS-style context labels based on manual annotation were available.

For the TTS condition, we used a FastSpeech-based acoustic model [13] with full-context label input to estimate the accent properly [9], which was complemented using ESPnet-TTS [14]. Although the full-context labels with the JSUT corpus consist of 575-dimensional features, we used simple 47-dimensional vectors constructed from 38-dimensional phoneme one-hot vectors and nine-dimensional accentual label vectors, as in [9].

The network architecture of HiFi-GAN was the same as that of the original implementation, and we used two models, referred to as V1 and V2, with 512 and 128 initial hidden channels, respectively [5]. As input features, we used two types of acoustic features, 80-dimensional log-mel spectrograms and 32-dimensional LPCNet features consisting of 30-dimensional BFCCs and 2-dimensional pitch features. To calculate the mel-spectrograms, we limited the frequency range (80 to 7,600 Hz) [2], and set the window and shift lengths to 42.7 ms and 10 ms, respectively. To calculate the BFCCs, spectrum analysis was performed with a window length of 20 ms and a frame shift of 10 ms, and a Bark-scale filter bank was employed. Pitch was calculated based on an open-loop cross-correlation search. Although the original HiFi-GAN used $256\times$ upsampling, we applied $240\times$ upsampling to obtain a 24 kHz resolution from input features with a 10-ms frame shift. Therefore, we set the upsampling rates of the transposed convolution layers to [5, 4, 3, 4] and the kernel sizes to [11, 8, 7, 8]. For the TTS condition, we fine-tuned the HiFi-GAN vocoders using acoustic features estimated by the FastSpeech-based acoustic model because, like [5], we found that synthetic speech substantially deteriorated without it. Our pre-trained models were trained with 1M steps and fine-tuned with 200k steps.

The network structure of the MWDLP vocoder was the same as that in [6]. The input features were 80-band log-mel spectrograms with a band-limited frequency range. The window and shift lengths were set to 42.7 ms and 10 ms, respectively. Eight data-driven LP coefficients and six bands were used. Preliminary experiments of the TTS conditions and fine-tuning of a vocoder model revealed the synthetic speech was of very poor quality; hence, we excluded these conditions from the subjective evaluations.

^{*}<https://github.com/mozilla/LPCNet>

[†]<https://github.com/patrickltobing/cyclevae-vc-neuralvoco>

[‡]<https://github.com/jik876/hifi-gan>

Fig. 1 Result of RTFs for inference using an Intel Xeon 6152 or Intel core i5-9400.

The network structure of LPCNet was the same as that in [10]. The input features were the 32-dimensional LPCNet features also used in HiFi-GAN.

4.2. RTF evaluation

Figure 1 shows the RTF results of inference using an Intel Xeon 6152 or Intel core i5-9400. We evaluated HiFi-GAN under two input feature conditions and found there was no difference in the RTFs. Therefore, Fig. 1 shows the results of RTFs for HiFi-GAN conditioned on mel-spectrograms. We confirmed that the RTF with a single-core CPU was less than 1.0 in all models, and they could realize real-time synthesis. We also confirmed that real-time synthesis was maintained even when the FastSpeech-based acoustic model (RTF = 0.05 [9]) and neural vocoders were combined. In all models except for HiFi-GAN V1, improvements in RTF were slight even when the number of CPU cores was increased. Regarding MWDLP and LPCNet, since they have autoregressive architectures, parallel processing cannot be performed even if the number of CPU cores is increased. Because HiFi-GAN V2 has a lightweight architecture, it realized sufficiently high-speed synthesis with a single-core CPU, and there was little improvement when the number of cores increased. HiFi-GAN V1 yielded a significantly improved RTF when the number of CPU cores increased: from 0.4 for a single core to 0.06 for 16 cores, as in [8]. MWDLP obtained an RTF of 0.87 in a single-core CPU of Intel Xeon 6152, whereas conventional MWDLP obtained an RTF of 0.61 [6]. We repeated the measurement with another CPU (Intel Core i5-9400), and obtained an RTF of 0.64, indicating that the inference speed of MWDLP depends on factors such as the type of CPU.

4.3. Subjective evaluation

We conducted mean opinion score (MOS) tests with a five-point scale [15] to evaluate the subjective perceptual quality of the synthesized speech waveform. Ten native Japanese speakers without hearing loss listened to the synthesized speech samples using headphones (18 utterances $\times$ 22 conditions = 396 utterances).

Figure 2 shows the result of MOS test with the single-speaker model. HiFi-GAN V1 obtained the highest analysis-synthesis score, and there was no difference in quality with respect to the input acoustic features. MWDLP and LPCNet outputs tended to have more blurring components than the original sound, and we think that this decreased their scores. MWDLP generated low levels of noise with frequency components near the band boundary due to multi-band

Fig. 2 Result of MOS test with the single-speaker model.

Fig. 3 Result of the MOS test with the multi-speaker model using the speech of an unseen female speaker (JVS004).

Fig. 4 Result of the MOS test with the multi-speaker model using the speech of an unseen male speaker (JVS001).

processing, and we also believe that this contributed to the low score. In the TTS conditions, there was no significant difference in the results of HiFi-GAN with respect to the analysis-synthesis, and LPCNet scored higher than the analysis-synthesis, as in [9]. As mentioned in Sect. 4.1, HiFi-GAN was fine-tuned with the output features from the acoustic model. Since we found the quality deteriorated without fine-tuning, we need to improve their robustness to such deteriorated features. LPCNet with TTS also scored higher than the analysis-synthesis, as in [9]. The output features from the acoustic model tend to have smoothed components in the high frequencies, and hence we believe that it suppressed the deterioration of the perceptual quality caused by the blurring component in the synthetic speech of LPCNet.

Figures 3 and 4 show the results of the MOS tests for analysis-synthesis by multi-speaker models. HiFi-GAN and MWDLP achieved significantly better scores than LPCNet, but the overall scores tended to be lower than the results of single-speaker models. HiFi-GAN V1 scored higher than HiFi-GAN V2 under all conditions. In the female-speaker

experiments, as shown in Fig. 3, the use of the mel-spectrogram as the input feature significantly improved the score when compared with the use of LPCNet features. By contrast, there was no difference between the results obtained using the mel-spectrograms and LPCNet features for the male speaker (Fig. 4), as in the results of the single-speaker models (Fig. 2). We believe that because the LPCNet features consist of compact 32-dimensions, and they were more compressed than the mel spectrograms, the robustness of the features to unseen female speakers might be insufficient. HiFi-GAN V2 and MWDLP obtained the same scores in the experiment with a female speaker in Fig. 3, and HiFi-GAN obtained a higher score under all other conditions.

Although an official implementation is not provided, a neural vocoder was recently proposed that achieves high-quality synthesis that is equal to or better than HiFi-GAN [16]. Comparison with this model is a topic for future work.

5. Conclusion

We evaluated the inference speed and synthesis quality of HiFi-GAN, MWDLP, and LPCNet vocoders. In terms of inference speed, we confirmed that all the models realized real-time synthesis with a single CPU core. Moreover, HiFi-GAN V1 can be accelerated using multiple CPU cores. In terms of synthesis quality, HiFi-GAN yielded higher synthesis quality than MWDLP and LPCNet under both single speaker and multi-speaker conditions. Furthermore, both single speaker and multi-speaker HiFi-GAN models obtained the same quality for unseen male speaker synthesis whether mel-spectrograms or LPCNet features were used.

References

- [1] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior and K. Kavukcuoglu, “WaveNet: A generative model for raw audio,” *Proc. SSW9*, p. 125 (2016).
- [2] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. J. Skerry-Ryan, R. A. Saurous, Y. Agiomyriannakis and Y. Wu, “Natural TTS synthesis by conditioning WavaNet on mel spectrogram predictions,” *Proc. ICASSP 2018*, pp. 4779–4783 (2018).
- [3] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda and T. Toda, “Speaker-dependent WaveNet vocoder,” *Proc. Interspeech 2017*, pp. 1118–1122 (2017).
- [4] J. L.-Trueba, T. Drugman, J. Latorre, T. Merritt, B. Putrycz, R. B.-Chicote, A. Moinet and V. Aggarwal, “Towards achieving robust universal neural vocoding,” *Proc. Interspeech 2019*, pp. 181–185 (2019).
- [5] J. Kong, J. Kim and J. Bae, “HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis,” *Proc. NeurIPS 2020*, pp. 17022–17033 (2020).
- [6] P. L. Tobing and T. Toda, “High-fidelity and low-latency universal neural vocoder based on multiband WaveRNN with data-driven linear prediction for discrete waveform modeling,” *Proc. Interspeech 2021*, pp. 2217–2221 (2021).
- [7] J. Valin and J. Skoglund, “LPCNet: Improving neural speech synthesis through linear prediction,” *Proc. ICASSP 2019*, pp. 5891–5895 (2019).
- [8] K. Matsubara, T. Okamoto, R. Takashima, T. Takiguchi, T. Toda, Y. Shiga and H. Kawai, “Investigation of training data size for real-time neural vocoders on CPUs,” *Acoust. Sci. & Tech.*, **42**, 65–68 (2021).
- [9] K. Matsubara, T. Okamoto, R. Takashima, T. Takiguchi, T. Toda, Y. Shiga and H. Kawai, “Full-band LPCNet: A real-time neural vocoder for 48 kHz audio with a CPU,” *IEEE Access*, **9**, 94923–94933 (2021).
- [10] K. Matsubara, T. Okamoto, R. Takashima, T. Takiguchi, T. Toda, Y. Shiga and H. Kawai, “High-intelligibility speech synthesis for dysarthric speakers with LPCNet-based TTS and CycleVAE-based VC,” *Proc. ICASSP 2021*, pp. 7058–7062 (2021).
- [11] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. van den Oord, S. Dieleman and K. Kavukcuoglu, “Efficient neural audio synthesis,” *Proc. ICML 2018*, pp. 2415–2424 (2018).
- [12] S. Takamichi, R. Sonobe, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji and H. Saruwatari, “JSUT and JVS: Free Japanese voice corpora for accelerating speech synthesis research,” *Acoust. Sci. & Tech.*, **41**, 761–768 (2020).
- [13] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao and T.-Y. Liu, “FastSpeech: Fast, robust and controllable Text to Speech,” *Proc. NeurIPS 2019*, pp. 3165–3174 (2019).
- [14] T. Hayashi, R. Yamamoto, K. Inoue, T. Yoshimura, S. Watanabe, T. Toda, K. Takeda, Y. Zhang and X. Tan, “ESPnet-TTS: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit,” *Proc. ICASSP 2020*, pp. 7654–7658 (2020).
- [15] ITU-T Recommendation P. 800, *Methods for subjective determination of transmission quality* (1996).
- [16] W. Jang, D. Lim, J. Yoon, B. Kim and J. Kim, “UnivNet: A neural vocoder with multi resolution spectrogram discriminators for high-fidelity waveform generation,” *Proc. Interspeech 2021*, pp. 2207–2211 (2021).