# Harmonic-Net: Fundamental Frequency and Speech Rate Controllable Fast Neural Vocoder

Keisuke Matsubara, Takuma Okamoto, *Member, IEEE*, Ryoichi Takashima, *Member, IEEE*, Tetsuya Takiguchi, *Member, IEEE*, Tomoki Toda, *Senior Member, IEEE*, and Hisashi Kawai, *Member, IEEE*

***Abstract*—There is a need to improve the synthesis quality of HiFi-GAN-based real-time neural speech waveform generative models on CPUs while preserving the controllability of fundamental frequency ($f_o$) and speech rate (SR). For this purpose, we propose Harmonic-Net and Harmonic-Net+, which introduce two extended functions into the HiFi-GAN generator. The first extension is a downsampling network, named the excitation signal network, that hierarchically receives multi-channel excitation signals corresponding to $f_o$. The second extension is the layer-wise pitch-dependent dilated convolutional network (LW-PDCNN), which can flexibly change its receptive fields depending on the input $f_o$ to handle large fluctuations in $f_o$ for the upsampling-based HiFi-GAN generator. The proposed explicit input of excitation signals and LW-PDCNNs corresponding to $f_o$ are expected to realize high-quality synthesis for the normal and $f_o$-conversion conditions and for the SR-conversion condition. The results of experiments for unseen speaker synthesis, full-band singing voice synthesis, and text-to-speech synthesis show that the proposed method with harmonic waves corresponding to $f_o$ can achieve higher synthesis quality than conventional methods in all (i.e., normal, $f_o$-conversion, and SR-conversion) conditions.**

***Index Terms*—Fundamental frequency control, neural vocoder, speech-rate conversion, speech synthesis.**

## I. Introduction

SPEECH synthesis (SS) is one of the most important speech communication technologies. In recent years, many SS approaches using deep neural networks have been developed and their synthesis quality has significantly improved, even approaching that of natural speech [1], [2], [3]. In particular, neural speech waveform generative models (neural vocoders that reconstruct speech waveforms from acoustic features [4], [5]) have greatly improved synthetic speech quality, compared with that of conventional digital signal processing (DSP)-based source-filter vocoders [6], [7], [8], [9]. To date, many types of fast neural vocoders that achieve real-time SS have been proposed. They mainly use lightweight autoregressive (AR) models [10], [11], [12] or non-AR models [13], [14], [15], [16], [17], [18], [19], [20]. In particular, HiFi-GAN [21], which is a non-AR neural vocoder based on generative adversarial networks (GANs) [22], realizes high-quality SS for both single-speaker and multi-speaker models. Compared with conventional models [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], HiFi-GAN can synthesize higher quality speech waveforms while achieving real-time inference on CPUs. Therefore, HiFi-GAN has recently become widely used in text-to-speech (TTS) [23], voice conversion [24], and singing voice synthesis [25]. Additionally, extended models—Fre-GAN [26], UnivNet [27], Multi-stream HiFi-GAN [28], CARGAN [29], and iSTFTNet [30]—have been proposed to further improve the synthesis quality and synthesis speed. Furthermore, a HiFi-GAN-based decoder is used in some end-to-end TTS models that can directly synthesize speech waveforms from phoneme sequences using a single neural network [3], [31].

Similarly to conventional DSP-based source-filter vocoders, neural vocoders are required to be able to flexibly control attributes such as fundamental frequency ($f_o$ [32]) and speech rate (SR), in addition to speech quality. However, the controllability of neural vocoders is usually inferior to that of DSP-based source-filter vocoders because most neural vocoders are data-driven.

To control the pitch of a speech waveform, acoustic features including $f_o$ are extracted from the original speech waveform, and the $f_o$-converted speech waveform is generated by scaling the $f_o$ values during the inference process. In the case of neural vocoders, the synthesis quality deteriorates when the input $f_o$ is not included in the range of the training data. Several approaches have been proposed to solve this problem [33], [34], [35], [36], [37], [38], [39]. In contrast to AR models [33], [34], [35], non-AR models [36], [37], [38], [39] can realize real-time inference. The neural source filter [36] introduces nonlinear filtering and dilated convolutional layers for parametrically generated source excitation signals corresponding to $f_o$ by source-filter modeling [40]. A method in [41], HiNet [42] and the neural homomorphic vocoder [43] introduce trainable linear-time-variant filters for impulse trains corresponding to $f_o$ and noise; these are based on mean squared error-based training in the time and frequency domains, GANs [22] and differentiable DSP [44]. Quasi-Periodic WaveNet (QPNet) [33] and

Manuscript received 15 July 2022; revised 4 January 2023 and 19 March 2023; accepted 30 April 2023. Date of publication 10 May 2023; date of current version 19 May 2023. This work was performed while Keisuke Matsubara was interning at NICT. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Kai Yu. *(Corresponding author: Takuma Okamoto.)*

Keisuke Matsubara, Ryoichi Takashima, and Tetsuya Takiguchi are with the Graduate School of System Informatics, Kobe University, Kobe 657–8501, Japan (e-mail: kmatsubara@stu.kobe-u.ac.jp; rtakashima@port.kobe-u.ac.jp; takigu@kobe-u.ac.jp).

Takuma Okamoto and Hisashi Kawai are with the National Institute of Information and Communications Technology, Kyoto 619–0289, Japan (e-mail: okamoto@nict.go.jp; hisashi.kawai@nict.go.jp).

Tomoki Toda is with the Information Technology Center, Nagoya University, Nagoya 464–8601, Japan (e-mail: tomoki@icts.nagoya-u.ac.jp).

Digital Object Identifier 10.1109/TASLP.2023.3275032

Quasi-Periodic Parallel WaveGAN (QPPWG) [37] introduce pitch-dependent dilated convolutional neural networks (PDCNNs), which flexibly change the dilation size of the dilated convolution kernel in response to $f_{\mathrm{o}}$ fluctuations. The unified source-filter GAN (uSFGAN) [38] improves QPPWG by introducing source-filter modeling that explicitly separates the generation of excitation signals from the filtering. The explicit input of the excitation signal into a generator based on source-filter modeling is highly effective for improving the control accuracy of $f_{\mathrm{o}}$. PeriodNet [39] uses sinusoidal waves and white noise as excitation signals to explicitly separate the generation of periodic waveforms from that of aperiodic waveforms. This approach is particularly effective for singing voice synthesis because the large $f_{\mathrm{o}}$ fluctuations contained in the singing voice can be efficiently captured by the input of the excitation signals. However, it cannot realize high-fidelity synthesis of normal speech [45]. Although these approaches have achieved high control accuracy of $f_{\mathrm{o}}$, the synthesis quality tends to be lower than that of the purely data-driven vocoders, such as HiFi-GAN. Additionally, most of them consist of very large convolution layers and require a high-end GPU for real-time synthesis.

SR conversion, which can expand or compress speech waveforms while preserving the pitch of the sound, is traditionally realized by signal-processing-based approaches, such as waveform similarity overlap-add (WSOLA) [46], time domain pitch synchronous overlap-add (TDPSOLA) [47], and source-filter vocoders [6], [7], [8], [9]. However, the synthesis quality of these models is not high. To improve synthesis quality for SR conversion, a neural-network-based approach with the multi-speaker AR WaveNet vocoder [48], which can be realized with time-compressed or stretched acoustic features by sinc interpolation-based resampling [49], outperforms conventional signal-processing-based models [50]. However, the AR WaveNet vocoder, even using a GPU, cannot realize real-time synthesis. Additionally, the synthesis quality for the slow-SR condition is particularly low, compared with that for the normal- and fast-SR conditions, because speech waveforms for slow speech are rarely included in training data. ScalerGAN, to perform real-time neural SR conversion with HiFi-GAN, has recently been proposed [51]. In ScalerGAN, input mel-spectrograms are non-uniformly compressed or stretched by a GAN, and SR-converted speech waveforms are synthesized by a multi-speaker HiFi-GAN generator with non-uniformly compressed or stretched features. In contrast to conventional neural vocoders, ScalerGAN cannot control $f_{\mathrm{o}}$ because mel-spectrograms are used as acoustic features.

In this article, we propose Harmonic-Net and Harmonic-Net+, which are real-time multi-speaker neural speech waveform generative models based on HiFi-GAN. They realize fast and high-quality SS on CPUs while preserving the controllability of $f_{\mathrm{o}}$ and SR. In the design of these models, we introduce two main extensions to the HiFi-GAN generator. First, we propose an excitation signal network with downsampling layers, which hierarchically receives multi-channel excitation signals for harmonic waves corresponding to $f_{\mathrm{o}}$, whereas the conventional methods only receive single channel sine waves or pulse trains. This explicit input of excitation signals is expected to improve synthesis quality when using scaled $f_{\mathrm{o}}$ input, similarly to PeriodNet. Second, we propose layerwise PDCNNs (LW-PDCNNs) for the upsampling-based HiFi-GAN generator, whereas the conventional PDCNNs are developed for CNN-based models and cannot be directly applied to the upsampling-based HiFi-GAN generator. As noted above, PDCNNs can incorporate $f_{\mathrm{o}}$ fluctuations into their model structure and we expect that the proposed method can further improve the synthesis quality when using scaled $f_{\mathrm{o}}$ input. Furthermore, the proposed models are expected to be used for high-quality and real-time neural SR conversion. This is because, although the $f_{\mathrm{o}}$ value itself is resampled along the time axis, the input excitation signals corresponding to $f_{\mathrm{o}}$ are not resampled but the number of repetitions of the input excitation signals is changed, and the direct input of the excitation signals can assist in the synthesis of SR-converted speech waveforms. The proposed models are expected to be particularly effective for slow-SR conversion, which corresponds to increasing the number of repetitions of the input excitation signals. The results of experiments for unseen speaker synthesis, full-band singing voice synthesis, and single-speaker TTS demonstrate that the proposed models with multi-channel harmonic waves can realize higher synthesis quality than conventional methods in all (normal, $f_{\mathrm{o}}$-conversion, and SR-conversion) conditions. The contributions of this article are as follows:

- An excitation signal network with downsampling layers, which hierarchically receives multi-channel excitation signals for harmonic waves corresponding to $f_{\mathrm{o}}$, is proposed to improve the $f_{\mathrm{o}}$ controllability for HiFi-GAN-based real-time neural vocoder on CPUs.
- LW-PDCNNs for the upsampling-based HiFi-GAN generator is additionally proposed to further improve the $f_{\mathrm{o}}$ controllability.
- We show that the proposed methods with explicit input of excitation signals are also effective for SR conversion.

The rest of this article is organized as follows. HiFi-GAN [21] is briefly introduced in Section II. Harmonic-Net and Harmonic-Net+ are then proposed in Section III. Section IV describes experiments to compare Harmonic-Net and Harmonic-Net+ with several conventional methods: WORLD [8], WaveNet [50], HiFi-GAN [21], uSFGAN [38], and PeriodNet [39]. Finally, conclusions are presented in Section V.

## II. Related Work: HiFi-GAN

HiFi-GAN is a GAN-based neural vocoder that consists of a high-speed generator with transposed convolution layers and two sophisticated discriminators. Similarly to Tacotron 2 [2], band-limited mel-spectrograms are used as input acoustic features. In contrast to typical neural vocoders with white noise input [13], [14], [15], [17], [18], HiFi-GAN directly upsamples input acoustic features and synthesizes speech waveforms without white noise input, similarly to MelGAN [16], [19]. The main component of the generator is an upsampling network with a few transposed convolution layers. The generator upsamples input acoustic features through transposed convolutions until the length of the output sequence matches the temporal resolution of the speech waveforms. After each transposed

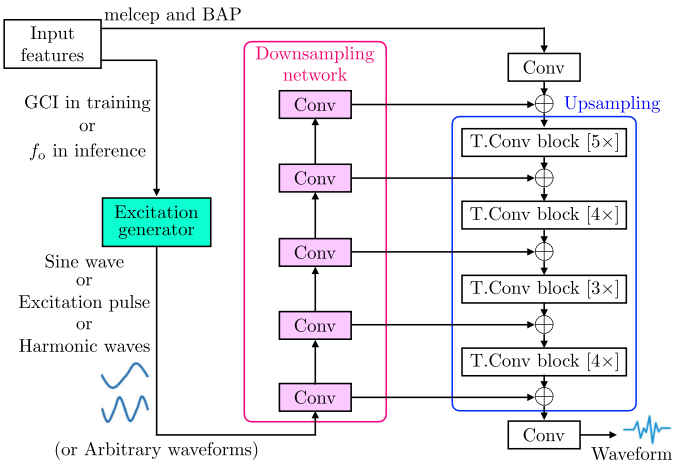



Fig. 1. Architecture of Harmonic-Net generator with only an excitation signal network.

convolution, multi-receptive field fusion (MRF) is performed. MRF is the aggregation of convolution layers with various receptive fields for efficiently capturing the various frequency components in speech waveforms. In contrast to typical neural vocoders, which have a large number of convolution layers, the HiFi-GAN generator achieves high-speed synthesis with only a few convolution layers, similarly to MelGAN [16], [19]. The two discriminators are a multi-scale discriminator (MSD) and a multi-period discriminator (MPD). The architecture of MSD, which was used in MelGAN [16], [19], 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 two-dimensional features 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 efficiently capture the various frequency components. In addition to adversarial training, mel-spectrogram loss [21] and feature matching loss [16], [52] (which is defined as the distance between the intermediate features of the discriminators) are used in MSD and MPD to train the generator effectively.

By using these sophisticated neural network models, HiFi-GAN has achieved high-quality and real-time SS, even for unseen speaker synthesis, using only a single CPU [21], and the synthesis speed can be further increased by using multiple CPU cores [28], [53]. HiFi-GAN can also be driven by acoustic features for source-filter vocoders instead of band-limited mel-spectrograms [53], such as features based on LPCNet [11].

## III. Proposed Models

### A. Harmonic-Net With Excitation Signal Network

Fig. 1 shows the architecture of the proposed Harmonic-Net generator with only an excitation signal network; this generator is a simple extension of the HiFi-GAN generator. The Harmonic-Net generator takes acoustic features that consist of mel-cepstra (melcep), binary-coded aperiodicity components (BAP), voiced/unvoiced vector (VUV), and glottal closure instants (GCI) in the training or $f_o$ value in the synthesis; these are used for controlling $f_o$ instead of mel-spectrograms. The upsampling network receives only melcep and BAP,[1] and upsamples them using transposed convolution blocks until the temporal resolution of the output sequence matches that of the audio waveform. The transposed convolution blocks (T.Conv blocks) consist of a transposed convolution layer and an MRF module, as used in HiFi-GAN.

In the training, the excitation generator receives GCI and generates the sinusoidal waves corresponding to the locations of GCI, similarly to PeriodNet [39]. These sinusoidal waves are then input to the downsampling layers. The purpose of using GCI is to input excitation signals that are in phase with the target speech waveforms, similarly to PeriodNet [39]. In the inference, we use $f_o$ instead of GCI, as proposed in [39]. The explicit input of $f_o$ features as time-domain waveform signals is expected to reduce the burden of modeling vocal fold vibration and improve the controllability of $f_o$. The downsampling network consists of five convolution layers with the same kernel size and stride as the transposed convolution layer of each T.Conv block (Fig. 1). In each layer, the excitation signals are converted to an intermediate feature whose temporal resolution corresponds to that of the output of the T.Conv blocks, and these features are added together.[2] This process was inspired by Fre-GAN [26], which uses a hierarchical output structure in the generator to maintain the consistency of the output audio at multiple resolutions. By introducing this process, we expect efficient training to be performed so that the generated speech waveforms maintain consistency with the excitation signals at various temporal resolutions.

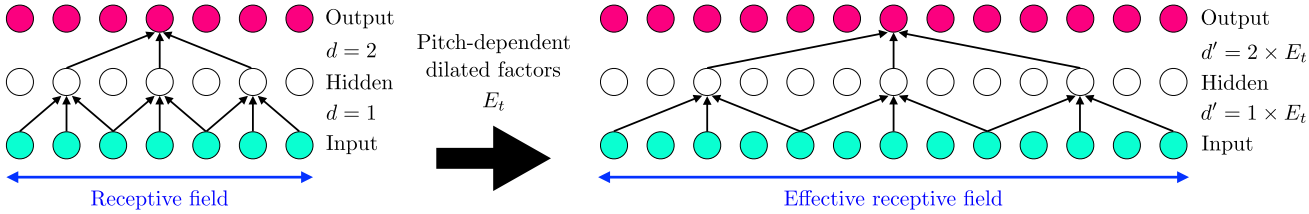
### B. Generation of Multi-Channel Excitation Signals Including Harmonic Wave Components

In most previous studies that introduced source-filter modeling to neural vocoders, sinusoidal waves or summational signals of harmonic components with fixed weighting values were used as excitation signals [36], [38], [39]. In contrast, we propose multi-channel excitation signals up to $I$th harmonic waves that consist of sinusoidal waves corresponding to GCI or to $f_o$ $(i = 1)$ and their harmonic components $(i = 2, 3, \ldots, I)$, where $i$ is the magnification rate of the harmonic signal. Then, the input channel of the proposed excitation signal network is $I$, and each $i$th harmonic wave is input to each $i$th channel of five trainable convolutional layers in a data-driven manner instead of using fixed weighting values [36]. Therefore, the proposed excitation signal network with multi-channel harmonic waves is expected to synthesize and control the harmonic components of output speech waveforms more accurately than the previous methods using simple sinusoidal waves [38], [39] or summational signals of harmonic components with fixed weighting values [36].

GCI can be defined as a sequence that collects the time points at which the most basic phases of the speech waveform

[1] Although $f_o$ values were also input to the upsampling network in preliminary experiments, they caused the controllability of $f_o$ to degrade. Therefore, in the final design, only melcep and BAP are input to the upsampling network.

[2] Although a downsampling network with one or two convolutional layers was initially investigated, the synthesis quality could not reach that achieved with five convolutional layers (Fig. 1).




(a) Dilated convolutional neural network (b) Pitch-dependent dilated convolutional neural network

Fig. 2. (a) Non-causal dilated convolutional neural network and (b) Non-causal pitch-dependent dilated convolutional neural network.

match. Let $\boldsymbol{g} = [g_1, \ldots, g_q, \ldots, g_Q]$ be the sequence obtained by multiplying each value of the GCI sequence by the sampling frequency $f_\mathrm{s}$. $\boldsymbol{g}$ is the index sequence of time points at which the phases of natural speech match. $\boldsymbol{v} = [v_1, \ldots, v_t, \ldots, v_T]$ is a one-hot vector sequence that indicates voiced/unvoiced speech at time step $t$. In the training, the $i$th harmonic excitation wave sequence $e_{t,i}$ is generated as follows:

$$l = \arg \min_{\{q: g_q < t\}} (t - g_q), \tag{1}$$

$$e_{t,i} = \begin{cases} \sin\left(2\pi i \dfrac{t - g_l}{g_{l+1} - g_l} + \phi\right) & v_t = 1, \\ 0 & v_t = 0, \end{cases} \tag{2}$$

where $\phi$ denotes the initial phase of the excitation signal at $t$. Conversely, in the inference, $e_{t,i}$ is generated using the $f_\mathrm{o}$ value sequence $\boldsymbol{f}_\mathrm{o} = [f_{\mathrm{o},1}, \ldots, f_{\mathrm{o},t}, \ldots, f_{\mathrm{o},T}]$ as follows:

$$e_{t,i} = \begin{cases} \sin\left(\sum_{t'=1}^{t} 2\pi i \frac{f_{\mathrm{o},t'}}{f_\mathrm{s}}\right) & f_{\mathrm{o},t} > 0, \\ 0 & f_{\mathrm{o},t} = 0. \end{cases} \tag{3}$$

HiFi-GAN-based $f_\mathrm{o}$-controllable SS can be realized by controlling $f_{\mathrm{o},l}$ in (3).

Additionally, arbitrary $I$ channel waveform signals can be input to the excitation signal network with input channel $I$. For example, single channel excitation pulse sequences used in source-filter vocoders or single channel waveforms synthesized by DSP-based source-filter vocoders (e.g. WORLD) can be input with input channel $I = 1$. These input waveforms were investigated in the experiments reported in Section IV.

### C. Pitch-Dependent Dilated Convolution Network

As the results of experiments conducted Section IV shown in Fig. 5, the Harmonic-Net generator with only an excitation signal network could realize high-fidelity SS while preserving the controllability of $f_\mathrm{o}$ for male speakers. However, the synthesis quality for female speakers with higher $f_\mathrm{o}$ conditions was degraded because the controlled $f_\mathrm{o}$ value was outside the range of the training data. However, collecting higher $f_\mathrm{o}$ speech data to extend the $f_\mathrm{o}$ range of the training data is costly and impractical due to the burden on the speakers compared to collecting normal $f_\mathrm{o}$ speech data. Therefore, investigating neural speech waveform generative models to extrapolate $f_\mathrm{o}$ component outside the range of the training data is important. To further improve the synthesis quality and controllability of $f_\mathrm{o}$ for female speaker synthesis, whose $f_\mathrm{o}$ range is quite large, we introduce PDCNNs, which were previously used in QPPWG and uSFGAN. The causal PDCNN was initially proposed for use in the AR model QPNet [33], as a sophisticated network to directly reflect fluctuations of $f_\mathrm{o}$ in the model structure. The non-causal PDCNN was subsequently proposed for use in non-AR models, such as QPPWG [37], and the synthesis quality was improved by combining it with source-filter modeling in uSFGAN [38]. Because of its higher synthesis quality, we incorporate the non-causal PDCNN into the Harmonic-Net generator.

Fig. 2 shows the architectures of the non-causal dilated convolutional neural network (DCNN) and non-causal PDCNN, in which DCNN has gaps between input samples; the length of each gap is a predefined hyperparameter called the dilation size. The non-causal DCNN can be formulated as follows:

$$\boldsymbol{y}_t^{(\mathrm{o})} = \sum_{k=0}^{K} \left( \boldsymbol{W}^{(k)} \otimes \boldsymbol{y}^{(\mathrm{i})}_{t-\left(\frac{K}{2}-k\right)d} \right), \tag{4}$$

where $\boldsymbol{y}_t^{(\mathrm{o})}$ is a 1D vector of the DCNN at sample $t$, $\boldsymbol{y}_t^{(\mathrm{i})}$ is a 1D vector of the DCNN input at sample $t$, $\otimes$ is the Hadamard product operator, $d$ is the dilation size and $K$ is the kernel size. $\boldsymbol{W}^{(k)}$ is a 1D vector of the $k$th trainable $1 \times 1$ convolution filter. Although $d$ is the time-invariant constant in the DCNN, the PDCNN extends $d$ to the $f_\mathrm{o}$-dependent dilation size $d'$. Specifically, the $f_\mathrm{o}$-dependent dilated factor $E_t$ and $d'$ are formulated as

$$E_t = \lfloor f_\mathrm{s} / (f_{\mathrm{o},t} \times a) \rfloor, \tag{5}$$

$$d' = E_t \times d, \tag{6}$$

where $\lfloor \cdot \rfloor$ is the floor function, $f_{\mathrm{o},t}$ is the $f_\mathrm{o}$ at time step $t$, and $a$ is the hyperparameter (named dense factor) that specifies the number of samples (in one cycle) that are taken as the inputs of a PDCNN. The model parameters are kept unchanged because only the dilation size $d'$ changes according to $f_{\mathrm{o},t}$ and the same filter $\boldsymbol{W}^{(k)}$ is used as shown in Fig. 2.

### D. Harmonic-Net+ With Excitation Signal Network and Layerwise Pitch-Dependent Dilated Convolution Neural Networks

The PDCNN has been proposed for CNN-based neural vocoders, such as Parallel WaveGAN [15], in which the temporal resolution is the sampling frequency of the speech waveforms. However, the PDCNN cannot be directly applied to the HiFi-GAN generator. In HiFi-GAN, the temporal resolution gradually increases as the number of transposed convolutions increases. Therefore, the PDCNN needs to be designed so that the density

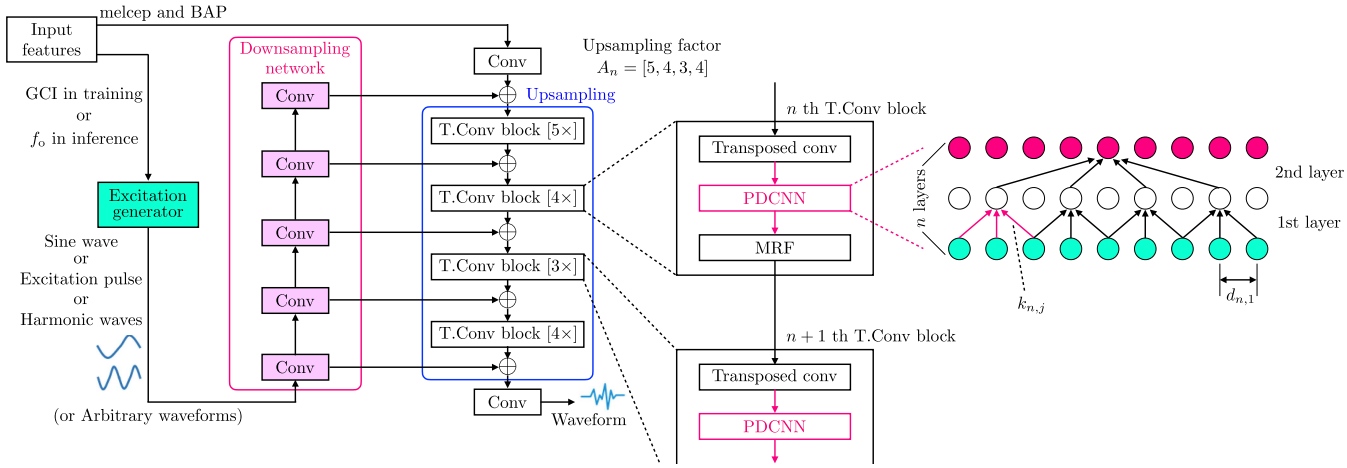



Fig. 3. Network architecture of Harmonic-Net+ generator with layerwise PDCNNs.

of the convolution networks increase in the same manner and all of them have the same receptive field. To introduce PDCNNs into the upsampling-based HiFi-GAN generator, we propose LW-PDCNNs. Fig. 3 shows the architecture of the proposed Harmonic-Net+ generator with LW-PDCNNs. Specifically, it is designed such that the PDCNN of the first T.Conv block has only one layer, and the number of layers of each PDCNN increases as the number of stages of the T.Conv block increases. This means that the $n$th T.Conv block has a PDCNN that consists of $n$ layers. The temporal resolution of the $n$th T.Conv block $F_{s,n}$, the kernel size of the $j$th layer of PDCNN $k_{n,j}$, and the dilation size $d_{n,j}$ are defined as follows:

$$f_{\mathrm{s},n} = f_{\mathrm{s},0} \prod_{m=1}^{n} A_m, \tag{7}$$

$$k_{n,j} = \begin{cases} A_j & j \neq 1, \\ 3 & j = 1, \end{cases} \tag{8}$$

$$d_{n,j} = \begin{cases} d_{n,2} \prod_{m=2}^{j-1} A_m & j > 2, \\ 2d_{n,1} & j = 2, \\ \lfloor f_{\mathrm{s},n} / (f_{\mathrm{o},t} \times a) \rfloor & j = 1, \end{cases} \tag{9}$$

where $f_{\mathrm{s},0}$ is the temporal resolution of the acoustic features and $A_n$ is the upsampling factor of the $n$th T.Conv block. We set $d_{n,2} = 2d_{n,1}$ in accordance with the results of preliminary experiments. The kernel size and dilation size of the first layer of the PDCNN are designed in the same manner as those of the conventional PDCNN. In the second and subsequent layers, these values are defined according to the upsampling factor.

The proposed LW-PDCNNs enable the upsampling-based HiFi-GAN generator to incorporate $f_{\mathrm{o}}$ fluctuations into its model structure, and the synthesis quality when using scaled $f_{\mathrm{o}}$ input is expected to be further improved. HiFi-GAN with only LW-PDCNNs but without an excitation signal network was also investigated in preliminary experiments. However, it could not outperform either Harmonic-Net or Harmonic-Net+ for $f_{\mathrm{o}}$ conversion conditions. Therefore, the proposed excitation signal network is important for $f_{\mathrm{o}}$ conversion.

### E. Speech-Rate Conversion

To control SR, the acoustic features (melcep, BAP, and $f_{\mathrm{o}}$ including VUV) extracted from target speech waveforms are resampled with a speech rate of $r$ along the time direction as proposed in [50]. The excitation signals are then generated from the resampled $f_{\mathrm{o,resampled}} = [f_{\mathrm{o},1}, \ldots, f_{\mathrm{o},t}, \ldots, f_{\mathrm{o},rT}]$ by the excitation generator (3) and input to the Harmonic-Net and Harmonic-Net+ generators. Although melcep and BAP are smoothed by resampling, the excitation signals are not resampled but the number of repetitions of the input excitation signal is changed. Therefore, the direct input of the excitation signals is also expected to improve the synthesis quality for SR conversion, compared with the quality of the conventional method that uses resampled mel-spectrograms.

## IV. Experiments

### A. Experimental Setup

We conducted three experiments to evaluate Harmonic-Net and Harmonic-Net+ in comparison with several conventional methods: WORLD [8] as a reference, HiFi-GAN [21], PeriodNet [39], WaveNet [50], and uSFGAN [38].[3] These experiments were conducted using a multi-speaker normal speech dataset, a single-speaker full-band singing voice dataset, and a single-speaker normal speech dataset for TTS. For $f_{\mathrm{o}}$ conversion with low and high $f_{\mathrm{o}}$, $0.5 \times f_{\mathrm{o}}$ and $1.5 \times f_{\mathrm{o}}$ were used, as in [37]. For SR conversion with fast and slow SRs, $0.8 \times T$ and $1.5 \times T$ were used, as in [50].[4] All the neural network models were implemented in PyTorch [54] and trained on an NVIDIA Tesla V100 GPU. Some of the speech samples used in the experiments are available online.[5]

[3]WSOLA [46] was not included in the experiments because it could not outperform WaveNet, Harmonic-Net and Harmonic-Net+ in the SR-conversion condition in preliminary experiments as [50].

[4]Although only uniform resampling of acoustic features was performed for SR conversion, non-uniform resampling will be investigated in future work by introducing the ScalerGAN framework [51] into the proposed method. This will further improve the quality of SR conversion.

[5]https://ast-astrec.nict.go.jp/demo_samples/harmonic-net/index.html

*1) Dataset:* The following open-source corpora were used in all the experiments to ensure reproducibility. For unseen speaker synthesis with multi-speaker models, we used the JVS corpus [55], a Japanese multi-speaker corpus with $f_s = 24$ kHz. In the training, we used 12,447 utterances by 96 speakers (jvs005 to jvs100). For evaluation, we used 120 non-parallel utterances by four speakers (jvs001 to jvs004), which were not included in the training. For full-band singing voice synthesis, we used 50 acapella songs (about 1 h) by a Japanese female singer, from the Tohoku Kiritan corpus [56] with $f_s = 96$ kHz. We then downsampled the audio to 48 kHz and clipped it into segments of appropriate length. We separated all 50 songs into phrases by using the provided labels and used two songs (05.wav and 30.wav), each of which includes 10 phrases, for evaluation; the remaining 48 songs, constructed from 376 phrases, were used for training, as in [45]. For single-speaker TTS, we used 7,497 utterances from the JSUT corpus [55], a Japanese single-speaker corpus, downsampled to $f_s = 24$ kHz, to train neural vocoders, and used the remaining 50 utterances (Basic5000-0001 to Basic5000-0050) and 150 utterances (Basic5000-0051 to Basic5000-0200) for evaluation and validation sets. To train a neural TTS model, we used 4,800 sentences (Basic5000-0201 to Basic5000-5000) from JSUT for which HTS-style context labels (based on manual annotation) were available,[6] as in [45].

*2) Neural Vocoders:* The network architecture of our implementation of HiFi-GAN was the same as that of the official implementation [21];[7] we used the V1 model in which the number of initial channel is 512 [21]. As input features, we used 50-dimensional melcep coefficients with warping coefficient $\alpha = 0.455$, three-dimensional BAP, and log-scaled continuous $f_o$ for unseen speaker synthesis and single-speaker TTS with $f_s = 24$ kHz. For full-band singing voice synthesis with $f_s = 48$ kHz, we used 50-dimensional melcep coefficients with warping coefficient $\alpha = 0.55$, five-dimensional BAP, and log-scaled continuous $f_o$, as in [45]. These features were extracted by cheaptrick [57], D4C [58], and Harvest [59] (based on WORLD [8]), respectively. 50-dimensional melcep coefficients, which are not affected by $f_o$, can be extracted from smooth vocal tract spectra analyzed by cheaptrick, while those based on the short time Fourier transform are affected by $f_o$. The window and shift lengths were set to 42.7 ms and 10 ms, respectively. Additionally, we used HiFi-GAN models with 80-dimensional log-mel spectrograms, HiFi-GAN (melspc), as used in [21], to compare the input features.[8] The window and shift lengths were also set to 42.7 ms and 10 ms: the same as those of the WORLD features. Although the original HiFi-GAN used 256-fold upsampling [21], we applied 240- or 480-fold upsampling to obtain a resolution of 24 kHz or 48 kHz from input features with a frame shift of 10 ms. 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 24-kHz synthesis, as in [53], and set the upsampling rates to $(10, 6, 2, 2, 2)$ and the kernel sizes to $(20, 12, 4, 4, 4)$ for 48-kHz synthesis.

The network structure of the PeriodNet was the same as that of the non-AR series model in [39]. Its implementation was based on that of Parallel WaveGAN [15],[9] as in [45], and we added two generators (to generate periodic and aperiodic signals) and discriminators that operate at multiple sampling frequencies. As input features, we used WORLD features, as used in HiFi-GAN. The network architecture of the uSFGAN was the same as that of the official implementation [38].[10] As input acoustic features, we used the same WORLD features as used in HiFi-GAN. The network structure of the WaveNet vocoder was based on [60] with an additional GRU unit for multi-speaker training, as in [61]. As input features for WaveNet, we also used WORLD features as used in HiFi-GAN. We also applied time-invariant noise shaping [60] to suppress the perceptual noise components caused by the prediction error; 35-dimensional melcep were used and a parameter to control noise energy in the formant regions was set to 0.5 for noise shaping, as in [60].

The network architecture of the proposed Harmonic-Net generator was based on the official implementation of HiFi-GAN [21], with the addition of the proposed excitation signal network. For the Harmonic-Net+ generator, the LW-PDCNNs were based on the official PyTorch implementation of QP-PWG [37].[11] The dense factor $a$ in (9) was set to 4.0 for the Harmonic-Net+ generator. The configuration of other modules, such as discriminators, was the same as that of the corresponding components of HiFi-GAN. As input features, we used 50-dimensional melcep coefficients, three-dimensional BAP, and GCI extracted by REAPER.[12] In the inference, we used linear $f_o$ extracted by Harvest [59] instead of GCI. We investigated four types of excitation signals: single-channel sine wave (*sine*) with $I = 1$, pulse sequence (*pulse*) with $I = 1$, speech waveform synthesized by WORLD vocoder (*world*) with $I = 1$ as a reference, and harmonic waves up to the fifth harmonic (*harm*) with $I = 5$,[13] as explained in Section III-B. Although the excitation signal network using 1-channel convolutional layers for summational signals of 5 harmonic components with fixed weighting values as [36] combined with LW-PDCNNs was initially investigated, it was not included in the experiments because the results of preliminary experiments indicated that it underperformed the proposed excitation signal network using 5-channel convolutional layers for 5-channel harmonic waves combined with LW-PDCNNs, especially for high $f_o$ conversion condition.

*3) Text-to-Speech:* As an acoustic model for TTS, we used a FastSpeech-based acoustic model [62] with full-context label input for Japanese, which was complemented by ESPnet-TTS [63]. We used simple 47-dimensional vectors constructed

[6]https://github.com/sarulab-speech/jsut-label

[7]https://github.com/jik876/hifi-gan

[8]Although we have initially investigated to introduce $f_o$ and mel-spectrograms as input for controlling $f_o$, it could not control $f_o$ accurately because mel-spectrograms include $f_o$ components.

[9]https://github.com/kan-bayashi/ParallelWaveGAN

[10]https://github.com/chomeyama/UnifiedSourceFilterGAN

[11]https://github.com/bigpon/QPPWG

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

[13]The number of harmonic waves was decided in accordance with the results of preliminary experiments

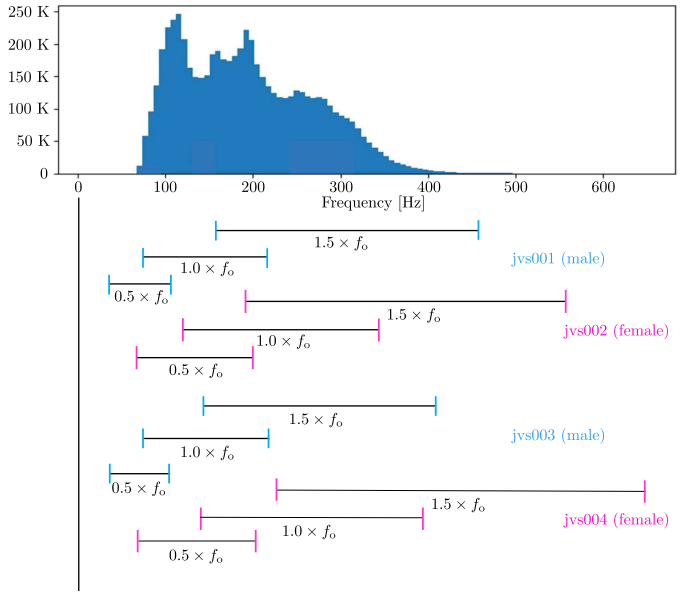



Fig. 4. Distribution of $f_o$ in training dataset and $f_o$ ranges of evaluation speakers for $0.5 \times f_o$, $1.0 \times f_o$, and $1.5 \times f_o$.

from 38-dimensional phoneme one-hot vectors and nine-dimensional accentual label vectors, as in [45]. For TTS, we fine-tuned HiFi-GAN-based neural vocoders using acoustic features (80-dimensional mel-spectrograms or 55-dimensional WORLD features) estimated by the trained acoustic models, as in [21]. The HiFi-GAN-based models were trained with 1,000,000 steps and fine-tuned with 200,000 steps.

*4) Objective Evaluation Criteria:* As objective evaluation criteria, we used signal-to-noise ratio (SNR), spectral distortion (SD), mel-cepstral distortion (MCD), root-mean-square error of linear $f_o$ ($f_o$-RMSE), and real-time factor (RTF) in the inference. To measure RTFs, we used an Intel Xeon 6152 CPU (with one core). To calculate $f_o$-RMSE with $f_o$-scaled conditions, we applied constant scaling to the $f_o$ extracted from natural speech and used it as the reference.

*5) Histogram of $f_o$ for Multi-Speaker Models:* Fig. 4 shows a histogram of the $f_o$ included in the training data, and the $f_o$ ranges of the evaluation set for multi-speaker models. The figure shows that $f_o$ values outside the training range are included for $0.5 \times f_o$ in male speech and $1.5 \times f_o$ in female speech.

### B. Evaluation of Unseen Speaker Synthesis With Multi-Speaker Models

Table I shows the results of the objective evaluations for unseen speaker synthesis in the normal condition ($1.0 \times f_o$ and $1.0 \times T$).[14] The uSFGAN showed the best performance with respect to SD, MCD and $f_o$-RMSE, and both Harmonic-Net and Harmonic-Net+ performed slightly better than HiFi-GAN. With respect to SNR, both Harmonic-Net and Harmonic-Net+ models except for *sine* excitation outperformed the other methods. With respect to RTF, all the proposed methods were able to perform real-time synthesis using a CPU, in contrast to WaveNet and uSFGAN. Although the synthesis speed of Harmonic-GAN+ with LW-PDCNNs was low because of the use of LW-PDCNNs, in comparison with the HiFi-GAN-based models without LW-PDCNNs, it could realize real-time inference with a CPU. PDCNN takes more computation time than normal CNN because it adaptively calculates the dilation size of CNN according to the input $f_o$. The detailed implementation of PDCNN in PyTorch can be found in the source code of QPPWG. Although the inference speed of Harmonic-GAN+ with PDCNNs could be increased by making the number of initial channels small (e.g., 256 or 128), this impaired the synthesis quality in preliminary experiments. Therefore, in future work, it is necessary to develop a lightweight PDCNN optimized for the HiFi-GAN structure.

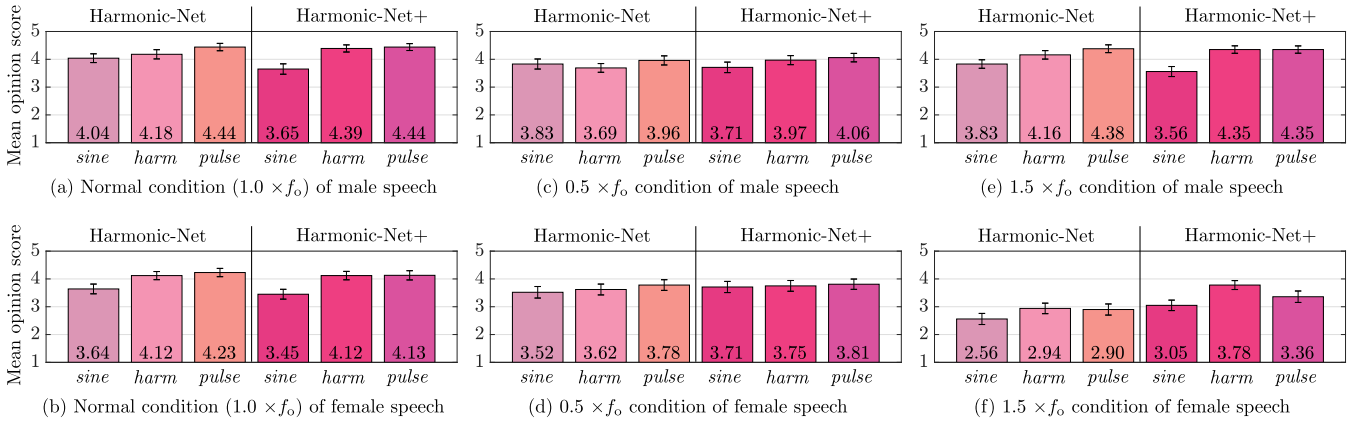
[14] As explained in Section I and [45], PeriodNet was not evaluated for normal SS because it cannot synthesize normal speech well.

TABLE I
RESULTS OF OBJECTIVE EVALUATIONS FOR UNSEEN SPEAKER SYNTHESIS IN NORMAL CONDITION

| Model | excitation | SNR [dB] | SD [dB] | MCD [dB] | $f_o$-RMSE [Hz] | RTF |
|---|---|---|---|---|---|---|
| WORLD (reference) | - | −1.27 | 8.34 | 3.63 | 20.3 | 0.10 |
| uSFGAN | *sine* | 0.87 | **8.00** | **3.46** | **21.9** | 3.63 |
| WaveNet | - | −0.42 | 8.82 | 3.82 | 26.8 | 3451 |
| HiFi-GAN | - | 0.18 | 8.62 | 3.81 | 26.6 | **0.38** |
| HiFi-GAN (melspc) | - | 1.09 | 8.66 | 4.43 | 25.7 | **0.38** |
| Harmonic-Net | *sine* | 0.41 | 8.45 | 3.61 | 27.3 | 0.39 |
| | *pulse* | 1.67 | 8.05 | 3.47 | 24.4 | 0.39 |
| | *world* | **1.71** | **8.00** | 3.52 | 24.8 | 0.51 |
| | *harm* | 1.33 | 8.19 | 3.52 | 22.9 | 0.40 |
| Harmonic-Net+ | *sine* | 0.33 | 8.48 | 3.81 | 25.9 | 0.66 |
| | *pulse* | 1.70 | 8.01 | 3.73 | 24.6 | 0.65 |
| | *world* | 1.62 | 8.06 | 3.72 | 24.5 | 0.75 |
| | *harm* | 1.35 | 8.09 | 3.56 | 23.2 | 0.66 |

TABLE II
RESULS OF OBJECTIVE EVALUATIONS FOR UNSEEN SPEAKER SYNTHESIS IN $f_o$-CONVERSION CONDITION

| Model | excitation | MCD (dB) | | $f_o$-RMSE (Hz) | |
|---|---|---|---|---|---|
| $0.5 \times f_o$ | | | | | |
| | | male | female | male | female |
| WORLD (reference) | - | 3.54 | 4.70 | 15.3 | 19.1 |
| uSFGAN | *sine* | 4.37 | 3.87 | **14.4** | 18.9 |
| HiFi-GAN | - | 4.45 | 3.94 | 20.7 | 20.2 |
| Harmonic-Net | *sine* | **3.92** | 3.79 | 29.2 | 20.1 |
| | *pulse* | 4.01 | **3.67** | 20.6 | 18.5 |
| | *world* | 4.00 | 3.76 | 21.4 | 17.9 |
| | *harm* | 4.08 | 3.72 | 25.3 | 19.1 |
| Harmonic-Net+ | *sine* | 4.28 | 3.93 | 22.3 | 18.3 |
| | *pulse* | 4.34 | 3.94 | 19.5 | 17.8 |
| | *world* | 4.38 | 3.98 | 18.6 | 17.6 |
| | *harm* | 4.00 | 3.79 | 20.6 | **17.5** |
| $1.5 \times f_o$ | | | | | |
| | | male | female | male | female |
| WORLD (reference) | - | 3.61 | 4.15 | 22.3 | 43.2 |
| uSFGAN | *sine* | **4.02** | **4.32** | 23.6 | **53.2** |
| HiFi-GAN | - | 4.30 | 5.10 | 31.5 | 66.3 |
| Harmonic-Net | *sine* | 4.15 | 4.63 | 29.8 | 58.9 |
| | *pulse* | 4.07 | 4.64 | 29.6 | 64.9 |
| | *world* | 4.04 | 4.71 | 27.2 | 58.9 |
| | *harm* | 4.13 | 4.73 | **22.6** | 53.9 |
| Harmonic-Net+ | *sine* | 4.38 | 4.83 | 25.7 | 54.3 |
| | *pulse* | 4.28 | 4.68 | 27.5 | 57.8 |
| | *world* | 4.23 | 4.72 | 30.2 | 54.8 |
| | *harm* | 4.03 | 4.55 | 23.4 | 53.9 |




(a) Normal condition ($1.0 \times f_o$) of male speech
(c) $0.5 \times f_o$ condition of male speech
(e) $1.5 \times f_o$ condition of male speech
(b) Normal condition ($1.0 \times f_o$) of female speech
(d) $0.5 \times f_o$ condition of female speech
(f) $1.5 \times f_o$ condition of female speech

Fig. 5. Results of MOS test for unseen speaker synthesis in normal and $f_o$-conversion conditions to compare excitation signals of Harmonic-Net and Harmonic-Net+. Confidence level of the error bars is 95%.

Table II shows the results of the objective evaluations for unseen speaker synthesis in the $f_o$-conversion condition. WaveNet was not included in the evaluations because a synthetic error occurred when scaling $f_o$. Overall, the proposed methods improved the controllability of $f_o$, compared with HiFi-GAN, particularly for the $1.5 \times f_o$ condition. The results show that the introduction of the excitation signal network contributed to the improvement of the controllability of $f_o$. Except for the $1.5 \times f_o$ condition for male speakers, the controllability of $f_o$ of the proposed methods was comparable to that of uSFGAN. Comparing the excitation signals, *harm* excitation achieved the best performance in most cases, particularly for the $1.5 \times f_o$ condition.

A mean opinion score (MOS) test with a five-point scale (5 for excellent, 4 for good, 3 for fair, 2 for poor, and 1 for bad) [64] was conducted to evaluate the subjective perceptual quality of the synthesized speech waveforms. First, Harmonic-Net and Harmonic-Net+ with *sine*, *harm* and *pulse* excitation signals were directly compared in $1.0 \times f_o$ for the normal condition, $0.5 \times f_o$ and $1.5 \times f_o$ for the $f_o$-conversion condition to compare the differences among the excitation signals. Twenty adult native Japanese speakers without hearing loss listened to the synthesized speech samples using headphones. Fig. 5 shows the results of the MOS test. Both Harmonic-Net and Harmonic-Net+ with *pulse* outperformed those with *sine* in all the $f_o$ conditions. Harmonic-Net+ with *harm* was comparable to both Harmonic-Net and Harmonic-Net+ with *pulse* except for $1.5 \times f_o$ condition of female speech. In $1.5 \times f_o$ condition of female speech, Harmonic-Net+ with *harm* significantly outperformed the other models. The results indicated that the proposed excitation signal network using trainable 5-channel convolutional layers for 5-channel harmonic waves combined with LW-PDCNNs was more suitable for high $f_o$ conversion condition than that using 1-channel convolutional layers for pulse trains, which is regarded as summational signals of infinite harmonic components with fixed weighting values, combined with LW-PDCNNs. As a result, only LW-PDCNNs with harmonic waves can extrapolate $f_o$ component outside the range of the training data while keeping the synthesis quality for high $f_o$ conversion condition. Therefore, Harmonic-Net+ with *harm* was introduced in the following MOS tests. Although the results in Fig. 5 indicate that Harmonic-Net with *harm* slightly lower than that with *pulse* except for $1.5 \times f_o$ condition of female speech, the results of preliminary experiments for full-band singing voice synthesis suggested that Harmonic-Net with *harm* significantly outperformed that with *pulse* for $1.5 \times f_o$ condition. The results indicated that the proposed excitation signal network using trainable 5-channel convolutional layers for 5-channel harmonic waves without LW-PDCNNs was also more suitable for high $f_o$ conversion condition in full-band singing voice synthesis than that using 1-channel convolutional layers for pulse trains without LW-PDCNNs.[15] Therefore, Harmonic-Net with *harm* was also introduced in the following MOS tests to match the type of excitation signal to Harmonic-Net+.

For unseen speaker synthesis with multi-speaker models, $1.0 \times f_o$ for the normal condition, $0.5 \times f_o$ and $1.5 \times f_o$ for the $f_o$-conversion condition, and $0.8 \times T$ and $1.5 \times T$ for the SR-conversion condition were evaluated.[16] Twenty adult native Japanese speakers without hearing loss also listened to the synthesized speech samples using headphones. There were 20 utterances for each model and each condition, where five sentences were randomly selected from each speaker of the evaluation set (jvs001, jvs002, jvs003, and jvs004). The total number of sentences evaluated by each listening subject was therefore 600 ($= 20$ utterances $\times (8 + 5 + 5 + 6 + 6)$ models). Fig. 6 shows the results of the MOS test for unseen speaker synthesis in the normal and $f_o$-conversion conditions. According to the results of the normal condition (Fig. 6(a) and (b)), Harmonic-Net+ achieved the highest quality synthesis for male speaker synthesis although it could not outperform WaveNet, which cannot realize real-time inference, for female speaker synthesis. Comparing the HiFi-GAN and Harmonic-Net models, both the excitation signal network and the LW-PDCNNs contributed to the improvement

[15] As described in Section IV-C, Harmonic-Net+ with LW-PDCNNs could not outperform Harmonic-Net without LW-PDCNNs for full-band singing voice synthesis due to the lack of training data.

[16] WaveNet was not evaluated in the $0.5 \times f_o$ and $1.5 \times f_o$ conditions because the $f_o$ controllability of AR WaveNet is lower than that of QPNet [33].

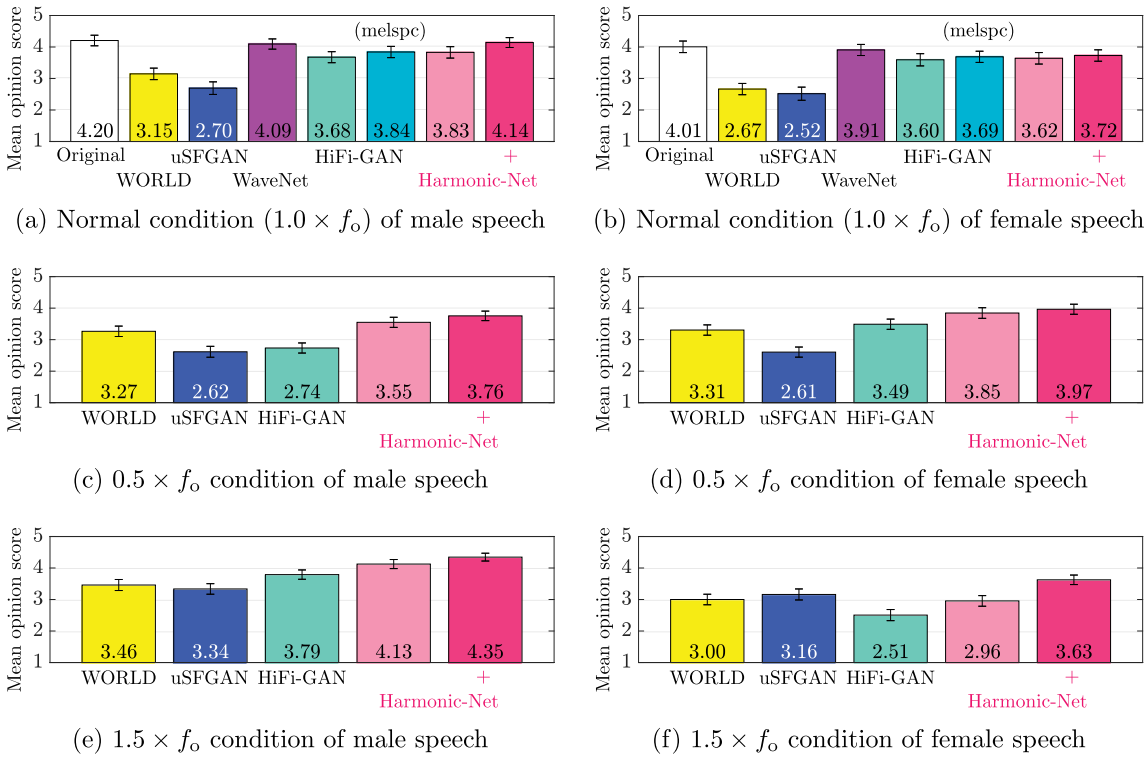



Fig. 6. Results of MOS test for unseen speaker synthesis in normal and $f_o$-conversion conditions. Confidence level of the error bars is 95%.

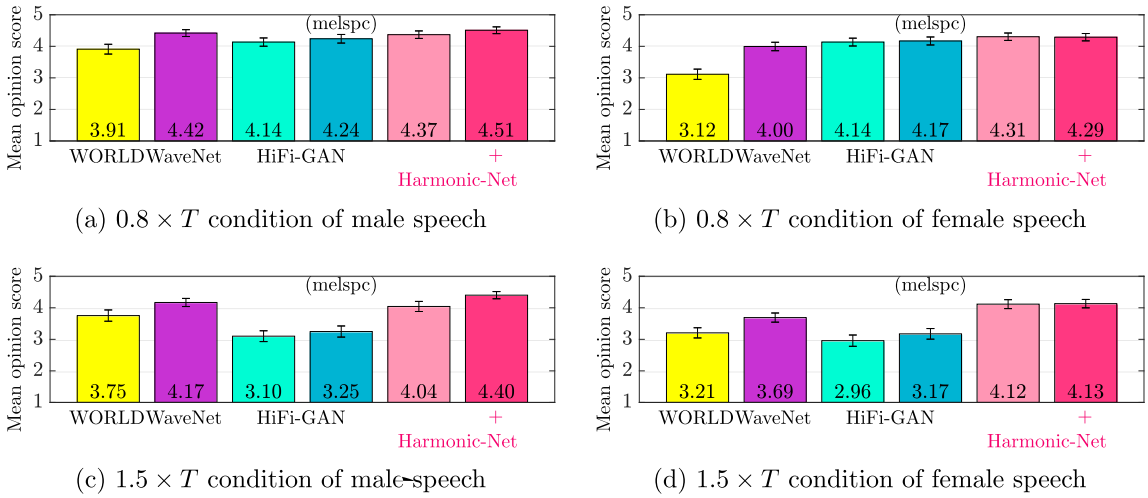



Fig. 7. Results of MOS test for unseen speaker synthesis in SR-conversion condition. Confidence level of the error bars is 95%.

of the synthesis quality for unseen speaker synthesis in the normal condition.

In the $f_o$-conversion condition (Fig. 6(c) to (f)), although the conventional HiFi-GAN and uSFGAN could not always outperform WORLD, Harmonic-Net with *harm* excitation signals improved the synthesis quality, compared with the conventional methods, except for the case of $1.5 \times f_o$ with female speech. This means that the excitation signal network worked effectively in low-$f_o$ synthesis and interpolation of high $f_o$, but it was unable to improve the extrapolation of high $f_o$. Conversely, Harmonic-Net+ with *harm* excitation signals further improved the synthesis quality and achieved the best score even in the case of $1.5 \times f_o$ with female speech.

Fig. 7 shows the results of the MOS test for unseen speaker synthesis in the SR-conversion condition. In the $0.8 \times T$ condition, although the conventional WaveNet and HiFi-GAN outperformed WORLD, Harmonic-Net+ achieved a significantly higher synthesis quality than the conventional methods. In the $1.5 \times T$ condition, the conventional HiFi-GAN models could not outperform WORLD, and WaveNet could not achieve high-quality synthesis for female speakers. Conversely, Harmonic-Net+ achieved the best performance of all the methods. As

TABLE III
RESULTS OF OBJECTIVE EVALUATIONS FOR FULL-BAND SINGING VOICE SYNTHESIS IN NORMAL CONDITION

| Model | excitation | SNR [dB] | SD [dB] | MCD [dB] | $f_o$-RMSE [Hz] | RTF |
|---|---|---|---|---|---|---|
| WORLD (reference) | - | 0.10 | 8.83 | 3.27 | 20.2 | 0.15 |
| uSFGAN | *sine* | 4.93 | 8.05 | 3.25 | 24.8 | 10.2 |
| WaveNet | - | 1.23 | 9.18 | 4.09 | 28.8 | 6322 |
| PeriodNet | *sine* | **6.40** | **8.02** | **2.96** | 24.3 | 20.4 |
| HiFi-GAN | - | 1.16 | 10.2 | 4.68 | 29.2 | **0.86** |
| HiFi-GAN (melspc) | - | 2.01 | 9.80 | 4.73 | 25.0 | **0.86** |
| Harmonic-Net | *sine* | 4.64 | 9.64 | 4.31 | 23.8 | 0.90 |
| | *pulse* | 4.63 | 9.63 | 4.48 | 23.9 | 0.91 |
| | *world* | 4.88 | 9.50 | 4.16 | 23.3 | 1.04 |
| | *harm* | 4.64 | 9.64 | 4.31 | 22.8 | 0.91 |
| Harmonic-Net+ | *sine* | 3.30 | 10.3 | 5.75 | **22.3** | 1.56 |
| | *pulse* | 3.67 | 10.1 | 5.44 | 24.6 | 1.55 |
| | *world* | 4.83 | 9.10 | 4.17 | 23.6 | 1.70 |
| | *harm* | 2.97 | 10.2 | 5.78 | 24.5 | 1.56 |

expected, artificial artifacts included in stretched features were avoided by introducing the excitation signal network because excitation signals are less susceptible to degradation resulting from interpolation. Consequently, Harmonic-Net+ outperformed the conventional models for unseen speaker synthesis in all (i.e., normal, $f_o$-conversion, and SR-conversion) conditions.

### C. Evaluation of Full-Band Singing Voice Synthesis

Table III shows the results of the objective evaluations for full-band singing voice synthesis in the normal condition ($1.0 \times f_o$ and $1.0 \times T$). With respect to SNR, SD and MCD, PeriodNet achieved the best score but the inference speed was insufficient for real-time synthesis on a CPU. With respect to $f_o$-RMSE, Harmonic-Net with *harm* excitation signals achieved higher score than the other methods except for Harmonic-Net+ with *sine* excitation, and it maintained real-time speed even for 48-kHz synthesis. Harmonic-Net+ could not realize real-time synthesis because of the large number of parameters associated with 48-kHz synthesis. Additionally, Harmonic-Net+ suffered deterioration in SNR, SD and MCD. We found that the speech waveforms synthesized by Harmonic-Net+ were buzzy throughout. Fig. 8 shows the spectrograms up to 12 kHz of an original speech waveform included in the speech samples and those synthesized by Harmonic-Net with *harm* and Harmonic-Net+ with *harm* for full-band singing voice synthesis. Compared with the spectrograms of the original (Fig. 8(a)) and Harmonic-Net (Fig. 8(b)), that of Harmonic-Net+ (Fig. 8(c)) includes horizontal stripes especially in aperiodic components surrounded by blue squares. These components sound buzzy and degrade the synthesized speech quality of Harmonic-Net+.

Table IV shows the results of the objective evaluations for full-band singing voice synthesis in the $f_o$-conversion condition. With respect to $f_o$-RMSE, Harmonic-Net+ models achieved higher $f_o$ conversion accuracy than the other models. However, with respect to MCD, Harmonic-Net+ models were lower than the other models, and the speech waveforms synthesized by Harmonic-Net+ were also buzzy throughout. Compared with multi-speaker model trained using the JVS corpus, the Tohoku Kiritan corpus only contains about 1 h although the $f_o$ range of

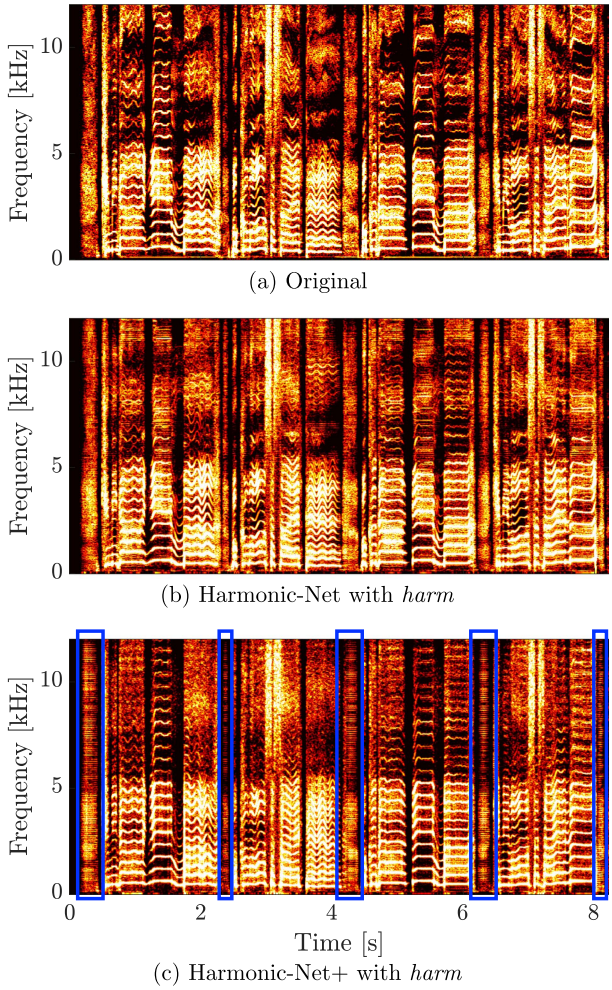
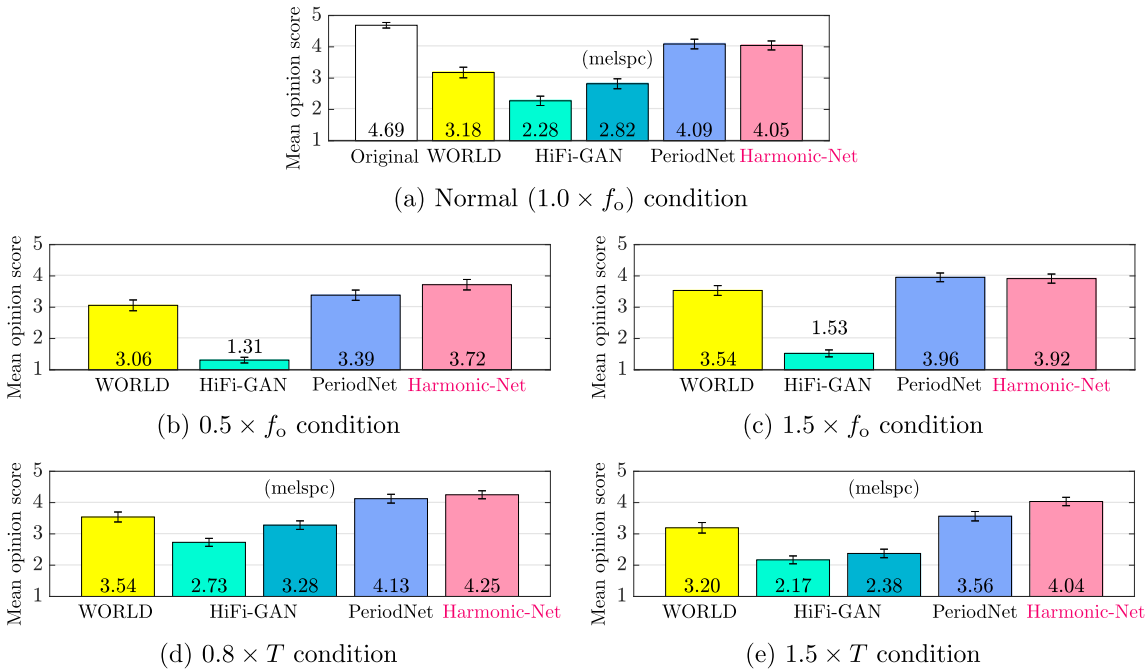



Fig. 8. Spectrograms of (a) original speech waveform and those synthesized by (b) Harmonic-Net with *harm* and (c) Harmonic-Net+ with *harm* for full-band singing voice synthesis. To show the buzzy components surrounded by blue squares in (c) more clearly, they are shown up to 12 kHz.

the Tohoku Kiritan corpus (58 to 793 Hz) is wider than that of the JVS corpus (Fig. 4). Then, LW-PDCNNs might not be able to be trained well due to the lack of training data. Therefore, further investigation of Harmonic-Net+ with a larger amount of training data for full-band singing voice synthesis is required as future work.

We also conducted a MOS test as a subjective evaluation. The evaluation conditions were the same as those for unseen speaker synthesis with multi-speaker models. According to the results of the objective evaluations and the preliminary MOS test, Harmonic-Net+ was not included in the MOS test. Additionally, although uSFGAN achieved high scores in the objective evaluations, it was not included in the MOS test because it could not outperform PeriodNet in preliminary experiments. Harmonic-Net with *harm* excitation signals was compared with WORLD, HiFi-GAN, and PeriodNet. Twenty subjects listened to all 20 phrases in the evaluation set for each model and each condition. Thus, the total number of phrases evaluated by each listening subject was 480 ($= 20$ phrases $\times$ $(6 + 4 + 4 + 5 + 5)$ models).




(a) Normal ($1.0 \times f_{\mathrm{o}}$) condition

(b) $0.5 \times f_{\mathrm{o}}$ condition

(c) $1.5 \times f_{\mathrm{o}}$ condition

(d) $0.8 \times T$ condition

(e) $1.5 \times T$ condition

Fig. 9. Results of MOS test for full-band singing voice synthesis in the normal, $f_{\mathrm{o}}$-conversion, and SR-conversion conditions. Confidence level of the error bars is 95%.

TABLE IV
RESULTS OF OBJECTIVE EVALUATIONS FOR FULL-BAND SINGING VOICE SYNTHESIS IN $f_{\mathrm{o}}$-CONVERSION CONDITION

| Model | excitation | MCD (dB) | $f_{\mathrm{o}}$-RMSE (Hz) |
|---|---|---|---|
| $0.5 \times f_{\mathrm{o}}$ | | | |
| WORLD (reference) | - | 3.91 | 12.9 |
| uSFGAN | *sine* | 4.25 | 12.9 |
| PeriodNet | *sine* | **3.77** | 13.3 |
| HiFi-GAN | - | 5.41 | 34.4 |
| Harmonic-Net | *sine* | 5.04 | 13.3 |
| | *pulse* | 4.57 | 17.1 |
| | *world* | 4.01 | 15.3 |
| | *harm* | 5.04 | 13.6 |
| Harmonic-Net+ | *sine* | 6.49 | 12.4 |
| | *pulse* | 5.86 | 12.6 |
| | *world* | 4.33 | **12.2** |
| | *harm* | 6.82 | 13.2 |
| $1.5 \times f_{\mathrm{o}}$ | | | |
| WORLD (reference) | - | 3.98 | 30.2 |
| uSFGAN | *sine* | **4.71** | 41.2 |
| PeriodNet | *sine* | 5.01 | 39.9 |
| HiFi-GAN | - | 6.43 | 94.5 |
| Harmonic-Net | *sine* | 5.59 | 38.5 |
| | *pulse* | 5.66 | 43.9 |
| | *world* | 5.13 | 43.8 |
| | *harm* | 5.59 | 38.5 |
| Harmonic-Net+ | *sine* | 6.73 | 39.1 |
| | *pulse* | 6.25 | 38.9 |
| | *world* | 5.34 | **36.2** |
| | *harm* | 6.81 | 37.3 |

Fig. 9 shows the results of the MOS test for full-band singing voice synthesis. According to Fig. 9(a), Harmonic-Net achieved almost the same score as PeriodNet. However, because PeriodNet has the problem of low inference speed, Harmonic-Net has the advantage of realizing fast and high-quality full-band singing voice synthesis. In the $f_{\mathrm{o}}$-conversion condition, Harmonic-Net also achieved a higher synthesis quality than HiFi-GAN which could not synthesize speech waveforms with $f_{\mathrm{o}}$-scaled features, and there was a significant difference between Harmonic-Net and PeriodNet in the $0.5 \times f_{\mathrm{o}}$ condition. In the SR-conversion condition, Harmonic-Net significantly achieved the best synthesis quality for both the $0.8 \times T$ and $1.5 \times T$ conditions. Therefore, the effectiveness of Harmonic-Net with *harm* excitation signals was validated for full-band singing voice synthesis.

### D. Evaluation of Text-to-Speech

Finally, we subjectively evaluated models using the FastSpeech-based TTS acoustic model. In the experiments, we compared Harmonic-Net+ with HiFi-GAN models that use mel-spectrograms or WORLD features.[17] In SR conversion, the phoneme durations, predicted by the duration predictor in FastSpeech, were changed for the $0.8 \times T$ and $1.5 \times T$ conditions. In preliminary experiments, there was no significant difference between Harmonic-Net+ and HiFi-GAN because the FastSpeech decoder could synthesize SR-converted acoustic features accurately, similarly to ScalerGAN [51] but differently from simple uniform resampling. Therefore, TTS was investigated only for $1.0 \times f_{\mathrm{o}}$ in the normal condition, and $0.5 \times f_{\mathrm{o}}$ and $1.5 \times f_{\mathrm{o}}$ in the $f_{\mathrm{o}}$-conversion condition. The evaluation conditions were the same as those for unseen speaker synthesis with multi-speaker models and full-band singing voice synthesis. Twenty subjects listened to ten randomly selected sentences from the evaluation set, for each model and each condition.

[17] Harmonic-Net was not included in the experiments because it could not outperform HiFi-GAN in the normal condition in preliminary experiments.

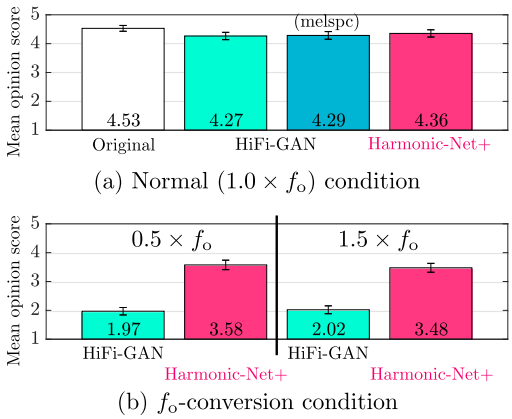



Fig. 10. Results of MOS test for single-speaker TTS in the normal and $f_o$-conversion conditions. Confidence level of the error bars is 95%.

Therefore, the total number of sentences evaluated by each listening subject was 80 (= 10 utterances × (4 + 2 + 2) models).

Fig. 10 shows the result of the MOS test for single-speaker TTS. According to Fig. 10(a), HiFi-GAN and Harmonic-Net+ achieved high performance and these was no significant difference between them. In the $f_o$-conversion condition, Harmonic-Net+ achieved a significantly higher synthesis quality even when scaled $f_o$ not included in the range of the training data was input, as unseen speaker synthesis with multi-speaker models and full-band singing voice synthesis. Therefore, the effectiveness of Harmonic-Net+ was confirmed for single-speaker TTS. Future work includes the integration of Harmonic-Net+ into an entire end-to-end TTS system, in a similar manner to [3], [31].

### *E. Discussion*

From the results of the MOS test shown in Fig. 5, the proposed excitation signal network with multi-channel harmonic waves corresponding to $f_o$ combined with the proposed LW-PDCNNs can realize high quality synthesis while keeping the $f_o$ controllability, whereas the conventional methods introduce sine waves or pulse trains. Especially for high $f_o$ conversion condition, only the proposed method with harmonic waves and LW-PDCNNs can realized high synthesis quality. The effectiveness of the proposed Harmonic-Net+ with multi-channel harmonic waves and LW-PDCNNs was validated for unseen speaker synthesis and TTS conditions from the results of the MOS tests depicted in Figs. 5 to 7, and 10. Although the effectiveness of the proposed LW-PDCNNs could not be validated for full-band singing voice synthesis due to the lack of training data according to the results of the objective evaluations (Table IV) and preliminary MOS test, the effectiveness of the proposed excitation signal network with multi-channel harmonic waves was validated from the results of the MOS test shown in Fig. 9. Further investigation of LW-PDCNNs with a larger amount of training data for full-band singing voice synthesis is required as future work. Additionally, the effectiveness of the proposed excitation signal network with multi-channel harmonic waves for SR conversion was also validated from the results of the MOS tests shown in Figs. 7 and 9.

## V. Conclusion

To realize fast and high-quality neural speech waveform generation while preserving the controllability of $f_o$ and SR, we proposed Harmonic-Net and Harmonic-Net+, which introduce an excitation signal network and non-AR LW-PDCNNs into the HiFi-GAN generator. The excitation signal network uses multi-channel harmonic waves corresponding to $f_o$ as excitation signals and we introduced a downsampling network that receives these excitation signals. LW-PDCNNs can flexibly change receptive fields corresponding to the input $f_o$, and we adjusted the network architecture to fit the structure of the HiFi-GAN generator. By introducing the proposed architectures, the controllability of $f_o$ is expected to be improved. Additionally, the direct input of the excitation signals is expected to improve the synthesis quality of SR conversion because the excitation signals are not resampled but the number of repetitions of the input excitation signals is changed. We conducted experiments for unseen speaker synthesis with multi-speaker models, full-band singing voice synthesis, and single-speaker TTS. The results of the experiments confirmed that the proposed excitation signal network and LW-PDCNNs worked effectively to improve the synthesis quality, compared with conventional models, while realizing real-time inference with a CPU in all (normal, $f_o$-conversion, and SR-conversion) conditions.

## References

[1] A. van den Oord et al., "WaveNet: A generative model for raw audio," in *Proc. ISCA Speech Synth. Workshop*, 2016, p. 125.
[2] J. Shen et al., "Natural TTS synthesis by conditioning WaveNet on MEL spectrogram predictions," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 4779–4783.
[3] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 5530–5540.
[4] J. Sotelo et al., "Char2Wav: End-to-end speech synthesis," in *Proc. Int. Conf. Learn. Representations*, 2017.
[5] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker-dependent WaveNet vocoder," in *Proc. Interspeech*, 2017, pp. 1118–1122.
[6] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Commun.*, vol. 27, no. 3/4, pp. 187–207, Apr. 1999.
[7] Y. Agiomyrgiannakis, "Vocaine the vocoder and applications in speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 4230–4234.
[8] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. Inf. Syst.*, vol. E99-D, no. 7, pp. 1877–1884, Jul. 2016.
[9] M. Airaksinen, L. Juvela, B. Bollepalli, J. Yamagishi, and P. Alku, "Comparison between STRAIGHT, glottal, and sinusoidal vocoding in statistical parametric speech synthesis," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 9, pp. 1658–1670, Sep. 2018.
[10] N. Kalchbrenner et al., "Efficient neural audio synthesis," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2415–2424.
[11] J.-M. Valin and J. Skoglund, "LPCNet: Improving neural speech synthesis through linear prediction," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 5891–5895.
[12] 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," in *Proc. Interspeech*, 2021, pp. 2217–2221.

[13] W. Ping, K. Peng, and J. Chen, "ClariNet: Parallel wave generation in end-to-end text-to-speech," in *Proc. Int. Conf. Learn. Representations*, 2019.

[14] R. Prenger, R. Valle, and B. Catanzaro, "WaveGlow: A flow-based generative network for speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 3617–3621.

[15] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 6199–6203.

[16] K. Kumar et al., "MelGAN: Generative adversarial networks for conditional waveform synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 14910–14921.

[17] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan, "WaveGrad: Estimating gradients for waveform generation," in *Proc. Int. Conf. Learn. Representations*, 2021.

[18] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "DiffWave: A versatile diffusion model for audio synthesis," in *Proc. Int. Conf. Learn. Representations*, 2021.

[19] G. Yang, S. Yang, K. Liu, P. Fang, W. Chen, and L. Xie, "Multi-band MelGAN: Faster waveform generation for high-quality text-to-speech," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2021, pp. 492–498.

[20] T. Okamoto, T. Toda, Y. Shiga, and H. Kawai, "Noise level limited sub-modeling for diffusion probabilistic vocoders," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 6029–6033.

[21] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 17022–17033.

[22] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[23] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden Markov models," *Proc. IEEE*, vol. 101, no. 5, pp. 1234–1252, May 2013.

[24] S. H. Mohammadi and A. Kain, "An overview of voice conversion systems," *Speech Commun.*, vol. 88, pp. 65–82, Apr. 2017.

[25] P. R. Cook, "Singing voice synthesis: History, current work, and future directions," *Comput. Music J.*, vol. 20, no. 3, pp. 38–46, 1996.

[26] J.-H. Kim, S.-H. Lee, J.-H. Lee, and S.-W. Lee, "Fre-GAN: Adversarial frequency-consistent audio synthesis," in *Proc. Interspeech*, 2021, pp. 2197–2201.

[27] 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," in *Proc. Interspeech*, 2021, pp. 2207–2211.

[28] T. Okamoto, T. Toda, and H. Kawai, "Multi-stream HiFi-GAN with data-driven waveform decomposition," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop*, 2021, pp. 610–617.

[29] M. Morrison, R. Kumar, K. Kumar, P. Seetharaman, A. Courville, and Y. Bengio, "Chunked autoregressive GAN for conditional waveform synthesis," in *Proc. Int. Conf. Learn. Representations*, 2022.

[30] T. Kaneko, K. Tanaka, H. Kameoka, and S. Seki, "iSTFTNet: Fast and lightweight mel-spectrogram vocoder incorporating inverse short-time Fourier transform," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 6207–6211.

[31] H. Chung, S.-H. Lee, and S.-W. Lee, "Reinforce-aligner: Reinforcement alignment search for robust end-to-end text-to-speech," in *Proc. Interspeech*, 2021, pp. 3635–3639.

[32] I. R. Titze et al., "Toward a consensus on symbolic notation of harmonics, resonances, and formants in vocalization," *J. Acoust. Soc. Amer.*, vol. 137, no. 5, pp. 3005–3007, May 2015.

[33] 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, Lang. Process.*, vol. 29, pp. 1134–1148, 2021.

[34] M. Morrison, Z. Jin, J. Salamon, N. J. Bryan, and G. J. Mysore, "Controllable neural prosody synthesis," in *Proc. Interspeech*, 2020, pp. 4437–4441.

[35] J. J. Webber, O. Perrotin, and S. King, "Hider-finder-combiner: An adversarial architecture for general speech signal modification," in *Proc. Interspeech*, 2020, pp. 3206–3210.

[36] X. Wang, S. Takaki, and J. Yamagishi, "Neural source-filter waveform models for statistical parametric speech synthesis," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 402–415, 2020.

[37] 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, Lang. Process.*, vol. 29, pp. 792–806, 2021.

[38] R. Yoneyama, Y.-C. Wu, and T. Toda, "Unified source-filter GAN: Unified source-filter network based on factorization of quasi-periodic parallel WaveGAN," in *Proc. Interspeech*, 2021, pp. 2187–2191.

[39] Y. Hono, S. Takaki, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "PeriodNet: A non-autoregressive raw waveform generative model with a structure separating periodic and aperiodic components," *IEEE Access*, vol. 9, pp. 137599–137612, 2021.

[40] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech., Signal Process.*, vol. 34, no. 4, pp. 744–754, Aug. 1986.

[41] O. Watts, C. Valentini-Botinhao, and S. King, "Speech waveform reconstruction using convolutional neural networks with noise and periodic inputs," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 7045–7049.

[42] Y. Ai and Z.-H. Ling, "A neural vocoder with hierarchical generation of amplitude and phase spectra for statistical parametric speech synthesis," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 839–851, 2020.

[43] Z. Liu, K. Chen, and K. Yu, "Neural homomorphic vocoder," in *Proc. Interspeech*, 2020, pp. 240–244.

[44] J. Engel, L. H. Hantrakul, C. Gu, and A. Roberts, "DDSP: Differentiable digital signal processing," in *Proc. Int. Conf. Learn. Representations*, 2020.

[45] K. Matsubara et al., "Full-band LPCNet: A real-time neural vocoder for 48 kHz audio with a CPU," *IEEE Access*, vol. 9, pp. 94923–94933, 2021.

[46] W. Verhelst and M. Roelands, "An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 1993, pp. 554–557.

[47] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Commun.*, vol. 9, no. 5/6, pp. 453–467, Dec. 1990.

[48] T. Hayashi, A. Tamamori, K. Kobayashi, K. Takeda, and T. Toda, "An investigation of multi-speaker training for WaveNet vocoder," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop*, 2017, pp. 712–718.

[49] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-Time Signal Processing*. Upper Saddle River, NJ, USA: Prentice Hall, 1999.

[50] T. Okamoto, K. Matsubara, T. Toda, Y. Shiga, and H. Kawai, "Neural speech-rate conversion with multispeaker WaveNet vocoder," *Speech Commun.*, vol. 138, pp. 1–12, Mar. 2022.

[51] E. Cohen, F. Kreuk, and J. Keshet, "Speech time-scale modification with GANs," *IEEE Signal Process. Lett.*, vol. 29, pp. 1067–1071, 2022.

[52] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1558–1566.

[53] K. Matsubara, T. Okamoto, R. Takashima, T. Takiguchi, T. Toda, and H. Kawai, "Comparison of real-time multi-speaker neural vocoders on CPUs," *Acoust. Sci. Technol.*, vol. 43, no. 2, pp. 121–124, Mar. 2022.

[54] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8024–8035.

[55] S. Takamichi et al., "JSUT and JVS: Free Japanese voice corpora for accelerating speech synthesis research," *Acoust. Sci. Technol.*, vol. 41, no. 5, pp. 761–768, Sep. 2020.

[56] I. Ogawa and M. Morise, "Tohoku Kiritan singing database: A singing database for statistical parametric singing synthesis using Japanese pop songs," *Acoust. Sci. Technol.*, vol. 42, no. 3, pp. 140–145, May 2021.

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

[58] M. Morise, "D4C, A band-aperiodicity estimator for high-quality speech synthesis," *Speech Commun.*, vol. 84, pp. 67–65, Nov. 2016.

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

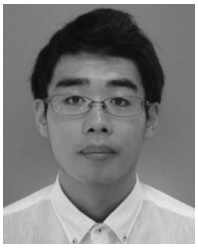
[60] K. Tachibana, T. Toda, Y. Shiga, and H. Kawai, "An investigation of noise shaping with perceptual weighting for WaveNet-based speech generation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5664–5668.

[61] J. Lorenzo-Trueba et al., "Towards achieving robust universal neural vocoding," in *Proc. Interspeech*, 2019, pp. 181–185.

[62] Y. Ren et al., "FastSpeech: Fast, robust and controllable text to speech," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 3165–3174.

[63] T. Hayashi et al., "Espnet-TTS: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 7654–7658.

[64] *Methods for Subjective Determination of Transmission Quality*, ITU-T Recommendation P. 800, 1996.

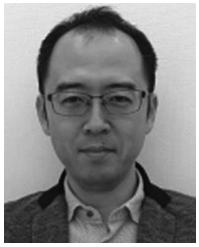


**Keisuke Matsubara** received the B.E. and M.S. degrees from Kobe University, Kobe, Japan, in 2020 and 2022, respectively. Since 2019, he has been an Internship Student with the National Institute of Information and Communications Technology, Kyoto, Japan. His research interests include speech synthesis and voice conversion. He was the recipient of the 21st Student Presentation Award from the Acoustical Society of Japan, in 2021.

**Takuma Okamoto** (Member, IEEE) received the B.E., M.S., and Ph.D. degrees from Tohoku University, Sendai, Japan, in 2004, 2006, and 2009, respectively. From 2009, he was a Postdoctoral Research Fellow with Tohoku University. From 2012 to 2020, he was a Researcher with the National Institute of Information and Communications Technology, Kyoto, Japan, where he is currently a Senior Researcher. His main research interests include sound field synthesis and speech synthesis. He is a Member of the Audio Engineering Society and Acoustical Society of Japan. He was the recipient of the 32nd Awaya Prize Young Researcher Award, 57th Sato Prize Paper Award, and 9th Society Activity Contribution Award from ASJ, in 2012, 2017 and 2022, respectively.

**Ryoichi Takashima** (Member, IEEE) received the B.E., M.E., and Dr.Eng. degrees in computer science from Kobe University, Kobe, Japan, in 2008, 2010, and 2013, respectively. From 2013 to 2018, he was a Researcher with Hitachi Ltd., Tokyo, Japan. From 2016 to 2018, he was on loan to the National Institute of Information and Communication Technology, Kyoto, Japan. He is currently an Associate Professor with Kobe University. His research interests include machine learning and signal processing. He is a Member of ASJ.

**Tetsuya Takiguchi** (Member, IEEE) received the M.Eng. and Dr.Eng. degrees from the Nara Institute of Science and Technology, Ikoma, Japan, in 1996 and 1999, respectively. From 1999 to 2004, he was a Researcher with the Tokyo Research Laboratory, IBM Research. From 2004 to 2016, he was an Associate Professor with Kobe University, Kobe, Japan. Since 2016, he has been a Professor with Kobe University. From May 2008 to September 2008, he was a Visiting Scholar with the Department of Electrical Engineering, University of Washington, Seattle, WA, USA. From March 2010 to September 2010, he was a Visiting Scholar with the Institute for Learning and Brain Sciences, University of Washington. From April 2013 to October 2013, he was a Visiting Scholar with the Laboratoire d'InfoRmatique en Image et Systémes d'information, INSA Lyon. His research interests include speech, image, and brain processing, and multimodal assistive technologies for people with articulation disorders. He was the recipient of the best paper award from IEEE ICME 2008.



**Tomoki Toda** (Senior Member, IEEE) received the B.E. degree from Nagoya University, Nagoya, Japan, in 1999, and the M.E. and D.E. degrees from the Nara Institute of Science and Technology (NAIST), Ikoma, Japan, in 2001 and 2003, respectively. He was an Assistant Professor from 2005 to 2011 and an Associate Professor from 2011 to 2015 with NAIST. Since 2015, he has been a Professor with the Information Technology Center, Nagoya University. His research interests include statistical approaches to speech and audio processing. From 2003 to 2005, he was a Research Fellow of the Japan Society for the Promotion of Science. He was the recipient of the more than ten article/achievement awards, including the IEEE SPS 2009 Young Author Best Paper Award and 2013 EURASIP-ISCA Best Paper Award (*Speech Communication journal*).

**Hisashi Kawai** (Member, IEEE) received the B.E., M.E., and D.E. degrees in electronic engineering from The University of Tokyo, Tokyo, Japan, in 1984, 1986, and 1989, respectively. In 1989, he joined Kokusai Denshin Denwa Company Ltd. From 2000 to 2004, he worked for ATR Spoken Language Translation Research Laboratories, where he engaged in the development of text-to-speech synthesis system. From October 2004 to March 2009 and from April 2012 to September 2014, he worked for KDDI Research and Development Laboratories, where he was engaged in the research and development of speech information processing, speech quality control for telephone, speech signal processing, acoustic signal processing, and communication robots. From April 2009 to March 2012 and since October 2014, he has been working with the National Institute of Information and Communications Technology, Kyoto, Japan, where he is engaged in development of speech technology for spoken language translation. He is a Member of the Acoustical Society of Japan and Institute of Electronics, Information and Communication Engineers.