Received August 29, 2021, accepted September 19, 2021, date of publication October 5, 2021, date of current version October 13, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3118033

PeriodNet: A Non-Autoregressive Raw Waveform Generative Model With a Structure Separating Periodic and Aperiodic Components

YUKIYA HONO, SHINJI TAKAKI, (Member, IEEE), KEI HASHIMOTO, (Member, IEEE), KEIICHIRO OURA, YOSHIHIKO NANKAKU, (Member, IEEE), AND KEIICHI TOKUDA, (Fellow, IEEE)

Department of Computer Science, Nagoya Institute of Technology, Nagoya 466-8555, Japan

Corresponding author: Yukiya Hono (hono@sp.nitech.ac.jp)

This work was supported in part by the Japan Society for the Promotion of Science (JSPS) KAKENHI under Grant JP18K11163 and Grant JP19H04136, and in part by the CASIO SCIENCE PROMOTION FOUNDATION.

ABSTRACT This paper presents PeriodNet, a non-autoregressive (non-AR) waveform generative model with a new model structure for modeling periodic and aperiodic components in speech waveforms. Non-AR raw waveform generative models have enabled the fast generation of high-quality waveforms. However, the variations of waveforms that these models can reconstruct are limited by training data. In addition, typical non-AR models reconstruct a speech waveform from a single Gaussian input despite the mixture of periodic and aperiodic signals in speech. These may significantly affect the waveform generation process in some applications such as singing voice synthesis systems, which require reproducing accurate pitch and natural sounds with less periodicity, including husky and breath sounds. PeriodNet uses a parallel or series model structure to model a speech waveform to tackle these problems. Two sub-generators connected in parallel or in series take an explicit periodic and aperiodic signal (sine wave and Gaussian noise) as an input. Since PeriodNet models periodic and aperiodic components by focusing on whether these input signals are autocorrelated or not, it does not require external periodic/aperiodic decomposition during training. Experimental results show that our proposed structure improves the naturalness of generated waveforms. We also show that speech waveforms with a pitch outside of the training data range can be generated with more naturalness.

INDEX TERMS Generative adversarial network, neural vocoder, signal processing, singing voice synthesis, waveform generative model.

I. INTRODUCTION

Speech synthesis technology has been rapidly advancing with the introduction of neural networks (NNs). Text-to-speech synthesis (TTS) and singing voice synthesis (SVS) are techniques for generating speech and singing voices on the basis of given input text and musical scores, respectively [1]–[7]. In typical TTS and SVS systems, a low-dimensional representation of speech and singing voices (sometimes termed an acoustic feature) is predicted by an acoustic model, and a corresponding waveform is generated by a vocoder. Conventional vocoders such as STRAIGHT [8] and WORLD [9] are carefully engineered and knowledge-based procedures that are based on signal processing. These vocoders rely on many oversimplified assumptions, such as a time-invariant linear filter and a stationary Gaussian process. Since these assumptions act as constraints on reconstructing speech waveforms and can cause detailed temporal structure and phase information to be lost, the generated speech waveforms can be degraded.

In recent years, NN-based raw waveform generative models have been highly successful. WaveNet [10] and SampleRNN [11] demonstrate remarkable performance by directly modeling the distribution of waveform samples. Since these models can be used as a vocoder by modeling waveforms by conditioning acoustic features [12], they have

The associate editor coordinating the review of this manuscript and approving it for publication was Valentina E. Balas.

succeeded in replacing the conventional vocoders by giving speech applications the benefit of generating high-quality speech waveforms [13]–[15]. They have a huge network architecture with AR mechanisms, which suffer from a slow inference speed. Although some compact AR models [16], [17] have been proposed to improve the inference speed, carefully engineered optimization is required for achieving an adequate inference speed. Therefore, such AR models are not suited for real-time applications.

Recently, significant efforts have been put into the development of non-AR waveform generative models. Flow-based models, including inverse autoregressive flows (IAFs) [18], [19], generative flows (Glows) [20]–[22], and continuous normalizing flows (CNFs) [23], [24], generative adversarial network (GAN)-based models [25]–[30], and variational auto-encoder (VAE)-based models [31] have been proposed. Although these models can efficiently generate waveform samples in parallel, the generated audio quality is sometimes inferior to that of AR models. Recent attempts [32], [33] have incorporated diffusion probabilistic models to reduce the gap between non-AR and AR models in terms of audio quality. However, as can be seen from the fact that most previous works used a moderate sampling rate such as 16 kHz or 24 kHz, it is still not easy to generate high-fidelity waveforms with high sampling rates (e.g., 48 kHz) in real time.

While NN-based vocoders (neural vocoders) can generate high-fidelity speech waveforms without being restricted by knowledge or by assumptions, the lack of acoustic controllability and robustness to input acoustic features are issues. In fact, if the pitch given to these models is outside the range of the training data, it is difficult to generate a speech waveform with a corresponding pitch. Even if the given input pitch is within the range of the training data, these vocoders sometimes generate a waveform with a different pitch. To address this problem, an explicit periodic signal such as a sine wave is used as the input signal for non-AR models [34], [35]. The authors of [36], [37] introduced a pitch-dependent convolution mechanism into AR and non-AR neural vocoders to tackle this problem.

When neural vocoders are used in SVS systems, they should also have the ability to generate components with less periodicity, such as husky voice and breath sounds, in addition to having a high pitch accuracy. Since it is known that a speech waveform contains a mixture of periodic and aperiodic components, proper reconstruction of both components is essential. There are several methods for decomposing the periodic and aperiodic components contained in natural waveforms [38]–[40]. However, it is difficult to decompose these components from natural speech waveforms accurately, and it is not optimal to use these waveforms with decomposition errors as the training data. To generate speech with less periodicity, the speech waveform should be modeled by taking into account the mixture of these components without explicitly separating them for high-quality waveform generation.

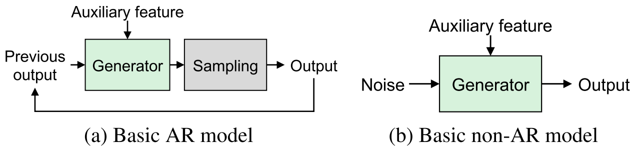
In this paper, we propose PeriodNet, a non-AR raw waveform generative model with a novel structure for appropriately modeling the periodic and aperiodic components in speech waveforms. Two sub-generators are connected in parallel or in series, and each generator reconstructs periodic and aperiodic waveforms from a sine wave and Gaussian noise. By using these two signals with different characteristics in terms of autocorrelation as the input signal, PeriodNet can model periodic and aperiodic components without explicit decomposition methods. While we previously considered these model structures using a female singer's singing voice corpus [41] (Sections IV-B1 and IV-B2), this paper additionally evaluates the effectiveness of PeriodNet with other singers' corpora and explicit decomposition techniques. The main contributions of this paper are summarized as follows.

- We give details on the training framework of PeriodNet (Section III-B).
- We show that PeriodNet can model a speech waveform while appropriately separating periodic and aperiodic components during the training process by comparing it with systems that use periodic and aperiodic waveforms pre-decomposed by using explicit decomposition techniques (Section V).
- We assess the naturalness of generated singing voice waveforms with different female and male singers and suggest that PeriodNet is suitable for generating singing voice waveforms (Section VI).

This paper is organized as follows. In Section II, we review the recent neural waveform generative models. Section III introduces our proposed model structure and its training framework. Sections IV, V, and VI present experimental evaluations. Finally, Section VII concludes this paper.

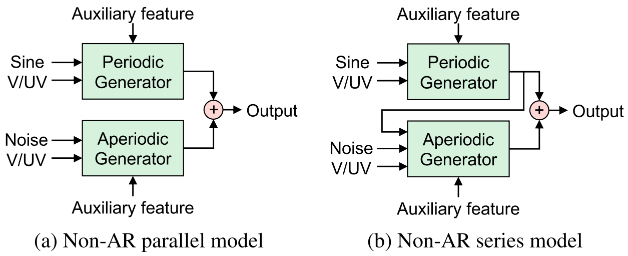
II. NEURAL WAVEFORM GENERATIVE MODELS

The modeling of speech waveforms is a challenging task because they have a high temporal resolution. The pioneer AR model, called WaveNet [10], enables the direct modeling of speech waveform samples through dilated causal convolution with a large receptive field. SampleRNN [11] is an alternative architecture that explicitly models speech waveforms at different temporal resolutions with multi-scale recurrent neural networks. A key idea of these models is using an autoregressive probabilistic model that describes the distribution of current samples conditioned on previous waveform samples. These models can be used as a vocoder by conditioning them with auxiliary features such as acoustic features, as shown in Fig. 1(a). These AR models can adequately predict speech waveforms because the sampling process is strictly serial. However, this serial sampling makes waveform generation slow, and thus, these models are impractical for real-time applications. While some compact AR models [16], [17] can reduce the computational cost at inference, there is a limitation to how much speed can be improved because these models are inherently serial and cannot avoid serial sampling.



(a) Basic AR model (b) Basic non-AR model

FIGURE 1. Structures of conventional neural waveform modeling. Acoustic feature representations, such as mel-spectrogram, mel-cepstrum, and fundamental frequency (F_0), are generally used as auxiliary features.



(a) Non-AR parallel model (b) Non-AR series model

FIGURE 2. Structures of proposed neural waveform modeling. Periodic and aperiodic generators take a sine wave and Gaussian noise, which have different characteristics in terms of autocorrelation, along with a sample-level voiced/unvoiced (V/UV) signal.

In contrast, various types of non-AR waveform generative models have been proposed to overcome the slow inference speed of AR models. These non-AR models generate waveforms in parallel from pre-generated signals such as Gaussian noise, as shown in Fig. 1(b). A teacher-student-based framework (e.g., Parallel WaveNet [18] and ClariNet [19]) distills a trained AR teacher WaveNet into an inverse flow-based student model. The authors of [20]–[22] incorporated a flow-based generative model based on Glow [42], which can be directly learned by minimizing the negative log-likelihood of data without a distillation process. Another group of non-AR models [25]–[29] is based on an adversarial training framework [43]. Combining adversarial loss and auxiliary loss, such as multi-resolution short-time Fourier transform (STFT) loss and feature matching loss, enables non-AR models to be learned efficiently. These non-AR methods often yield lower quality samples than AR models because they need to generate speech waveforms, which are inherently serial, with fewer sequential operations. The authors of WaveGrad [32] and DiffWave [33] recently utilized diffusion probabilistic models with denoising score matching for waveform generation to overcome this deterioration. Although these models can achieve high-fidelity waveform generation, the inference speed tends to be slower than other non-AR models.

In addition to the data-driven approach described above, neural waveform generation methods incorporating speech-related knowledge and assumptions have also been proposed for both AR and non-AR models. For example, approaches generating glottal excitation signals [44]–[47] or linear predictive residual signals [48], [49] have been proposed to facilitate neural waveform modeling by reducing the burden of speaker identity and spectral information modeling. The authors of [34] proposed a non-AR vocoder taking mixed sine-based excitation inputs made from the fundamental frequency (F_0) and Gaussian noise as inputs, which can provide accurate pitch control via the manipulation of F_0 values. Meanwhile, our previous work [35] presented a deep auto-encoder-based framework with tailored periodic and aperiodic inputs. The speech waveform is modeled as a sum of a periodic component and 24 frequency-banded aperiodic components by using a carefully designed auto-encoder-based model; however, this structured model limits the flexibility of the generator. Although introducing knowledge and assumptions from signal processing brings some benefits such as improving controllability, interpretability, and training efficiency, this may affect the quality of the generated waveform.

III. PeriodNet

We propose PeriodNet, which has a parallel or series model structure to model a speech waveform considering its periodicity and aperiodicity. PeriodNet consists of two sub-generators called the *periodic generator* and *aperiodic generator*. They are connected in parallel or in series for separating periodic and aperiodic components in speech waveforms, as shown in Fig. 2. Typically, target periodic and aperiodic waveforms are required when explicitly modeling periodic and aperiodic waveforms. However, it is not easy to separate them accurately from natural waveforms. To model periodic and aperiodic components without a pre-decomposed waveform, we introduce two simple assumptions:

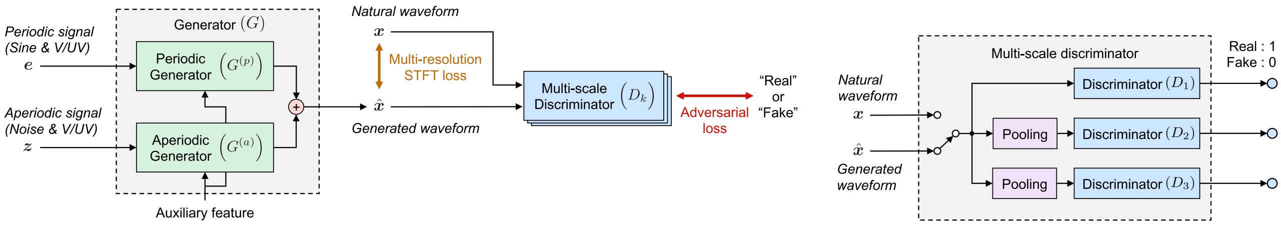
1) A speech waveform can represent the sum of periodic and aperiodic waveforms.
2) Periodic and aperiodic waveforms of speech can be easily generated from an explicit periodic signal with autocorrelation (such as a sine wave) and an explicit aperiodic signal without it (such as noise), respectively.

The proposed parallel and serial model structures are based on these assumptions and do not require any explicit periodic/aperiodic decomposition technique during training. PeriodNet is a non-AR waveform generative model incorporating these structures into a recent GAN-based waveform generative model [25]. Moreover, since PeriodNet can synthesize a periodic waveform from a sine-based periodic signal, it achieves high-fidelity waveform generation while attaining pitch controllability.

We explain the model structures of the PeriodNet generators in Section III-A and a training framework in Section III-B.

A. MODEL STRUCTURES

We introduce two types of model structures based on the different assumptions regarding the dependencies between periodic and aperiodic waveforms. The first one is the *parallel model structure*, as shown in Fig. 2(a). This structure is based on the assumption that periodic and aperiodic waveforms are independent of each other. A periodic signal



(a) Training framework of PeriodNet parallel model

(b) Details of multi-scale discriminator

FIGURE 3. Overview of PeriodNet parallel model. Generator has two sub-generators connected in parallel, and generated waveform is given as sum of both sub-generators. Generator is trained to minimize multi-resolution STFT loss and adversarial loss with multi-scale discriminator.

consisting of a sine wave and a sample-level voiced/unvoiced (V/UV) signal is fed into the periodic generator to predict a periodic waveform. An aperiodic signal consisting of Gaussian noise and a sample-level V/UV signal is fed into the aperiodic generator to predict an aperiodic waveform. The second type is the *series model structure*, as shown in Fig. 2(b). This is a model structure that can model the dependency between aperiodic and periodic waveforms, taking into account the possibility that the aperiodic waveform can change synchronously with the periodic waveform. To model the dependencies between the two components, we take advantage of the property that the sine-based periodic signal is deterministic. The former generator takes a periodic signal, and the latter generator takes an aperiodic signal and the output of the former generator. A residual connection is introduced between the two sub-generators so that the latter can predict only aperiodic waveforms. Hence, the former generator works as a periodic generator, and the latter works as an aperiodic generator.

To use PeriodNet as a vocoder, acoustic features are given to it as auxiliary features. In the parallel and series models, different acoustic features can be selected for the auxiliary features of the periodic and aperiodic generators, making it possible to obtain a more robust neural vocoder with proper conditioning.

B. DETAILS OF TRAINING FRAMEWORK

We now give details on the training framework for PeriodNet. An overview of the PeriodNet parallel model is shown in Fig. 3. The PeriodNet generator G is composed of two sub-generator modules: a periodic generator $G^{(p)}$ and an aperiodic generator $G^{(a)}$. The generated samples $\hat{\boldsymbol{x}}$ are obtained by

$$\hat{\boldsymbol{x}} = G(\boldsymbol{e}, \boldsymbol{z}) = G^{(p)}(\boldsymbol{e}) + G^{(a)}(\boldsymbol{z}), \tag{1}$$

where $\boldsymbol{e}$ is an explicit periodic signal concatenated with a sine wave and V/UV signal, and $\boldsymbol{z}$ is an explicit aperiodic signal concatenated with Gaussian noise and a V/UV signal. Note that various architectures can be used for periodic and aperiodic generators. In this paper, we adopt a WaveNet [10]-like architecture with non-causal dilated convolution and skip connections.

PeriodNet learns a distribution of realistic waveforms following an adversarial game between a generator and a discriminator. It uses multiple discriminators $(D_1, D_2, \ldots, D_K)$ with different temporal resolutions to discriminate between natural and generated waveform samples. This multi-scale architecture is motivated by the success of MelGAN [26] and helps to evaluate features for a different frequency range of waveform samples. The discriminator D_k discriminates speech samples with a temporal resolution of $1/k$ times the original waveform. The loss function of each discriminator D_k can be written as

$$\mathcal{L}_D(D_k) = \mathbb{E}_{\boldsymbol{x}}\left[\,(1 - D_k(\boldsymbol{x}))^2\,\right] + \mathbb{E}_{\boldsymbol{e},\boldsymbol{z}}\left[D_k(\hat{\boldsymbol{x}})^2\right], \tag{2}$$

where $\boldsymbol{x}$ denotes natural waveform samples.

The generator is trained to minimize adversarial loss, which is designed on the basis of least squares GAN [50]. The adversarial loss for the generator G to deceive the discriminator D_k is given by

$$\mathcal{L}_{adv}(G, D_k) = \mathbb{E}_{\boldsymbol{e},\boldsymbol{z}}\left[\left(1 - D_k(\hat{\boldsymbol{x}})\right)^2\right]. \tag{3}$$

To improve the stability and efficiency of the adversarial training, multi-resolution STFT loss is used along with adversarial loss. The STFT loss of the m-th temporal resolution $\mathcal{L}_{sp}^{<m>}$ is calculated as:

$$\mathcal{L}_{sp}^{<m>}(G) \ = \mathbb{E}_{\boldsymbol{e},\boldsymbol{z},\boldsymbol{x}}\left[\mathcal{L}_{sc}^{<m>}(\boldsymbol{x}, \hat{\boldsymbol{x}}) + \mathcal{L}_{mag}^{<m>}(\boldsymbol{x}, \hat{\boldsymbol{x}})\right], \tag{4}$$

where $\mathcal{L}_{sc}^{<m>}$ and $\mathcal{L}_{mag}^{<m>}$ denote spectral convergence loss and amplitude spectral loss, respectively. They are given by

$$\mathcal{L}_{sc}^{<m>}(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \frac{\left\| |\mathrm{STFT}(\boldsymbol{x})| - |\mathrm{STFT}(\hat{\boldsymbol{x}})| \right\|_F}{\left\| |\mathrm{STFT}(\boldsymbol{x})| \right\|_F}, \tag{5}$$

$$\mathcal{L}_{mag}^{<m>}(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \frac{1}{N}\left\| \log|\mathrm{STFT}(\boldsymbol{x})| - \log|\mathrm{STFT}(\hat{\boldsymbol{x}})| \right\|_1, \tag{6}$$

where $\|\cdot\|_F$ and $\|\cdot\|_1$ denote Frobenius and L_1 norms, respectively, and STFT$(\cdot)$ and N denote the amplitude spectrum and the number of frequency bins, respectively. Finally, the generator G is optimized to minimize a loss function $\mathcal{L}_G(G, D)$ given by

$$\begin{aligned}\mathcal{L}_G&(G, D)\\ &= \frac{1}{M}\sum_{m=1}^{M}\mathcal{L}_{sp}^{<m>}(G) + \frac{\lambda_{adv}}{K}\sum_{k=1}^{K}\mathcal{L}_{adv}(G, D_k),\end{aligned} \tag{7}$$

where λ_{adv} denotes a hyperparameter of the adversarial loss.

In the PeriodNet series model, the generated samples are obtained by

$$\hat{\boldsymbol{x}} = G'(\boldsymbol{e}, \boldsymbol{z}) = G'^{(p)}(\boldsymbol{e}) + G'^{(a)}(G'^{(p)}(\boldsymbol{e}), \boldsymbol{z}), \tag{8}$$

where $G'^{(p)}$ and $G'^{(a)}$ denote the periodic and aperiodic generator in the series model, respectively. The series model is also trained using the adversarial training framework in the same fashion as the parallel model described above.

IV. EXPERIMENT 1

Following the explanation on our PeriodNet, we discuss the experiments. In this section, we will investigate the effectiveness of PeriodNet using a single female singer dataset.

A. EXPERIMENTAL CONDITIONS

Seventy Japanese children's songs performed by one female singer (F01) were used. Singing voice signals were sampled at 48 kHz, and each sample was quantized by 16 bits. Sixty songs (approx. 65 min.) were used for training, and the rest (approx. 5 min.) were used for testing. The auxiliary features consisted of 50-dimensional mel-cepstral coefficients, 25-dimensional mel-cepstral analysis aperiodicity measures, 1-dimensional continuous log F_0 values, and 1-dimensional V/UV binary code. To reduce F_0 extraction errors such as V/UV detection errors and octave confusions, voting results from three F_0 estimators were used as the log F_0. [51]. Mel-cepstral coefficients were extracted from smoothed spectra analyzed by WORLD [9]. Feature vectors were extracted with a 5-ms shift, and the features were normalized to have zero mean and unit variance before training.

The following seven systems were compared.

- **AR**: An AR model based on WaveNet [10].
- **BM1**: The non-AR baseline model shown in Fig. 1(b). The generator took Gaussian noise and a V/UV signal as input and was conditioned on all auxiliary features.
- **BM2**: The modified version of the non-AR baseline model, in which the generator took a sine wave and a V/UV signal as input and was conditioned on all auxiliary features.
- **BM3**: The modified version of the non-AR baseline model, in which the generator took a sine wave, Gaussian noise, and a V/UV signal as the input and was conditioned on all auxiliary features.
- **PM1**: The non-AR parallel model shown in Fig. 2(a). The periodic generator took a sine wave and a V/UV signal as input, and the aperiodic generator took Gaussian noise and a V/UV signal as input. Both generators were conditioned on all auxiliary features.
- **PM2**: The non-AR parallel model shown in Fig. 2(a). Unlike **PM1**, the aperiodic generator was conditioned by auxiliary features other than F_0.
- **SM**: The non-AR series model shown in Fig. 2(b). The periodic generator took a sine wave and a V/UV signal as input, and the aperiodic generator took Gaussian noise, a V/UV signal, and the output signal of the periodic generator as input. Both generators were conditioned on all auxiliary features.

TABLE 1. Parameter settings of multi-resolution STFT loss. Hanning window was applied before the FFT process.

STFT loss	FFT size	Window size	Frame shift
$\mathcal{L}_{sp}^{<1>}$	1024	600 (12.5 ms)	120 (2.5 ms)
$\mathcal{L}_{sp}^{<2>}$	2048	1200 (25 ms)	240 (5 ms)
$\mathcal{L}_{sp}^{<3>}$	4096	2400 (50 ms)	480 (10 ms)

AR consisted of 30 layers of dilated residual convolution blocks with causal convolution. The dilations of **AR** were set to 1, 2, 4, . . . , 512, and 10 dilation layers were stacked three times. The channel size for the dilations, residual blocks, and skip-connections in **AR** was set to 256, and the filter size was set to 2. For **AR**, the waveform samples were quantized from 16 bits to 8 bits by using the μ-law algorithm [52].[1]

The generators of **BM1**, **BM2**, and **BM3** and the periodic generators of **PM1**, **PM2**, and **SM** consisted of 30 layers of dilated residual convolution blocks with 3 dilation cycles, similar to **AR**. The aperiodic generators of **PM1**, **PM2**, and **SM** consisted of 10 layers of dilated residual convolution blocks without dilation cycles. The channel size for the dilations, residual blocks, and skip-connections was set to 64, and the filter size was set to 3. Note that the sizes of the two sub-generators are different because we assumed that the generation of the aperiodic waveform does not require a larger receptive field and is relatively straightforward than that of the periodic waveform. The discriminators of **BM1**, **BM2**, **BM3**, **PM1**, **PM2**, and **SM** had a multi-scale architecture with three discriminators. The discriminators took 48-kHz full-resolution waveforms and 24-kHz and 16-kHz downsampled waveforms. The downsampling was performed using average pooling. Each discriminator consisted of 10 non-causal dilated convolutions with a leaky ReLU activation function. We applied weight normalization [53] to all convolutional layers.

At the training stage, the multi-resolution STFT loss was calculated as the sum of three different STFT losses, as shown in Table 1. The hyperparameter λ_{adv} in Eq. (7) was set to 4.0. All models were trained using the RAdam optimizer [54] with 1000K iterations. Specifically, for **BM1**, **BM2**, **BM3**, **PM1**, **PM2**, and **SM**, the discriminators were not used for the first 100K iterations, and then both the generator and discriminator were jointly trained afterward.

The sine waves for the input of **BM2**, **BM3**, **PM1**, **PM2**, and **SM** were generated on the basis of glottal closure points extracted from natural speech using REAPER [55] in the training stage. The purpose of this was to input a sine wave that is close in phase to the target's natural waveform during

[1] Recent TTS systems such as [14] use AR WaveNet without μ-law quantization by modeling waveform samples as a mixture of logistic distributions. However, since it is not easy to train a WaveNet that can generate 48-kHz waveform samples from only about 1 hour of recorded speech, we applied μ-law quantization to train AR WaveNet.

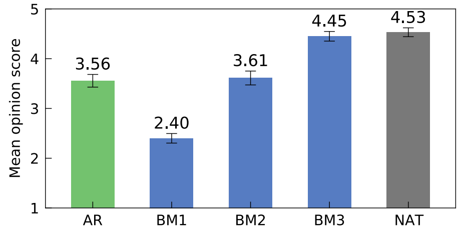


FIGURE 4. Subjective evaluation results obtained using a single female-singer data with 95% confidence intervals.

training. Meanwhile, sine waves were generated on the basis of the F_0 values in the synthesis stage. Note that the sample-level V/UV signal included in the input signal of the non-AR models was smoothed by applying a moving average in advance.

B. SUBJECTIVE EVALUATIONS

1) COMPARISON OF AR/NON-AR MODELS AND THE INPUT SIGNALS OF NON-AR MODELS

To evaluate the naturalness of the generated singing voice, we conducted mean opinion score (MOS) tests. We first used **AR**, **BM1**, **BM2**, **BM3**, and **NAT** to compare the waveform generative models with and without the AR structure and the input signals for the non-AR models. Note that **NAT** indicates a recorded natural waveform. The participants were 16 native Japanese speakers, and each participant evaluated 10 phrases randomly selected from the test data. After listening to each test sample in the MOS test, the participants were asked to score the naturalness of the samples out of five (1 = Bad; 2 = Poor; 3 = Fair; 4 = Good; and 5 = Excellent). The demo phrases can be found on the demo page [56].

The results of the subjective evaluation are shown in Fig. 4. **BM1** yielded a lower MOS value than **AR**, indicating that it is difficult to generate a high-quality singing voice from Gaussian noise. **BM2** showed a score comparable to **AR**. This indicates that inputting periodic signals into non-AR models is an alternative approach to the AR structure for reconstructing waveforms with periodicity. **BM3**, which inputs both explicit periodic and aperiodic signals, got a better score than **AR** and reached a MOS value close to **NAT**. This shows the effectiveness of using both explicit periodic and aperiodic signals as inputs for non-AR waveform generative models.

2) COMPARISON OF MODEL STRUCTURES OF NON-AR WAVEFORM GENERATIVE MODELS

To compare the model structures of non-AR waveform generative models, we conducted two subjective evaluation experiments using **BM3**, **PM1**, **PM2**, and **SM**. Neural vocoders need to appropriately generate waveforms with a pitch outside the range of training data. In these experiments, samples were generated by each model conditioned on two different scales of F_0: original F_0 and F_0 upward-shifted by 1200 cents. Note that upward shifting by 1200 cents is equivalent to a double scale operation. The upward-shifted F_0 contained F_0 that were outside the range of the training data, as shown in Fig. 5. The F_0 range of a singing voice is typically wide because the pitch of the singing voice greatly changes in accordance with the note pitch in a musical score. Thus, the range of the upward-shifted F_0 partly overlaps with that of the original F_0. In the experiment with the original F_0 scale, the natural waveform **NAT** was also used for comparison.

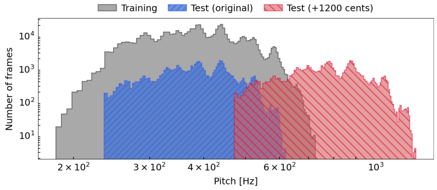
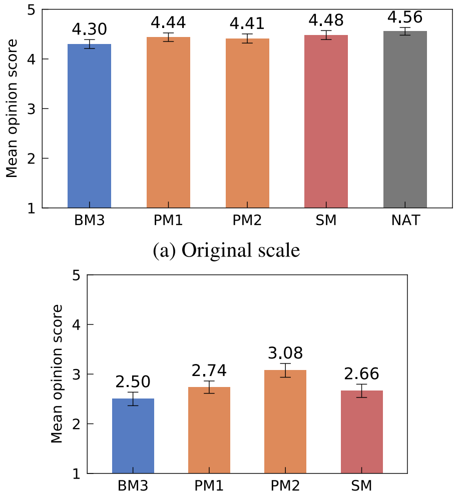


FIGURE 5. F_0 range of training data and test data at original scale, and upward-shifted test data in `F01`.

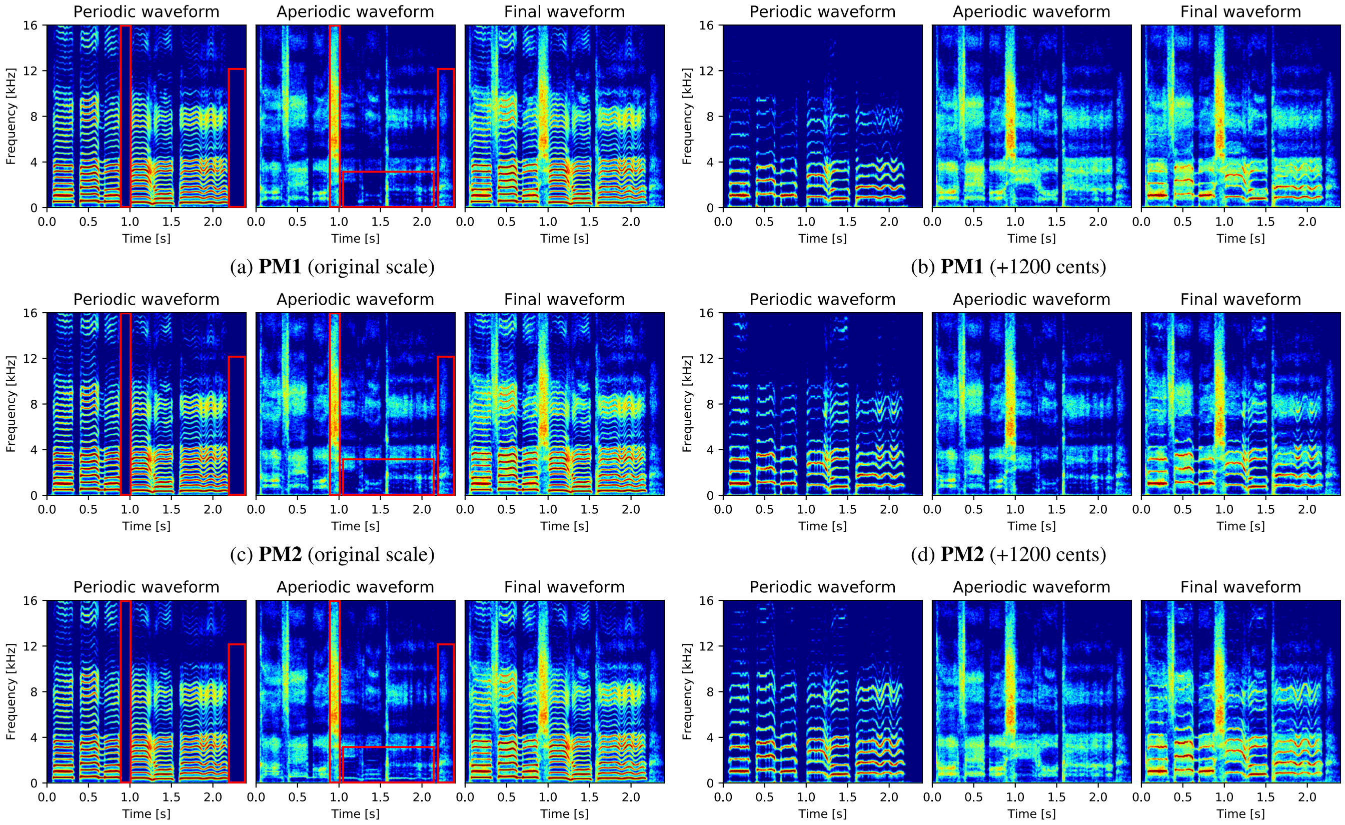


(a) Original scale

(b) +1200 cents

FIGURE 6. Subjective evaluation results obtained using single female-singer data with 95% confidence intervals. Top figure shows results obtained using original scale F_0. Bottom figure shows results obtained using F_0 upward-shifted by 1200 cents.

The results obtained using the original and upward-shifted F_0 are presented in Figs. 6(a) and 6(b), respectively. These figures show that **PM1**, **PM2**, and **SM** attained higher naturalness than **BM3**. This indicates that it is effective for the non-AR models to introduce a parallel or series structure. Here, examples of spectrograms for **PM1**, **PM2**, and **SM** are shown in Fig. 7. The highlighted boxes near 1.0 and 2.3 seconds in each figure represent parts of the unvoiced fricative "/s/" and the breath. It can be seen that the spectra of these unvoiced sounds only appeared in the output of the aperiodic generator. In addition, aperiodic components mixed in voiced sounds also appeared in other areas. These results indicate that the two sub-generators in the parallel model and the series model work on modeling the transformation from a sine wave and a noise sequence to periodic and aperiodic waveforms. Comparing the highlighted boxes in the lower right of Figs. 7(a), 7(b), and 7(d), the output waveform of



(a) **PM1** (original scale)

(b) **PM1** (+1200 cents)

(c) **PM2** (original scale)

(d) **PM2** (+1200 cents)

(e) **SM** (original scale)

(f) **SM** (+1200 cents)

FIGURE 7. Spectrograms of waveforms generated by PeriodNet. Three spectrograms of each system show periodic generator's output, aperiodic generator's output, and final output after sum of two signals.

the aperiodic generator in **SM** contained more harmonic components than in **PM1**. This suggests that the periodic waveforms used as the input of the aperiodic generator in **SM** may have leaked into the output of the aperiodic generator because the output waveform of the periodic generator was fed into the aperiodic generator. It should be noted that some harmonic components were also included slightly in the aperiodic waveform generated by **PM1** and **PM2** because the periodic and aperiodic waveforms were not explicitly decomposed in the training stage.

For the upward-shifted F_0 scenario, proposed systems **PM1**, **PM2**, and **SM** outperformed **BM3**. This result indicated that the proposed model structures were effective in generating waveforms with pitches outside the range of training data. Comparing **PM1** with **SM**, **PM1** had a slightly better score than **SM**. For **SM**, since the aperiodic generator should generate an aperiodic waveform in synchronization with a periodic waveform generated by the periodic generator, it was more affected by the input pitch than **PM1**. Comparing the two types of parallel models, **PM2** outperformed **PM1** significantly. The waveform samples generated by **PM1** tended to contain more noisy aperiodic waveforms than those generated by **PM2**. In the case of using the upward-shifted F_0, the aperiodic generator in **PM2** can avoid generating excessive aperiodic components, unlike **PM1**, as shown in Figs. 7(b) and 7(d). For **PM1**, although the periodic and aperiodic components in the singing voice waveforms were modeled separately, both aperiodic generators were conditioned on auxiliary features, including F_0. It was assumed that **PM1** could not generate proper aperiodic waveforms when these vocoders took out-of-range F_0. In contrast, since the aperiodic generator in **PM2** does not depend on either periodic signals or F_0, **PM2** showed the most robustness for unseen F_0 outside the range of the training data.

Proposed models (**PM1**, **PM2**, and **SM**) achieved a real-time factor (RTF) of 0.081 on an NVIDIA GTX 1080 Ti. Therefore, PeriodNet can generate high-quality 48 kHz singing voice waveforms more than 10 times faster than in real time.

V. EXPERIMENT 2

This section compares PeriodNet with systems that train using pre-decomposed periodic and aperiodic waveforms. This comparison aims to evaluate the performance of PeriodNet in modeling speech waveforms while appropriately separating periodic and aperiodic components during the training process. We used the harmonic plus residual model (HPR) [40] for periodic/aperiodic decomposition. The HPR defines a noise component in the harmonic plus noise model [57] as the difference between an original waveform and harmonic components.

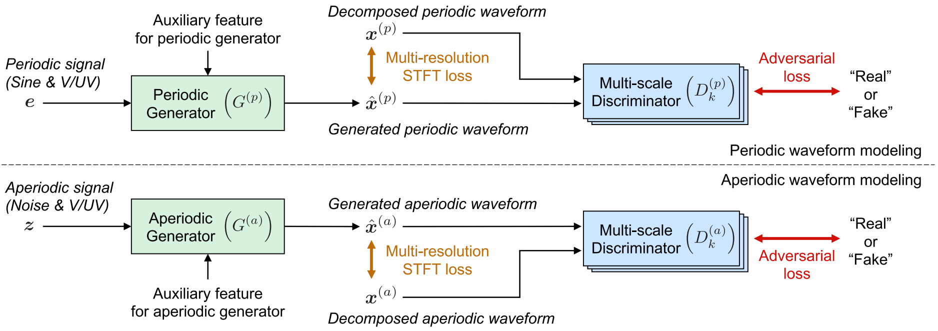

Auxiliary feature
for periodic generator
Decomposed periodic waveform
$x^{(p)}$
Periodic signal
(Sine & V/UV)
e
Periodic Generator $\left(G^{(p)}\right)$
Multi-resolution
STFT loss
$\hat{x}^{(p)}$
Generated periodic waveform
Multi-scale Discriminator $\left(D_k^{(p)}\right)$
Adversarial loss
"Real" or "Fake"
Periodic waveform modeling
Aperiodic waveform modeling
Aperiodic signal
(Noise & V/UV)
z
Aperiodic Generator $\left(G^{(a)}\right)$
Generated aperiodic waveform
$\hat{x}^{(a)}$
Multi-resolution
STFT loss
Multi-scale Discriminator $\left(D_k^{(a)}\right)$
"Real" or "Fake"
Adversarial loss
Auxiliary feature
for aperiodic generator
$x^{(a)}$
Decomposed aperiodic waveform

FIGURE 8. Overview of non-AR waveform generative model framework incorporating explicit periodic/aperiodic decomposition, denoted as HPR/HPR in Section V-B. Periodic and aperiodic generators are separately trained using pre-decomposed periodic and aperiodic waveforms by HPR [40] as target waveforms.

A. OVERVIEW OF METHODS FOR COMPARISON WITH EXPLICIT PERIODIC/APERIODIC DECOMPOSITION

We built a system incorporating explicit periodic/aperiodic decomposition techniques in the PeriodNet parallel model. Figure 8 shows a system incorporating an explicit periodic/aperiodic decomposition technique. Unlike the PeriodNet parallel model shown in Fig. 3, a periodic generator $G^{(p)}$ and an aperiodic generator $G^{(a)}$ were trained separately using pre-decomposed periodic and aperiodic waveforms $\boldsymbol{x}^{(p)}$, $\boldsymbol{x}^{(a)}$ as the target waveforms instead of the natural waveform $\boldsymbol{x}$. Moreover, we used spectral parameters extracted from decomposed waveforms as a part of the auxiliary features. While the standard aperiodic measure represents aperiodic components as the ratio of power to a speech signal, these parameters, extracted from the decomposed periodic and aperiodic waveforms, can represent the pure change in periodic and aperiodic components directly. Specifically, 50-dimensional mel-cepstrum coefficients extracted from decomposed periodic and aperiodic waveforms were used for the periodic and aperiodic generators, respectively, instead of the standard mel-cepstrum coefficients and aperiodicity measures described in Section IV-A.

As shown in Fig 8, by introducing two discriminators $D^{(p)}$ and $D^{(a)}$, both generators can be trained in the same fashion as PeriodNet described in Section III-B. The loss functions for the periodic and aperiodic generators $\mathcal{L}_G(G^{(p)}, D^{(p)})$ and $\mathcal{L}_G(G^{(a)}, D^{(a)})$ are given by

$$\mathcal{L}_G(G^{(p)}, D^{(p)}) = \frac{1}{M}\sum_{m=1}^{M} \mathcal{L}_{sp}^{<m>}(G^{(p)}) + \frac{\lambda_{adv}^{(p)}}{K}\sum_{k=1}^{K} \mathcal{L}_{adv}(G^{(p)}, D_k^{(p)}), \tag{9}$$

$$\mathcal{L}_G(G^{(a)}, D^{(a)}) = \frac{1}{M}\sum_{m=1}^{M} \mathcal{L}_{sp}^{<m>}(G^{(a)}) + \frac{\lambda_{adv}^{(a)}}{K}\sum_{k=1}^{K} \mathcal{L}_{adv}(G^{(a)}, D_k^{(a)}), \tag{10}$$

where $\lambda_{adv}^{(p)}$ and $\lambda_{adv}^{(a)}$ denote hyperparameters of the adversarial loss terms. $\lambda_{adv}^{(p)}$ and $\lambda_{adv}^{(a)}$ were set to 4.0.

B. SUBJECTIVE EVALUATION

This experiment used natural waveforms and waveforms decomposed by HPR belonging to a female singer F01. We compared the PeriodNet parallel model and several variants of systems incorporating HPR as follows.

- **ORG/ORG**: The parallel model without any decomposition techniques. This is the same system as **PM1** in Section IV.
- **HPR/ORG**: A system that independently models periodic and aperiodic waveforms decomposed by two separate generators. Both generators were conditioned by the same auxiliary features extracted from the natural waveforms.
- **ORG/HPR**: A system that models natural waveforms with the PeriodNet parallel model. Unlike **ORG/ORG**, the periodic and aperiodic generators were conditioned by different auxiliary features that contained spectral parameters extracted from decomposed periodic and aperiodic waveforms, respectively.
- **HPR/HPR**: A system that combines **HPR/ORG** and **ORG/HPR**, as shown in Fig. 8. Two separate generators, conditioned by auxiliary features with spectral parameters extracted from decomposed waveforms, independently model decomposed periodic and aperiodic waveforms.

For **ORG/HPR** and **HPR/HPR**, the smoothed spectra were calculated by analyzing decomposed periodic and aperiodic waveforms using WORLD, followed by mel-cepstral coefficients being extracted from both spectra. A continuous log F_0 extracted from the natural waveforms and V/UV binary code were used for all systems. The periodic and aperiodic generators in **HPR/ORG**, **ORG/HPR**, and **HPR/HPR** used the same architectures as those in **ORG/ORG**.

We evaluated the naturalness of the synthesized waveforms by conducting a five-point MOS test. The natural waveform **NAT** was also used for comparison. Fifteen subjects evaluated 15 phrases randomly selected from 10 test songs for each method.

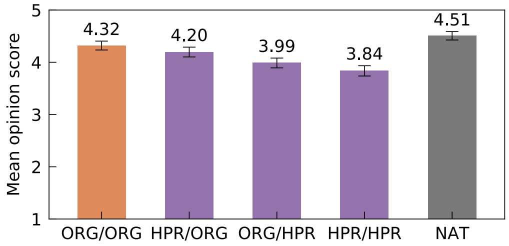


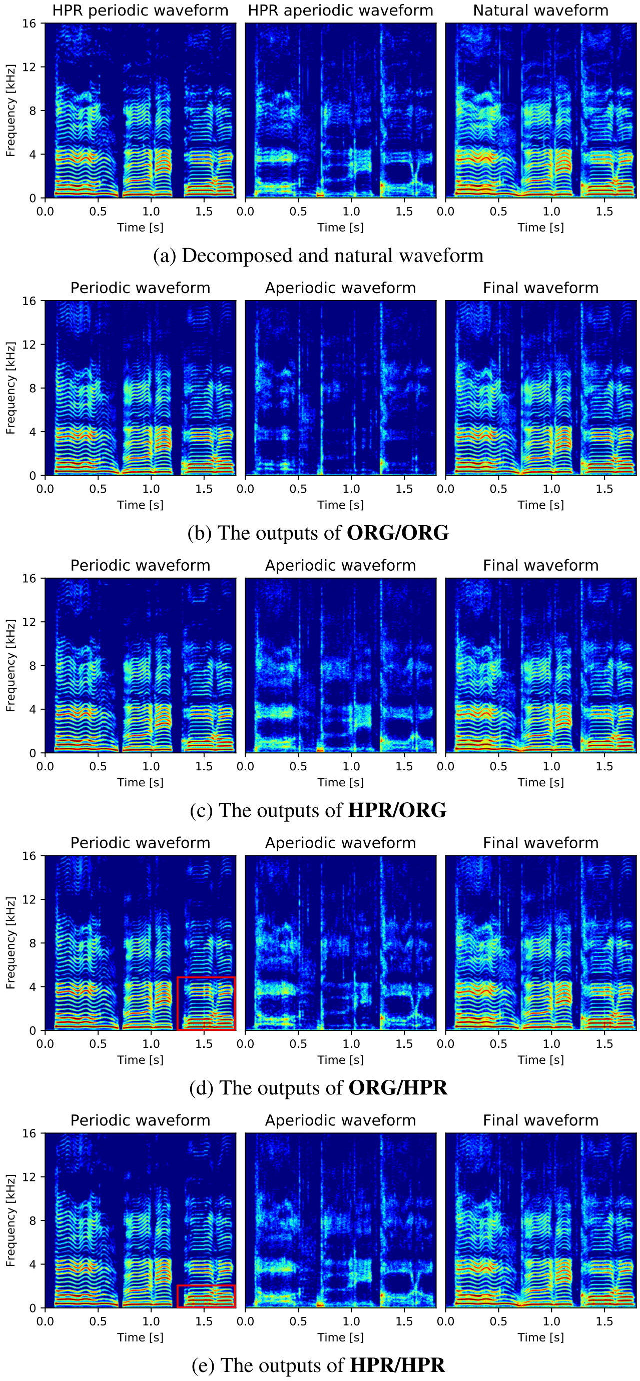
FIGURE 9. Subjective evaluation results with 95% confidence intervals.

Figure 9 shows the results of the subjective evaluation. From this figure, **ORG/ORG** had a higher MOS value compared with **HPR/ORG**. **HPR/ORG** may have been affected by decomposition errors because the target waveforms of **HPR/ORG** were waveforms decomposed using HPR. In addition, **ORG/HPR** and **HPR/HPR** showed lower MOS values than **ORG/ORG** and **HPR/ORG**. This indicates the difficulty of extracting appropriate acoustic features from decomposed waveforms that contain decomposition errors and using them to train the generator. Moreover, this factor affects not only the training process but also the inference process. Here, the log-amplitude spectrograms of the natural waveform, periodic and aperiodic waveforms obtained by HPR from the natural waveform, and the outputs of $G^{(p)}(\boldsymbol{e})$, $G^{(a)}(z)$, and $G^{(p)}(\boldsymbol{e}) + G^{(a)}(z)$ for each method are shown in Fig. 10. It can be seen from Figs. 10(d) and 10(e) that **ORG/HPR** and **HPR/HPR** had unnatural spectra, such as around the 1.5-second point of the periodic waveform, which degraded the naturalness. Figure 10(a) showed that HPR aperiodic waveform still had some harmonic structures due to the decomposed error. Consequently, the output of $G^{(a)}(z)$ in **HPR/ORG**, **ORG/HPR** and **HPR/HPR** has more harmonics structures than that in **ORG/ORG**. In contrast, **ORG/ORG** was able to model the aperiodic components without using the periodic/periodic decomposition method, as shown in Fig. 10(b). In addition to this result, the MOS value of **ORG/ORG** was closest to that of **NAT**. These results indicate that PeriodNet with the proposed structures can model periodic and aperiodic components appropriately without any explicit decomposition process for these components.

VI. EXPERIMENT 3

In this section, we examined the effectiveness of the proposed method when using different singers' datasets. We used one other female singer (F02) and two male singers (M01 and M02). The dataset of each singer consisted of 70 Japanese children's songs, which were the same songs as F01, while the key and tempo of some of the songs differed for each singer.

To investigate robustness with the singers, we conducted MOS tests. **BM1**, **BM3**, **PM1**, **PM2**, and **SM**, described in Section IV-A, were used for comparison. These models were individually trained for each singer. With **SM** for the male singers, we found that the aperiodic generator tended to predict some of the periodic components instead of the periodic generator. Therefore, we trained only the periodic generator for the first 10K iterations, followed by both sub-generators in **SM** for the male singers. **BM3**, **PM1**, **PM2**, and **SM** were evaluated by using both the original scale and upward-shifted F_0. The F_0 range of the three singers is shown in Fig. 11. The range that could be covered by the training data of M01 and M02 was wider than that of F01 and F02. Therefore, for M01 and M02, the amount of F_0 shifting was set to 1600 cents instead of 1200 cents. In the experiment



(a) Decomposed and natural waveform

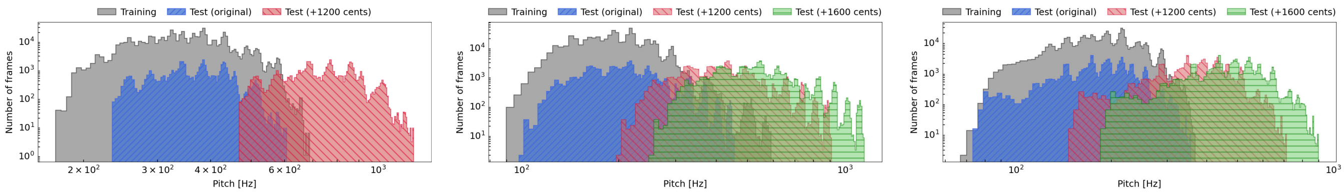
(b) The outputs of **ORG/ORG**

(c) The outputs of **HPR/ORG**

(d) The outputs of **ORG/HPR**

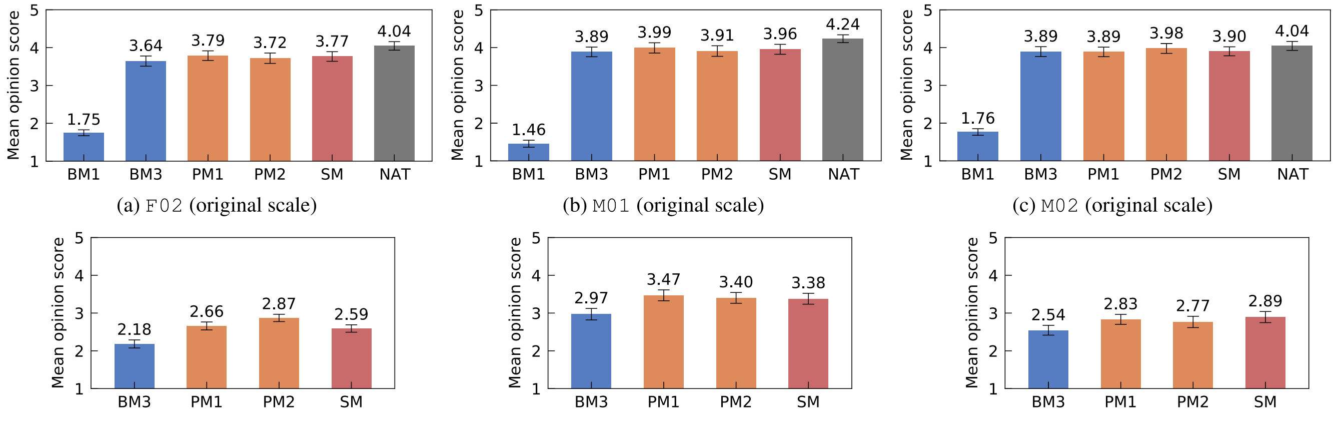
(e) The outputs of **HPR/HPR**

FIGURE 10. Spectrograms of waveforms generated by non-AR parallel model.



(a) F02 (b) M01 (c) M02

FIGURE 11. *F_0* **range of singing voice data using subjective evaluation described in Section VI.**



(a) F02 (original scale) (b) M01 (original scale) (c) M02 (original scale)

(d) F02 (+1200 cents) (e) M01 (+1600 cents) (f) M02 (+1600 cents)

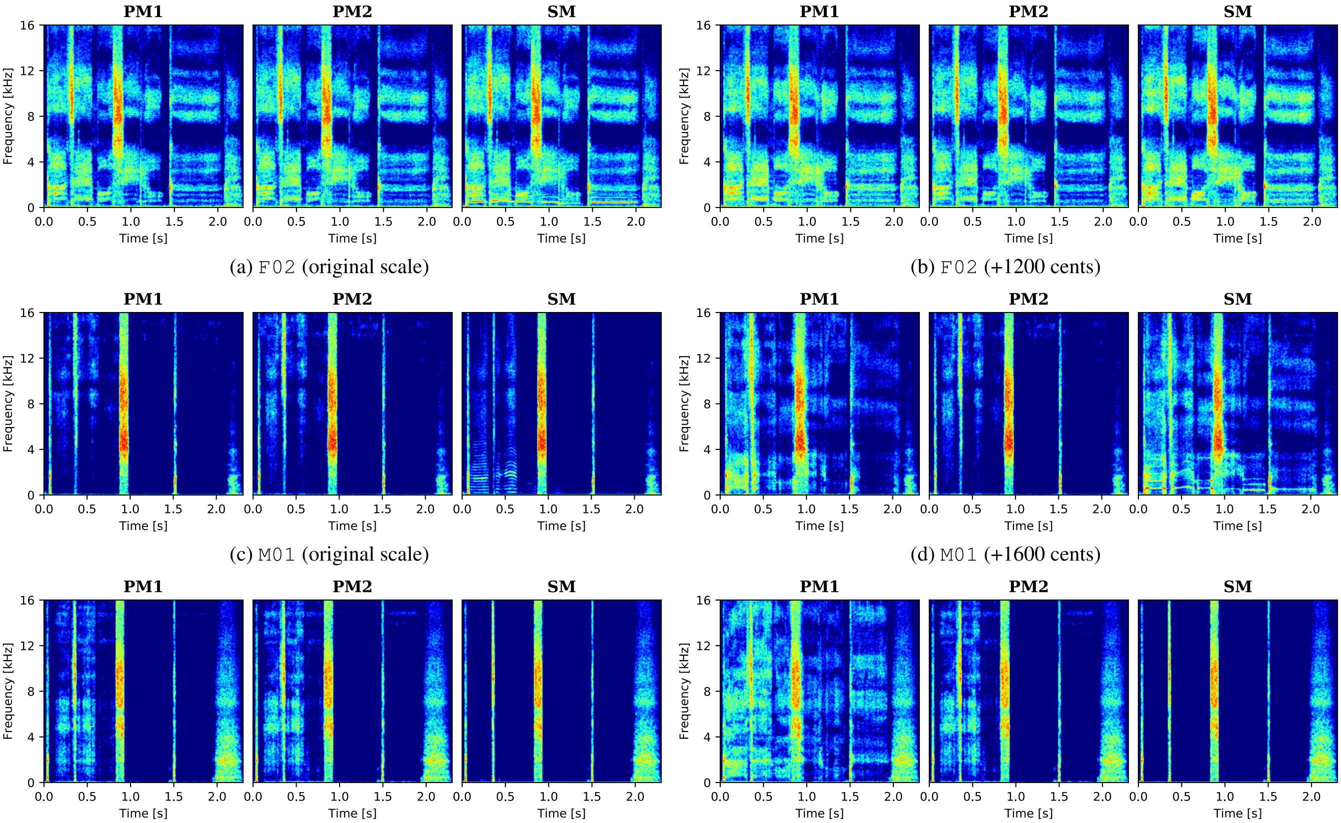
FIGURE 12. Subjective evaluation results obtained using different singers' data with 95% confidence intervals. Upper row shows case using original *F_0* **extracted from test data, and lower row shows case using** *F_0* **upward-shifted by 1200 cents for female singers or 1600 cents for male singers.**

with the original F_0 scale, the natural waveform **NAT** was also used for comparison. The participants were 11 native Japanese speakers for each MOS test, and each participant evaluated 12 phrases randomly selected from the test data.

Figure 12 shows the results of the MOS test. In the experiments using the original F_0, the systems that take periodic signals as input outperformed **BM1** significantly for the case of all singers. Proposed systems **PM1**, **PM2**, and **SM** showed slightly better results than **BM3**. These results were similar to those for F01.

In the experiments using upward-shifted F_0, the proposed systems showed a better result than **BM3**; however, the trend between **PM1**, **PM2**, and **SM** was different for each singer. While **PM2** achieved the best MOS score for F02, there was no significant difference between **PM1**, **PM2**, and **SM** for M01 and M02. Here, examples of spectrograms of the aperiodic waveforms generated by the aperiodic generator in each proposed system are shown in Fig. 13. For F02, the outputs of the aperiodic generator in **PM1**, **PM2**, and **SM** contained more aperiodic components than that in the case of male singers. By focusing on spectrograms below 1 kHz, the generated aperiodic waveform of **SM** had clearer periodic components than **PM1** and **PM2** due to the periodic waveforms leaking into the output of the periodic generator. In addition, the aperiodic components were excessively emphasized for **PM1** and **SM** under the influence of the upward shift of F_0. Since these phenomena are similar to F01, the results were also similar to F01. For the male singers, although the aperiodic generators in **PM1**, **PM2**, and **SM** were able to model conspicuous aperiodic components such as unvoiced consonants and breath, they did not capture many aperiodic components mixed in voiced sounds, unlike the case of the female singers. This difference in trend was caused by the fact that the pitch of the male singing voice is lower than that of the female singing voice. Aperiodic components in voiced sounds appear in the spectral valleys between the harmonic components. For the waveforms with a low pitch, it is expected to be difficult to model these components because the intervals between harmonic components tend to be narrower. In addition, for M02, the aperiodic components within the voiced sound were not seen much in the output of the aperiodic generator in **SM**, as shown in Figs. 13(e) and 13(f). This means that these aperiodic components were modeled by the periodic generator, which depends on the given periodic signal and F_0, in place of the aperiodic generator. In fact, comparing the spectrograms of M01 and M02 when using upward-shifted F_0 in Fig. 14, the output of the periodic generator of **SM** in M02 contained a few aperiodic components, especially in the area between the harmonics below 4 kHz. Thus, the decomposition performance of PeriodNet is influenced by the characteristics of the singers.

Another reason for the different trends between the subjective impressions of the male and female singers is that the generated waveforms of the male singers sometimes contained perceptual artifacts, which may have been caused



(a) F02 (original scale)

(b) F02 (+1200 cents)

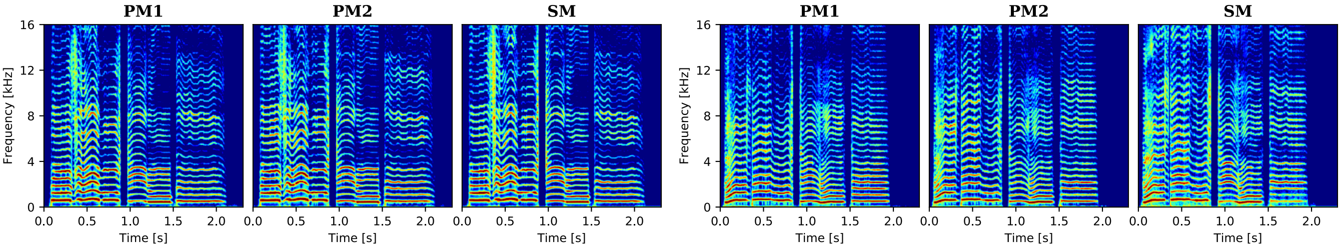
(c) M01 (original scale)

(d) M01 (+1600 cents)

(e) M02 (original scale)

(f) M02 (+1600 cents)

FIGURE 13. Spectrograms of aperiodic waveform generated by aperiodic generator in each PeriodNet system. All figures indicate spectrogram of generated waveform of same test phrase.



(a) M01 (+1600 cents)

(b) M02 (+1600 cents)

FIGURE 14. Spectrograms of periodic waveforms generated by periodic generator in each PeriodNet system. Both figures correspond to spectrograms of generated aperiodic waveforms, as shown in Figs. 13(d) and 13(f), respectively.

by a sine-based input signal. This is because the male voice has more low-pitch voices where the shape of the input signal has a more significant perceptual effect [58], [59]. The generated waveforms of **PM1** and **SM** had more aperiodic components than that of **PM2** in the case of using the upward-shifted F_0, as shown in Figs. 13(d), 13(f), and 14(b). Since this rather led to a relative reduction in the perception of such artifacts, **PM1**, **PM2**, and **SM** obtained similar naturalness for the male singers.

It has been reported that perceptual artifacts tend to occur when generating the waveform of a male speaker with other non-AR waveform generative models that use sine-based input signals [34]. The authors of [60] introduced a quasi-periodic cyclic noise signal based on the convolution of a pulse train and exponentially decaying Gaussian noise sequence to tackle this problem. However, it may not be appropriate to use a quasi-periodic signal instead of a sine-based periodic signal for PeriodNet because PeriodNet is a method that focuses on the presence or absence of the autocorrelation of input signals. To generate high-quality waveforms for male singers, further study using other kinds of input signals is a topic of future work.

VII. CONCLUSION

We proposed a novel non-AR neural waveform generative model with a structure separating periodic and aperiodic components in speech waveforms called "PeriodNet." PeriodNet consists of two sub-generators connected in parallel or in series that take a sine-based input signal and a Gaussian noise signal, respectively, and it represents a speech waveform as the sum of the outputs of both sub-generators. Since these input signals have different characteristics in terms of autocorrelation, the two sub-generators can model periodic and aperiodic components in speech waveforms without any explicit decomposition techniques. In particular, the proposed model structures bring the advantage of robustness to input pitch to PeriodNet. Thus, PeriodNet is highly suited for the vocoder in SVS systems.

The experimental results showed that PeriodNet was able to generate high-fidelity singing voice waveforms and improve the ability to generate waveforms with a pitch outside the range of the training data. Compared with systems trained using pre-decomposed periodic and aperiodic waveforms, PeriodNet makes it possible to model periodic and aperiodic components appropriately without explicit decomposition. The results obtained using several singer's data indicated that it was more challenging to model the periodic and aperiodic components of the male singers' waveforms than those of the female singers. Further investigating other periodic signals and improving the training criteria are topics for future work. Moreover, we intend to investigate whether proposed model structures can help improve performance in other types of neural waveform generative models.

In this paper, we focused on modeling singing voice waveforms. In SVS systems, vocoder parameter representations such as mel-cepstrum and F_0 are mainly used because accurate pitch control is required [4]–[7], whereas some systems use a combination of mel-spectrogram and F_0 [61]. Evaluating the proposed models with different acoustic feature representations is also included in future work. Furthermore, comparison with other types of recent neural waveform generative models and different kinds of waveform signals such as speech and musical instrument sound is also an important task for future work to further understand the performance of PeriodNet.

REFERENCES

[1] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Commun.*, vol. 51, no. 11, pp. 1039–1064, Nov. 2009.

[2] H. Ze, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 7962–7966.

[3] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Proc. Interspeech*, Aug. 2017, pp. 4004–4010.

[4] K. Oura, A. Mase, T. Yamada, S. Muto, Y. Nankaku, and K. Tokuda, "Recent development of the HMM-based singing voice synthesis system—Sinsy," in *Proc. ISCA SSW*, 2010, pp. 211–216.

[5] Y. Hono, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Sinsy: A deep neural network-based singing voice synthesis system," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 2803–2815, Aug. 2021.

[6] M. Blaauw and J. Bonada, "A neural parametric singing synthesizer," in *Proc. Interspeech*, Aug. 2017, pp. 4001–4005.

[7] Y. Hono, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Singing voice synthesis based on generative adversarial networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 6955–6959.

[8] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, "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, pp. 187–207, 1999.

[9] 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. 99, no. 7, pp. 1877–1884, 2016.

[10] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," in *Proc. ISCA SSW*, 2016, p. 125.

[11] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. Courville, and Y. Bengio, "SampleRNN: An unconditional end-to-end neural audio generation model," in *Proc. ICLR*, 2017.

[12] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker-dependent WaveNet vocoder," in *Proc. Interspeech*, Aug. 2017, pp. 1118–1122.

[13] A. Gibiansky, S. Arik, G. Diamos, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou, "Deep voice 2: Multi-speaker neural text-to-speech," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 2962–2970.

[14] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 4779–4783.

[15] K. Kobayashi, T. Hayashi, A. Tamamori, and T. Toda, "Statistical voice conversion with WaveNet-based waveform generation," in *Proc. Interspeech*, Aug. 2017, pp. 1138–1142.

[16] 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," 2018, *arXiv:1802.08435*. [Online]. Available: http://arxiv.org/abs/1802.08435

[17] Z. Jin, A. Finkelstein, G. J. Mysore, and J. Lu, "Fftnet: A real-time speaker-dependent neural vocoder," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 2251–2255.

[18] A. van den Oord, Y. Li, and I. Babuschkin, "Parallel WaveNet: Fast high-fidelity speech synthesis," in *Proc. ICML*, 2018, pp. 3918–3926.

[19] W. Ping, K. Peng, and J. Chen, "ClariNet: Parallel wave generation in end-to-end text-to-speech," in *Proc. ICML*, 2019.

[20] 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. (ICASSP)*, May 2019, pp. 3617–3621.

[21] S. Kim, S.-g. Lee, J. Song, J. Kim, and S. Yoon, "FloWaveNet: A generative flow for raw audio," 2018, *arXiv:1811.02155*. [Online]. Available: http://arxiv.org/abs/1811.02155

[22] W. Ping, K. Peng, K. Zhao, and Z. Song, "WaveFlow: A compact flow-based model for raw audio," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 7706–7716.

[23] H. Kim, H. Lee, W. Hyun Kang, S. Jun Cheon, B. Jin Choi, and N. Soo Kim, "WaveNODE: A continuous normalizing flow for speech synthesis," 2020, *arXiv:2006.04598*. [Online]. Available: http://arxiv.org/abs/2006.04598

[24] N.-Q. Wu and Z.-H. Ling, "WaveFFJORD: FFJORD-based vocoder for statistical parametric speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 7214–7218.

[25] 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. (ICASSP)*, May 2020, pp. 6199–6203.

[26] 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. Adv. Neural Inf. Process. Syst.*, 2019, pp. 14910–14921.

[27] J. Yang, J. Lee, Y. Kim, H.-Y. Cho, and I. Kim, "VocGAN: A high-fidelity real-time vocoder with a hierarchically-nested adversarial network," in *Proc. Interspeech*, Oct. 2020, pp. 200–204.

[28] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020.

[29] 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 (SLT)*, Jan. 2021, pp. 492–498.

[30] M. Bińkowski, J. Donahue, S. Dieleman, A. Clark, E. Elsen, N. Casagrande, L. C. Cobo, and K. Simonyan, "High fidelity speech synthesis with adversarial networks," in *Proc. ICLR*, 2020.

[31] K. Peng, W. Ping, Z. Song, and K. Zhao, "Non-autoregressive neural text-to-speech," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 7586–7598.

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

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

[34] 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, Nov. 2019.

[35] K. Oura, K. Nakamura, K. Hashimoto, Y. Nankaku, and K. Tokuda, "Deep neural network based real-time speech vocoder with periodic and aperiodic inputs," in *Proc. 10th ISCA Workshop Speech Synth. (SSW)*, Sep. 2019, pp. 13–18.

[36] 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, Feb. 2021.

[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, Jan. 2021.

[38] X. Serra and J. Smith, "Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition," *Comput. Music J.*, vol. 14, no. 4, pp. 12–24, 1990.

[39] P. Zubrycki and A. Petrovsky, "Accurate speech decomposition into periodic and aperiodic components based on discrete harmonic transform," in *Proc. 15th Eur. Signal Process. Conf.*, Sep. 2007, pp. 2336–2340.

[40] *SMS-Tools: Sound Analysis/Synthesis Tools for Music Applications*. Accessed: 2021. [Online]. Available: https://github.com/MTG/sms-tools

[41] Y. Hono, S. Takaki, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Periodnet: A non-autoregressive waveform generation model with a structure separating periodic and aperiodic components," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 6049–6053.

[42] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1×1 convolutions," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 10215–10224.

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

[44] L. Juvela, V. Tsiaras, B. Bollepalli, M. Airaksinen, J. Yamagishi, and P. Alku, "Speaker-independent raw waveform model for glottal excitation," in *Proc. Interspeech*, Sep. 2018, pp. 2012–2016.

[45] Y. Cui, X. Wang, L. He, and F. K. Soong, "A new glottal neural vocoder for speech synthesis," in *Proc. Interspeech*, Sep. 2018, pp. 2017–2021.

[46] L. Juvela, B. Bollepalli, V. Tsiaras, and P. Alku, "GlotNet—A raw waveform model for the glottal excitation in statistical parametric speech synthesis," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 6, pp. 1019–1030, Jun. 2019.

[47] L. Juvela, B. Bollepalli, J. Yamagishi, and P. Alku, "Waveform generation for text-to-speech synthesis using pitch-synchronous multi-scale generative adversarial networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 6915–6919.

[48] J.-M. Valin and J. Skoglund, "LPCNET: Improving neural speech synthesis through linear prediction," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 5891–5895.

[49] L. Juvela, B. Bollepalli, J. Yamagishi, and P. Alku, "GELP: GAN-excited linear prediction for speech synthesis from mel-spectrogram," in *Proc. Interspeech*, Sep. 2019, pp. 694–698.

[50] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2794–2802.

[51] K. Sawada, C. Asai, K. Hashimoto, K. Oura, and K. Tokuda, "The NITech text-to-speech system for the Blizzard challenge 2016," in *Proc. Blizzard Challenge Workshop*, 2016, pp. 1–6.

[52] *Pulse Code Modulation (PCM) of Voice Frequencies*, ITU-T Recommendation G.711, Int. Telecommun. Union, Geneva, Switzerland, 1988.

[53] T. Salimans and D. P. Kingma, "Weight normalization: A simple reparameterization to accelerate training of deep neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 901–909.

[54] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, "On the variance of the adaptive learning rate and beyond," in *Proc. ICLR*, Apr. 2020.

[55] *REAPER: Robust Epoch and Pitch Estimator*. Accessed: 2021. [Online]. Available: https://github.com/google/REAPER

[56] Y. Hono. *PeriodNet Demo*. Accessed: 2021. [Online]. Available: https://www.sp.nitech.ac.jp/~hono/demos/access2021/

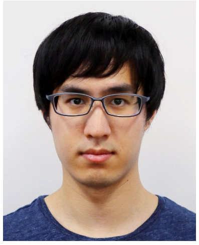
[57] A. Syrdal, Y. Stylianou, L. Garrison, A. Conkie, and J. Schroeter, "TD-PSOLA versus harmonic plus noise model in diphone based speech synthesis," in *Proc. ICASSP*, vol. 1, 1998, pp. 273–276.

[58] A. E. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels," *J. Acoust. Soc. Amer.*, vol. 49, no. 2B, pp. 583–590, Feb. 1971.

[59] T. Raitio, L. Juvela, A. Suni, M. Vainio, and P. Alku, "Phase perception of the glottal excitation and its relevance in statistical parametric speech synthesis," *Speech Commun.*, vol. 81, pp. 104–119, Jul. 2016.

[60] X. Wang and J. Yamagishi, "Using cyclic noise as the source signal for neural source-filter-based speech waveform model," in *Proc. Interspeech*, Oct. 2020, pp. 1992–1996.

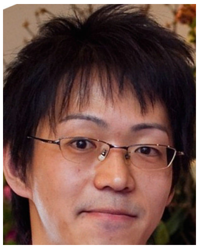
[61] J. Chen, X. Tan, J. Luan, T. Qin, and T.-Y. Liu, "HiFiSinger: Towards high-fidelity neural singing voice synthesis," 2020, *arXiv:2009.01776*. [Online]. Available: http://arxiv.org/abs/2009.01776



YUKIYA HONO received the B.E. and M.E. degrees in computer science from Nagoya Institute of Technology, Nagoya, Japan, in 2017 and 2019, respectively, where he is currently pursuing the Ph.D. degree. From July to August 2019, he was an Intern at Microsoft Development Company Ltd., Tokyo, Japan. He was a Visiting Researcher with The University of Edinburgh, U.K., from October 2019 to December 2019 and The University of Sheffield, U.K., from January 2020 to February 2020. His research interests include statistical speech synthesis, singing voice synthesis, and machine learning. He is a member of the Acoustical Society of Japan (ASJ). He was a recipient of the 18th Student Presentation Award from ASJ, the 2019 Information and Communication Engineers (IEICE) Tokai Section Student Award, and the 2021 IEEE Nagoya Section Excellent Student Award.

SHINJI TAKAKI (Member, IEEE) received the B.Eng. degree in computer science and the M.Eng. and Ph.D. degrees in scientific and engineering simulation from Nagoya Institute of Technology (NITech), Nagoya, Japan, in 2009, 2011, and 2014, respectively. From 2013 to 2014, he was a Visiting Researcher with The University of Edinburgh, Edinburgh, U.K. From 2014 to 2019, he was a Project Researcher with the National Institute of Informatics, Chiyoda, Japan. Since 2019, he has been a Project Researcher with NITech. His research interests include statistical machine learning and speech synthesis. His awards include the 1st Best Student Presentation Award from the Acoustical Society of Japan (ASJ), the Vice-President Prize from NITech, the IPSJ Yamashita SIG Research Award, and the Awaya Prize Young Researcher Award from ASJ.

KEI HASHIMOTO (Member, IEEE) received the B.E., M.E., and Ph.D. degrees in computer science, computer science and engineering, and scientific and engineering simulation from Nagoya Institute of Technology, Nagoya, Japan, in 2006, 2008, and 2011, respectively. From October 2008 to January 2009, he was an Intern Researcher at the National Institute of Information and Communications Technology (NICT), Kyoto, Japan. From April 2010 to March 2012, he was a Research Fellow of the Japan Society for the Promotion of Science (JSPS), Nagoya Institute of Technology. From May 2010 to September 2010, he was a Visiting Researcher at The University of Edinburgh and Cambridge University. From April 2012 to March 2017, he was a specially-appointed Assistant Professor at Nagoya Institute of Technology. From April 2017 to December 2018, he was a specially-appointed Associate Professor at Nagoya Institute of Technology, where he is currently an Associate Professor. His research interests include statistical speech synthesis and speech recognition. He is a member of the IEICE and the Acoustical Society of Japan.



KEIICHIRO OURA received the Ph.D. degree in computer science and engineering from Nagoya Institute of Technology, Nagoya, Japan, in 2010. He was a specially-appointed Assistant Professor at Nagoya Institute of Technology, from April 2010 to March 2017. From April 2017 to May 2020, he was a specially-appointed Associate Professor at Nagoya Institute of Technology. He is currently a Project Associate Professor at Nagoya Institute of Technology and the CEO of the Techno-Speech, Inc. His research interests include statistical speech recognition and synthesis. He received the ISCSLP Best Student Paper Award, in 2008, the IPSJ Yamashita SIG Research Award, in 2010, the ASJ Itakura Award, in 2013, the IPSJ Kiyasu Special Industrial Achievement Award, in 2013, the ASJ Awaya Prize Young Researcher Award, in 2019, and the IPSJ Microsoft Faculty Award, in 2020. He is a member of the Acoustical Society of Japan and the Information Processing Society of Japan.

YOSHIHIKO NANKAKU (Member, IEEE) received the B.E. degree in computer science and the M.E. and Ph.D. degrees from the Department of Electrical and Electronic Engineering, Nagoya Institute of Technology, Nagoya, Japan, in 1999, 2001, and 2004, respectively. After a year as a Postdoctoral Fellow at Nagoya Institute of Technology, where he became an Associate Professor. He was a Visiting Researcher at the Department of Engineering, University of Cambridge, U.K., from May to October 2011. His research interests include statistical machine learning, speech recognition, speech synthesis, image recognition, and multi-modal interface. He is a member of the Institute of Electronics, Information and Communication Engineers and the Acoustical Society of Japan.

KEIICHI TOKUDA (Fellow, IEEE) received the B.E. degree in electrical and electronic engineering from Nagoya Institute of Technology, Nagoya, Japan, and the M.E. and Dr.Eng. degrees in information processing from the Tokyo Institute of Technology, Tokyo, Japan, in 1984, 1986, and 1989, respectively. From 1989 to 1996, he was a Research Associate at the Department of Electronic and Electric Engineering, Tokyo Institute of Technology. From 1996 to 2004, he was an Associate Professor at the Department of Computer Science, Nagoya Institute of Technology, where he is currently a Professor. He is also an Honorary Professor at The University of Edinburgh. He was an Invited Researcher at the ATR Spoken Language Translation Research Laboratories, Japan, from 2000 to 2013 and a Visiting Researcher at Carnegie Mellon University, from 2001 to 2002, and Google, from 2013 to 2014. He published over 80 journal papers and over 200 conference papers, and received six paper awards and three achievement awards. He acts as an organizer and a reviewer for many major speech conferences, workshops, and journals. His research interests include speech coding, speech synthesis and recognition, and statistical machine learning. He was a member of the Speech Technical Committee, IEEE Signal Processing Society, from 2000 to 2003, and the ISCA Advisory Council. He is a fellow of ISCA. He was an Associate Editor of IEEE Transactions on Audio, Speech and Language Processing.