# FIRNET: FUNDAMENTAL FREQUENCY CONTROLLABLE FAST NEURAL VOCODER WITH TRAINABLE FINITE IMPULSE RESPONSE FILTER

*Yamato Ohtani*[1], *Takuma Okamoto*[1], *Tomoki Toda*[2,1], *Hisashi Kawai*[1]

[1]National Institute of Information and Communications Technology, Japan
[2]Information Technology Center, Nagoya University, Japan

## ABSTRACT

Some neural vocoders with fundamental frequency ($f_0$) control have succeeded in performing real-time inference on a single CPU while preserving the quality of the synthetic speech. However, compared with legacy vocoders based on signal processing, their inference speeds are still low. This paper proposes a neural vocoder based on the source-filter model with trainable time-variant finite impulse response (FIR) filters, to achieve a similar inference speed to legacy vocoders. In the proposed model, FIRNet, multiple FIR coefficients are predicted using the neural networks, and the speech waveform is then generated by convolving a mixed excitation signal with these FIR coefficients. Experimental results show that FIRNet can achieve an inference speed similar to legacy vocoders while maintaining $f_0$ controllability and natural speech quality.

***Index Terms*—** Speech synthesis, neural vocoder, source-filter model, finite impulse response, fundamental frequency control

## 1. INTRODUCTION

A neural vocoder is a well-known neural waveform generation technique that allows us to convert acoustic features to high-quality speech waveforms. Since the invention of WaveNet [1], many neural vocoders have been proposed [2–6] and applied in speech generation systems, such as text-to-speech (TTS), voice conversion, and singing voice synthesis. For practical use, they require fundamental frequency ($f_0$) controllability and real-time generation speed on a single CPU. Therefore, it is important to develop neural vocoders that satisfy these requirements.

HiFi-GAN [5] is a fast and high-fidelity neural vocoder. Although HiFi-GAN achieves real-time inference, it does not have flexible $f_0$ controllability. To solve this problem, Harmonic-Net+ [7] introduced the source-filter model [8] and the quasi-periodic (QP) architecture [9, 10] into HiFi-GAN. In Harmonic-Net+, the downsampling network converts a source excitation signal generated from the $f_0$ to downsampled latent representations. The speech waveform is then generated using the QP-HiFi-GAN module. Another approach based on HiFi-GAN, Source-filter HiFi-GAN (SiFi-GAN) [11], has been proposed. This framework has two networks: the source network, which applies the QP architecture, and the filter network, which comprises the HiFi-GAN module. In SiFi-GAN, the source network converts the $f_0$-dependent sine signal to the source excitation representation, from which the filter network generates the speech waveform. Although these neural vocoders achieve real-time generation speed on a single CPU and flexible $f_0$ controllability, their real-time factors (RTFs) for 24-kHz sampling waveforms range from approximately 0.4 to 0.8 [7, 11] because of the high complexity of the filtering process. In contrast, the TTS acoustic models reported in [12] are much faster than the neural vocoders described above on a single CPU. Therefore, for real-time operation in speech generation systems, source-filter-based neural vocoders require higher inference speed than those described above.

According to the theory of the source-filter model [8], a speech waveform is generated by convolving a source excitation signal with an impulse response of the vocal tract. That is, there is a possibility of realizing a high-speed neural vocoder based on the source-filter model by estimating appropriate impulse responses using neural networks. To test this hypothesis, we propose a neural vocoder with trainable time-variant finite impulse response (FIR) filters. In the proposed model, which we call FIRNet, neural networks convert $f_0$-independent acoustic features to multiple FIR coefficients. The speech waveform is then generated through the intermediate residual signal by filtering the source excitation signal with these FIR coefficients. Because the FIR filters do not depend on $f_0$ parameters, flexible $f_0$ control is possible by changing source excitation signals. In addition, this model achieves very high inference speed because of the low computational cost incurred by simple linear filtering. Experimental results show that the inference speed of the proposed model on a single CPU is almost the same as that of legacy vocoders, such as the WORLD synthesizer [13], while the speech quality and $f_0$ controllability are preserved [1].

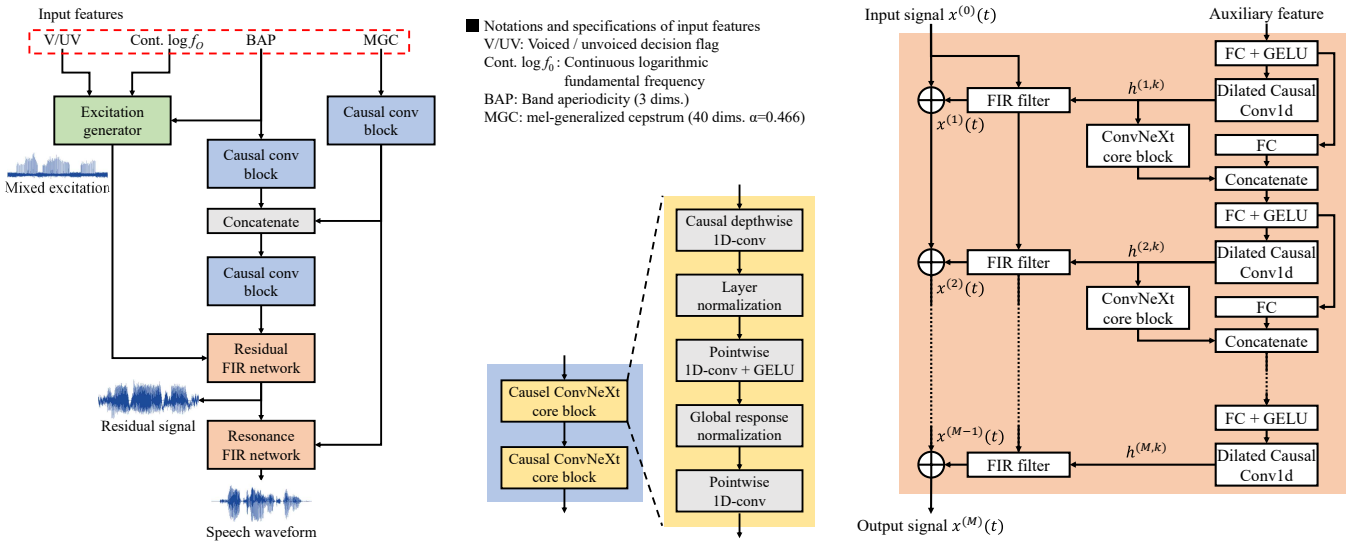
## 2. RELATED WORK

Several neural vocoders based on the source-filter model have been proposed previously. In [7, 11, 14, 15], speech waveforms are generated by filtering excitation signals nonlinearly with neural networks. LPCNet [4, 16] predicts residual signals using neural networks and then generates speech waveforms using linear prediction with linear predictive coding parameters, that is, infinite impulse response (IIR) filtering. The NITECH end-to-end TTS system [17] introduced the source-filter model using FIR filtering into the waveform generation module. In this module, speech waveforms are generated from predicted residual signals by the differentiable mel-cepstral synthetic filter [18]. This synthetic filter employs the cascade connection of multiple FIR filters instead of the IIR filter because of the need to satisfy bounded-input bounded-output (BIBO) stability.

In contrast to the above neural vocoders, FIRNet generates speech only by conversion from source excitation signals to residual signals and speech waveforms, in a step-by-step manner, by linear filtering with multiple FIR filters whose coefficients are predicted by neural networks. In addition, FIRNet guarantees stable speech generation because of its BIBO stability.

Corresponding author: yamato.ohtani@nict.go.jp

[1]Speech samples are available on `https://ast-astrec.nict.go.jp/demo_samples/firnet_icassp2024/`




(a) Network architecture (b) Causal convolution block (c) Residual/resonance FIR network

**Fig. 1**. Overview of FIRNet and sub-modules.

## 3. PROPOSED METHOD: FIRNET

Fig. 1 shows an overview of FIRNet and its sub-modules. FIRNet (Fig. 1(a)) is composed of the excitation generator, three causal convolution blocks (Fig. 1(b)), and the residual FIR network and resonance FIR network (Fig. 1(c)). As input features, we employ continuous $f_0$, voiced/unvoiced flag (V/UV), band aperiodicity (BAP) [19], and mel-generalized cepstrum coefficients (MGC) [20].

In inference, the excitation generator first generates a mixed excitation signal [21] shown in Fig. 2(a), using continuous $f_0$, V/UV, and BAP. BAP and MGC are converted to their latent representations by causal convolution blocks. Moreover, latent representations for the residual FIR network are generated from the BAP and MGC latent representations. Because the frequency characteristics of the mixed excitation signal differ from those of the actual residual signal, the residual FIR network conditioned by its corresponding latent representations converts the mixed excitation signal to the residual signal, as illustrated in Fig. 2(b). Finally, the residual signal is converted to a speech waveform by the resonance FIR network conditioned by the MGC latent representation, as shown in Fig. 2(c).

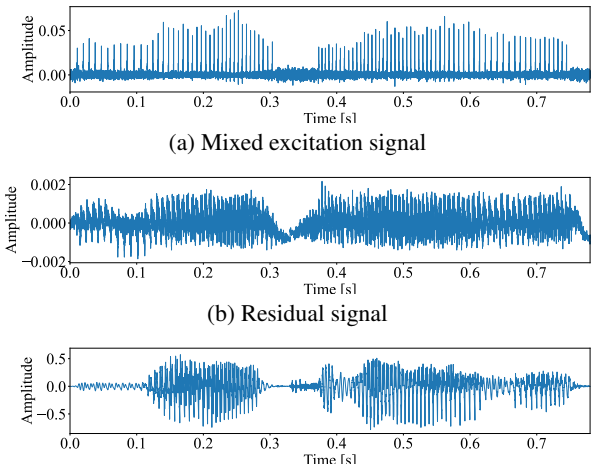
### 3.1. Details of generator modules

#### *3.1.1. Excitation generator*

The excitation generator generates a mixed excitation signal [13,21], which consists of a weighted sum of the $f_0$-dependent pulse train and the Gaussian noise based on BAP. The $t$th mixed excitation signal $s(t)$ is defined as follows:

$$s(t) = \begin{cases} g_p v^{(k)} * p(t) + g_n u^{(k)} * n(t) & \text{if voiced} \\ g_n n(t) & \text{if unvoiced} \end{cases}, \quad (1)$$

where $p(t)$, $n(t)$, $g_p$, and $g_n$ denote the $f_0$-dependent pulse train, Gaussian noise at time $t$, gain values for $p(t)$, and gain values for $n(t)$, respectively. $v^{(k)}$ and $u^{(k)}$ denote impulse responses of voiced and unvoiced segments, respectively. These are calculated by applying the inverse Fourier transform to BAP at the $k$th frame corresponding to $t$. $*$ denotes the convolution operation. Following [14], $g_p$ and $g_n$ are set to 0.1 and 0.003, respectively.




(a) Mixed excitation signal

(b) Residual signal

(c) Speech waveform

**Fig. 2**. Generated signals from FIRNet.

#### *3.1.2. Causal convolution block*

Fig. 1(b) illustrates the structure of our causal convolution block. To obtain highly accurate and effective latent representations, we employ the core module of the ConvNeXt V2 architecture [22]. This consists of an input depthwise convolution layer, layer normalization [23], pointwise convolution with Gaussian error linear unit (GELU) activation [24], global response normalization [22], and output pointwise convolution. In this paper, we apply causality to the input depthwise convolution layer and set its kernel size to 5. The causal convolution blocks include two ConvNeXt core blocks. The numbers of output channels for BAP, MGC, and the residual FIR network are 128, 256, and 128, respectively.

#### *3.1.3. FIR network*

Fig. 1(c) illustrates the structure of the FIR network architecture. In this network, an input signal is converted to an output signal using multiple FIR filters. In the figure, $h^{(1,k)}, h^{(2,k)}, \cdots, h^{(M,k)}$ represent the impulse response coefficients of the corresponding FIR

filters, $x^{(0)}(t)$ and $x^{(M)}(t)$ are the input and output signals, respectively, and $x^{(1)}(t), x^{(2)}(t), \cdots, x^{(M-1)}(t)$ represent the intermediate output signals, which are defined as follows:

$$x^{(m)}(t) = h^{(m,k)} * x^{(m-1)}(t) + x^{(m-1)}(t). \quad (2)$$

The impulse response coefficients of the $m$th FIR filter are estimated from the latent representations of an auxiliary feature and previous impulse response coefficients using the dilated causal convolution layer. In this paper, the number of latent channels is set to 128, the kernel sizes of the dilated causal convolution layers are set to 3, and the others are set to 1. The residual and resonance FIR networks include 8 FIR filters with tap sizes of 256. The dilation sizes of the 8 dilated causal convolution layers are 1, 2, 4, 8, 1, 2, 4, and 8.

### 3.2. Training criteria

FIRNet is a generative adversarial network (GAN) [25]. In this paper, we employ four losses as the objective for the generator: an adversarial loss with least squares criteria $\mathcal{L}_{adv}$ [26], a feature matching loss $\mathcal{L}_{fm}$ [27], a mel-spectral L1 loss $\mathcal{L}_{mel}$ for speech waveforms, and a source excitation regularization loss $\mathcal{L}_{reg}$ [11] for residual signals. The objective function $\mathcal{L}_G$ is defined as follows:

$$\mathcal{L}_G = \mathcal{L}_{adv} + \lambda_{fm}\mathcal{L}_{fm} + \lambda_{mel}\mathcal{L}_{mel} + \lambda_{reg}\mathcal{L}_{reg}, \quad (3)$$

where $\lambda_{fm}$, $\lambda_{mel}$, and $\lambda_{reg}$ denote the balancing hyperparameters for $\mathcal{L}_{fm}$, $\mathcal{L}_{mel}$, and $\mathcal{L}_{reg}$, respectively. We empirically set $\lambda_{fm}$, $\lambda_{mel}$, and $\lambda_{reg}$ to 2.0, 50.0, and 20.0, respectively. In addition, we employed HiFi-GAN's discriminator [5] according to the results of preliminary experiments.

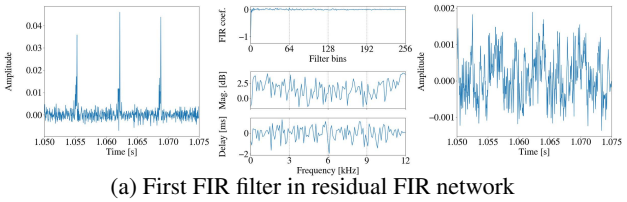
## 4. EXPERIMENTAL EVALUATION

### 4.1. Experimental setup

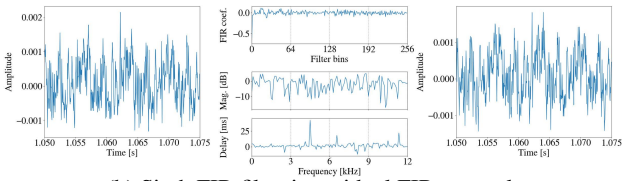
In experiments, we objectively and subjectively evaluated the performance of FIRNet in the analysis–synthesis scenario. We used a Japanese male speaker from the Hi-Fi-CAPTAIN corpus [28].We randomly selected 1,000 utterances from the parallel and non-parallel training subsets to use as training data. We used 100 validation and 100 evaluation utterances from their corresponding subsets. We downgraded all recording data from 48 kHz and 24 bits to 24 kHz and 16 bits and normalized them to $-17$ dB. In feature extraction, we set the frame shift to 5.0 ms and the FFT size to 1024. In $f_0$ extraction, we calculated $f_0$ contours by using DIO [13], HARVEST [29], SWIPE [30], RAPT [31], and REAPER[2]. From these contours, we then extracted the median of $f_0$ and the majority vote of V/UV in each frame. Spectra were calculated by CheapTrick [32] and converted to 40-dimensional MGCs whose warping coefficients were 0.466. BAP was extracted by D4C [19] and the number of BAP dimensions was 3.

The conventional systems that we used were the WORLD synthesizer [13] and SiFi-GAN [11]. All neural vocoders used in the experiments were constructed using the Adam optimizer [33]. The learning rate, $\beta_1$, $\beta_2$, and $\epsilon$ were set to 0.0002, 0.5, 0.8, and $1.0 \times 10^{-8}$, respectively. These optimizer settings were made for both the generator and discriminator. The number of update iterations was 500,000, and the learning rate was halved every 100,000 iterations. While the minibatch size of SiFi-GAN was the original setting in [11], that of FIRNet set to one utterance.

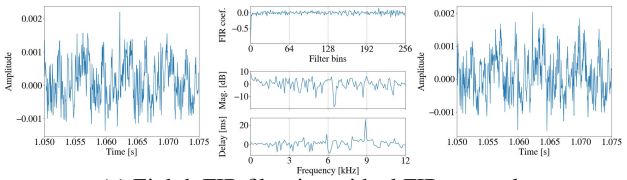
[2] https://github.com/google/REAPER



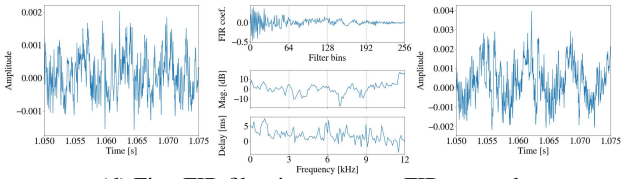

(a) First FIR filter in residual FIR network



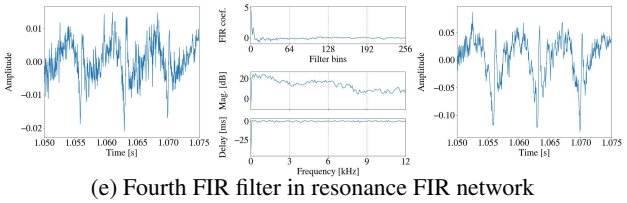

(b) Sixth FIR filter in residual FIR network



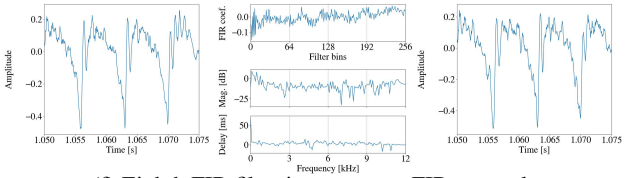

(c) Eighth FIR filter in residual FIR network




(d) First FIR filter in resonance FIR network




(e) Fourth FIR filter in resonance FIR network




(f) Eighth FIR filter in resonance FIR network

**Fig. 3**. Impulse responses of FIR filters at the vowel /a/ and the corresponding input and output signals. The left and right parts of the figure show the input and output signals, respectively. The central parts of the figure show (from top to bottom) the FIR coefficients, amplitude, and group delay characteristics.

### 4.2. Effectiveness of FIR filters

Fig. 3 shows the impulse responses of various FIR filters and their corresponding input and output signals. In the impulse responses of the residual FIR network, the first FIR filter in Fig. 3(a) dramatically changes the waveform shape of the source excitation signal. The remaining FIR filters, after the first, only slightly modify the waveform shapes of the input signals. This suggests that we may be able to reduce the number of FIR filters in the residual FIR network.

**Table 1**. Results of objective evaluation and numbers of model parameters. The best score in each $f_0$ condition is highlighted in bold.

| Model | MCD [dB] | RMSE | VUVE [%] | RTF |
|---|---|---|---|---|
| $1.0 \times f_0$ | | | | |
| WORLD | 6.87 | **0.029** | 4.47 | 0.106 |
| SiFiGAN | 5.85 | 0.037 | 4.16 | 0.576 |
| FIRNet (proposed) | **4.41** | 0.030 | **3.65** | **0.103** |
| $0.00 \times f_0$ | | | | |
| WORLD | 7.46 | - | - | 0.115 |
| SiFiGAN | 7.90 | - | - | 0.591 |
| FIRNet (proposed) | **6.23** | - | - | **0.103** |
| $0.25 \times f_0$ | | | | |
| WORLD | 11.21 | **0.073** | **20.61** | **0.100** |
| SiFiGAN | 8.06 | 0.096 | 21.14 | 0.566 |
| FIRNet (proposed) | **6.64** | 0.095 | 22.31 | 0.104 |
| $0.5 \times f_0$ | | | | |
| WORLD | 8.39 | **0.038** | **5.03** | **0.102** |
| SiFiGAN | 6.74 | 0.056 | 6.59 | 0.569 |
| FIRNet (proposed) | **6.04** | 0.192 | 16.42 | 0.103 |
| $2.0 \times f_0$ | | | | |
| WORLD | 7.19 | **0.026** | **5.15** | 0.12 |
| SiFiGAN | 6.70 | 0.097 | 8.18 | 0.574 |
| FIRNet (proposed) | **5.09** | 0.068 | 6.13 | **0.103** |
| $4.0 \times f_0$ | | | | |
| WORLD | 8.82 | **0.065** | **10.53** | 0.138 |
| SiFiGAN | 8.22 | 0.258 | 42.22 | 0.572 |
| FIRNet (proposed) | **6.99** | 0.179 | 25.62 | **0.104** |
| $8.0 \times f_0$ | | | | |
| WORLD | 10.30 | - | - | 0.180 |
| SiFiGAN | 9.20 | - | - | 0.574 |
| FIRNet (proposed) | **8.57** | - | - | **0.101** |
| Number of model parameters [M] | | | | |
| SiFi-GAN | 11.29 | | | |
| FIRNet (proposed) | 9.21 | | | |

In contrast, in the resonance FIR network, the filtered waveform shapes gradually change by convolving with successive FIR filters. In addition, the impulse responses in the resonance FIR network have shapes that differ from those in the residual FIR network. This implies that we need to make the impulse responses as long as possible by using many FIR filters or a FIR filter with a long tap size to obtain fine spectral envelopes.

### 4.3. Objective evaluation

We compared FIRNet with conventional methods in aspects of mel-cepstral distortion (MCD), root mean squared error of log $f_0$ (RMSE), V/UV decision error (VUVE), and RTF on a single CPU (AMD EPYC 7542). We used seven $f_0$ scaling conditions: 1.0, 0.0, 0.25, 0.5, 2.0, 4.0, and 8.0.

Table 1 shows the objective results. $f_0$ and V/UV could not be calculated in the $f_0$ scaling conditions of 0.0 and 8.0. The proposed method achieved the lowest MCD of all three methods in all $f_0$ conditions. With respect to generation speed, FIRNet was five times faster than SiFi-GAN because FIRNet does not require complex processing of time-domain signals due to the classical FIR filterings. Moreover, FIRNet has a higher generation speed than the WORLD synthesizer in the cases of large $f_0$ scaling conditions because the waveform generation algorithm in WORLD depends on pitch intervals.

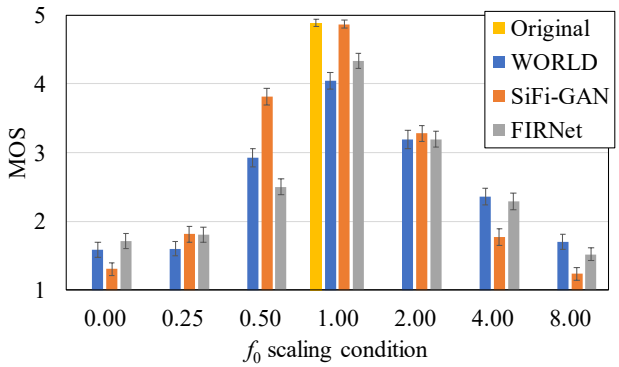



**Fig. 4**. Subjective results. Error bars show 95% confidence intervals.

### 4.4. Subjective evaluation

We conducted five-level mean opinion score (MOS) tests to evaluate the naturalness of the synthetic speech. Using the $f_0$ scaling conditions in Section 4.3, we compared WORLD, SiFi-GAN, FIRNet, and the original waveforms. We employed 20 listeners, who each evaluated 12 samples for each method and each $f_0$ scaling condition.

Fig. 4 shows the subjective results. For SiFi-GAN, although the synthetic speech with $f_0$ scaling conditions of 1.0, 0.5, and 2.0 had high quality, synthetic speech with other conditions was degraded dramatically. In contrast, FIRNet keeps a similar speech quality to WORLD in the $f_0$ scaling conditions of 0.0, 0.25, 4.0, and 8.0. In addition, FIRNet had higher speech quality than WORLD in the $f_0$ scale condition of 1.0. Therefore, FIRNet could achieve both good robustness for $f_0$ control and speech quality of the neural vocoder.

SiFi-GAN achieved higher speech quality than FIRNet in the case of the $f_0$ scaling conditions of 1.0. Moreover, the speech quality of FIRNet with a $f_0$ scaling condition of 0.5 was worse than that generated by any other method. By checking some samples of generated speech, we found that some FIR coefficients include slight harmonic structures. We conclude that these structures caused this degradation of speech quality in FIRNet. Therefore, we need to improve the speech quality further in future work.

## 5. CONCLUSION

This paper proposed a neural waveform generation model, FIRNet, which achieves fast waveform generation and $f_0$ controllability. Because the inference speed of FIRNet is comparable to that of the WORLD synthesizer on a single CPU, we believe that FIRNet can be applied in speech generation systems such as TTS, singing voice synthesis, and voice conversion without a bottleneck caused by processing speed. In the future, to improve speech quality, we will introduce multiple collaborative discriminators [34] and more effective structures into FIRNet. In addition, we will apply FIRNet to full-band neural waveform generation and in various end-to-end speech generation systems.

## 6. 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," in *Proc. SSW9*, Sept. 2016, p. 125.

[2] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and

T. Toda, "Speaker-dependent WaveNet vocoder," in *Proc. INTERSPEECH*, Aug. 2017, pp. 1118–1122.

[3] 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," in *Proc. ICML*, July 2018, pp. 2415–2424.

[4] J.-M. Valin and J. Skoglund, "LPCNet: Improving neural speech synthesis through linear prediction," in *Proc. ICASSP*, May 2019, pp. 5826–7830.

[5] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proc. NeurIPS*, Dec. 2020, pp. 17022–17033.

[6] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan, "WaveGrad: Estimating gradients for waveform generation," in *Proc. ICLR*, May 2021.

[7] K. Matsubara, T. Okamoto, R. Takashima, T. Takiguchi, T. Toda, and H. Kawai, "Harmonic-net: Fundamental frequency and speech rate controllable fast neural vocoder," *IEEE/ACM Trans. on Audio, Speech, and Lang. Proc.*, vol. 31, pp. 1902–1915, 2023.

[8] T. Chiba and M. Kajiyama, *The Vowel*, Tokyo-Kaiseikan Pub. Co., Ltd., 1942.

[9] Y.-C. Wu, T. Hayashi, P. L. Tobing, K. Kobayashi, and T. Toda, "Quasi-Periodic WaveNet: An autoregressive raw waveform generative model with pitch-dependent dilated convolution neural network," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 29, pp. 1134–1148, feb 2021.

[10] Y.-C. Wu, T. Hayashi, T. Okamoto, H. Kawai, and T. Toda, "Quasi-Periodic Parallel WaveGAN: A non-autoregressive raw waveform generative model with pitch-dependent dilated convolution neural network," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 29, pp. 792–806, Jan. 2021.

[11] R. Yoneyama, Y.-C. Wu, and T. Toda, "Source-Filter HiFi-GAN: Fast and pitch controllable high-fidelity neural vocoder," in *Proc. ICASSP*, June 2023.

[12] R. Luo, X. Tan, R. Wang, T. Qin, J. Li, S. Zhao, E. Chen, and T.Y. Liu, "Lightspeech: Lightweight and fast text to speech with neural architecture search," in *Proc. ICASSP*, 2021, pp. 5699–5703.

[13] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based highquality speech synthesis system for real-time applications," *IEICE Trans. Info. & Syst.*, vol. E99-D, no. 7, pp. 1877–1884, July 2016.

[14] X. Wang, S. Takaki, and J. Yamagishi, "Neural Source-Filter Waveform Models for Statistical Parametric Speech Synthesis," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 28, pp. 402–415, 2020.

[15] R. Yoneyama, Y.-C. Wu, and T. Toda, "Unified Source-Filter GAN with Harmonic-plus-Noise Source Excitation Generation," in *Proc. Interspeech*, 2022, pp. 848–852.

[16] K. Subramani, J.-M. Valin, U. Isik, P. Smaragdis, and A. Krishnaswamy, "End-to-end LPCNet: A neural vocoder with fully-differentiable lpc estimation," in *Proc. Iinterspeech*, 2022, pp. 818–822.

[17] T. Yoshimura, S. Takaki, K. Nakamura, K. Oura, Y. Hono, K. Hashimoto, Y. Nankaku, and K. Tokuda, "Embedding a differentiable mel-cepstral synthesis filter to a neural speech synthesis system," in *Proc. ICASSP*, 2023, pp. 1–5.

[18] T. Yoshimura, T. Fujimoto, K. Oura, and K. Tokuda, "SPTK4: An open-source software toolkit for speech signal processing," in *12th ISCASpeech Synthesis Workshop (SSW 2023)*, 2023, pp. 211–217.

[19] M. Morise, "D4C, A band-aperiodicity estimator for high-quality speech synthesis," *Speech Comm.*, vol. 84, pp. 57–65, Nov. 2016.

[20] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis," in *Proc. ICASSP*, 1994, pp. 1043–1046.

[21] H. Kawahara, Jo Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT," in *MAVEBA 2001*, Sept. 2001.

[22] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I.-S. Kweon, and S. Xie, "ConvNeXt V2: Co-designing and scaling convnets with masked autoencoders," *Proc. CVPR*, pp. 16133–16142, 2023.

[23] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv*, 2016.

[24] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv*, 2016.

[25] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. NIPS*, Dec. 2014, pp. 2672–2680.

[26] X. Mao, Q. Li, H. Xie, R. K. Lau, Z. Wang, and S. Smolley, "Least squares generative adversarial networks," in *Proc. ICCV*, Oct. 2017, pp. 2813–2821.

[27] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. C Courville, "MelGAN: Generative adversarial networks for conditional waveform synthesis," in *Proc. NeurIPS*, Dec. 2019, pp. 14910–14921.

[28] 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.

[29] M. Morise, "Harvest: A high-performance fundamental frequency estimator from speech signals," in *Proc. INTERSPEECH*, 2017, pp. 2321–2325.

[30] A. Camacho, "SWIPE: A sawtooth waveform inspired pitch estimator for speech and music," *Ph.D. Thesis, University of Florida*, 2007.

[31] D. Talkin, *A Robust Algorithm for Pitch Tracking (RAPT)*, chapter Speech Coding & Synthesis, pp. 495–518, Elsevier, 1995.

[32] M. Morise, "CheapTrick, A spectral envelope estimator for high-quality speech synthesis," *Speech Comm.*, vol. 67, pp. 1–7, Mar. 2015.

[33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.

[34] D. S. Dang, T. L. Nguyen, B. T. Ta, T. T. Nguyen, T. N. A. Nguyen, D. L. L., N. M. Le, and V. H. Do, "LightVoc: An Upsampling-Free GAN Vocoder Based On Conformer And Inverse Short-time Fourier Transform," in *Proc. Interspeech*, Aug. 2023, pp. 3043–3047.