High-Fidelity and Pitch-Controllable Neural Vocoder Based on Unified Source-Filter Networks

Reo Yoneyama, Yi-Chiao Wu, and Tomoki Toda, *Senior Member, IEEE*

***Abstract*—We introduce unified source-filter generative adversarial networks (uSFGAN), a waveform generative model conditioned on acoustic features, which represents the source-filter architecture in a generator network. Unlike the previous neural-based source-filter models in which parametric signal process modules are combined with neural networks, our approach enables unified optimization of both the source excitation generation and resonance filtering parts to achieve higher sound quality. In the uSFGAN framework, several specific regularization losses are proposed to enable the source excitation generation part to output reasonable source excitation signals. Both objective and subjective experiments are conducted, and the results demonstrate that the proposed uSFGAN achieves comparable sound quality to HiFi-GAN in the speech reconstruction task and outperforms WORLD in the F_0 transformation task. Moreover, we argue that the F_0-driven mechanism and the inductive bias obtained by source-filter modeling improve the robustness against unseen F_0 in training as shown by the results of experimental evaluations. Audio samples are available at our demo site at https://chomeyama.github.io/PitchControllableNeuralVocoder-Demo/.**

***Index Terms*—Speech synthesis, neural vocoder, source-filter model, unified source-filter networks.**

I. Introduction

SPEECH synthesis is a technology of generating speech waveforms on the basis of text or acoustic features. In particular, models conditioned on acoustic features are called vocoders. Vocoders have been widely adopted in many voice applications, such as text-to-speech (TTS), singing voice synthesis (SVS), and voice conversion (VC). In the applications, the quality of the final generated waveform strongly depends on the performance of the vocoder. Specifically, vocoders are required to generate speech of high sound quality in addition to functions for flexibly and independently controlling the generated speech in accordance with given acoustic features (e.g., F_0, timbre, and periodicity). Furthermore, vocoders should be robust to unseen data, such as unseen speakers and F_0, and continuously generate high-fidelity speech. For example, the vocoder of a multi-speaker TTS for few-shot voice cloning should be able to tackle a wide range of F_0 because of the frequent adaptations of arbitrary speakers since it is impractical to collect a corpus covering all unseen speakers, and even the utterances of the seen speakers may include out-of-range F_0 values. Moreover, SVS often requires a significant deviation of F_0 to generate a singing voice that transcends physical limitations. Therefore, the vocoder of SVS should have high robustness to unseen F_0 over an extensive range.

However, most vocoders do not meet the mentioned requirements. Specifically, conventional vocoders [1], [2] based on source-filter models [3], [4], [5] can flexibly control speech characteristics, but the quality of the generated speech is low because of their over-simplified speech production process. Recent high-fidelity neural vocoders [6], [7], [8], [9], [10], [11], [12], [13], [14], [15] lack the robustness to unseen data because of their purely data-driven training-manners. For example, the state-of-the-art neural vocoder, HiFi-GAN [12], fails to generate high-fidelity speech when the input features include F_0 values deviating from the F_0 range of the training data. Furthermore, compared with the conventional source-filter models, those neural vocoders have poorer interpretability and less flexible controllability of speech characteristics. One reasonable way for neural vocoders to satisfy the above requirements is to introduce source-filter modeling to obtain sufficient flexibility and inductive bias for human speech production. Several approaches [16], [17], [18], [19], [20], [21], [22], [23] have been investigated to combine the source-filter architecture with deep neural networks using signal-processing-based modules. However, there are several problems with incorporating the parametric (signal-processing-based) module and strong constraints into neural vocoders. For instance, the partial utilization of signal processing makes the optimization of the entire speech generation process difficult and degrades sound quality and F_0 controllability since the neural networks must compensate for the incomplete output of the parametric modules.

To achieve flexibility and interpretability of source-filter modeling while maintaining the high sound quality of neural vocoders, we propose a novel framework of source-filter modeling on a single neural network, significantly reducing the effects of the ad hoc designs. Unlike previous approaches that model either the source excitation generation part or the resonance filtering part on the basis of signal processing as described in Section II, our approach enables the simultaneous optimization

Manuscript received 3 January 2023; revised 19 July 2023; accepted 25 August 2023. Date of publication 11 September 2023; date of current version 20 October 2023. This work was supported in part by the JST, CREST under Grant JPMJCR19A3 and in part by the Japan Society for the Promotion of Science (JSPS) Grants-in-Aid for Scientific Research (KAKENHI) under Grant JP21H05054. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Hung-yi Lee. *(Corresponding author: Reo Yoneyama.)*

Reo Yoneyama and Yi-Chiao Wu are with the Graduate School of Informatics, Nagoya University, Nagoya 464-8601, Japan (e-mail: yoneyama.reo@g.sp.m.is.nagoya-u.ac.jp; yichiao.wu@g.sp.m.is.nagoya-u.ac.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.2023.3313410

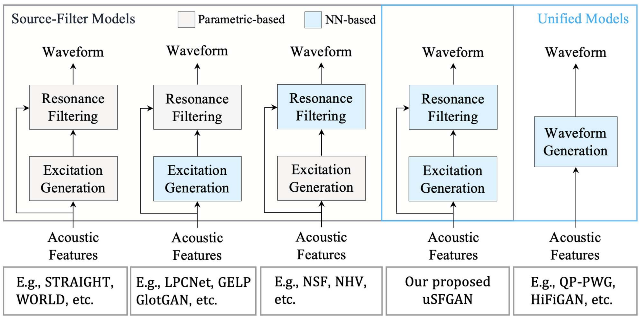

Fig. 1. Comparison of the architectures of conventional and neural vocoders in terms of the source-filter modeling.

of these two parts of the entire network, which leads to better sound quality. Our approach described in Section III is based on generative adversarial networks (GANs) [24] and separates the generator network into a source excitation generation network and a resonance filtering network using a regularization loss on the intermediate output of the network. To further obtain better F_0 robustness and F_0 controllability, we adopt a F_0-driven source excitation generation mechanism. Our experimental results shown in Section IV demonstrate that our proposed unified source-filter GAN (uSFGAN) achieves comparable sound quality to HiFi-GAN with much greater robustness against unseen F_0. Moreover, uSFGAN achieves better sound quality than other models presented in several previous works, such as WORLD [2], neural source-filter (NSF) [21], and quasi-periodic parallel waveGAN (QP-PWG) [25], [26] with high F_0 controllability. Furthermore, we demonstrate that uSFGAN models output reasonable source excitation signals via visualization. In this article, the previously proposed techniques proposed in [27], [28] are organized, improved, and evaluated in a unified manner. Additionally, we newly considered another network structure and input acoustic features, and thoroughly assessed their behaviors with more detailed experimental evaluations than our previous works. We provide the code of our model at https://github.com/chomeyama/HN-UnifiedSourceFilterGAN.

II. Related Work

In this section, we systematically introduce previous studies on vocoders based on the framework of the source-filter architecture [5]. The advantages and disadvantages of each approach are discussed. Fig. 1 shows the architectures in the previous studies and our proposed method.

A. Conventional Source-Filter Model

The source-filter architecture [5] is based on the idea that the human speech production process can be approximated by the modulation of a source excitation generated by vocal fold vibrations and a spectral filter model of vocal tract resonances. The assumption of independence between the two parts provides us with high interpretability and flexible controllability of speech characteristics such as F_0, timbre, and aperiodicity. The conventional source-filter-based vocoders such as STRAIGHT [1] and WORLD [2] achieve flexible controllability of speech characteristics while maintaining reasonable sound quality. Both models involve several assumptions to simplify the mathematical source-filter modeling of the speech production process. For example, these vocoders make assumptions based on prior knowledge, such as time-invariant linear filter and stationary Gaussian process. The source signals are modeled by a mixed excitation source that switches between pulse trains and white noise as it switches between voiced and unvoiced intervals, representing periodicity as a binary series of voiced or unvoiced parts. However, the mixed excitation source modeling loses detailed temporal and phase information of the original speech, which often deteriorates the sound quality of the synthesized speech. Because of the simplified and ad hoc mathematical modeling, the conventional source-filter models achieve low sound quality.

B. Neural Vocoders Based on Generative Models

Because of the more powerful modeling capacity of current neural networks, neural vocoders have markedly improved the naturalness of synthesized speech. WaveNet [6], which recursively predicts samples at sample level with the dilated convolution neural network (DCNN), has shown an impressively high sound quality. WaveNet, originally designed for text-to-speech (TTS) applications, was initially conditioned on linguistic features. It is also possible to condition WaveNet on acoustic features, similar to conventional vocoders. To replace traditional vocoders with WaveNet, WaveNet vocoder [29], which is conditioned on acoustically derived features, was proposed. Taking WaveNet as a vocoder significantly improves the synthetic speech quality and greatly reduces the required training data, making the WaveNet vocoder feasible for practical TTS systems.

However, WaveNet has a low speech generation speed owing to its autoregressive generation mechanism. WaveRNN [30] adopts a lightweight recurrent neural network (RNN) structure with acoustic feature conditions and hardware-friendly designs to achieve real-time generation. These autoregressive models often use teacher forcing, a technique that provides the correct values instead of the output from previous steps. Although it is very effective in stabilizing training, the mismatches between the training and inference stages cause the exposure bias [31] problem, resulting in quality degradation.

Non-autoregressive models using inverse autoregressive flows (IAF) have been investigated as an alternative real-time waveform generation approach. These IAF-based models [32], [33], [34] achieve higher inference speed through parallel waveform generation. Distillation techniques are adopted to alleviate the low training efficiency of IAF models due to their autoregressive training manner. However, distillation requires complex two-stage training, and connected teacher and student networks necessitate a large-scale memory for training.

To address these issues, GAN-based vocoders have been widely explored to take advantage of the compact generator size because the discriminator greatly helps the compact generator achieve high-fidelity speech generation. Parallel WaveGAN (PWG) [35] and MelGAN [36] are the recent most popular GAN-based vocoders, and many subsequent GAN-based vocoders are based on them [11], [12], [13], [14], [25], [26], [27], [37], [38], [39], [40], [41]. Non-autoregressive models without GAN are also proposed. WaveGlow [7] and WaveFlow [8] are normalizing flow-based neural vocoders. These flow-based neural vocoders achieve high speech generation speed with convincing sound quality. Moreover, denoising diffusion probabilistic-based neural vocoders such as DiffWave [9] and WaveGrad [10] have been proposed that iteratively refine Gaussian noise into speech via a Markov chain. Although the above-mentioned vocoders achieve impressive high-fidelity speech generation, independently controlling speech characteristics similarly to the conventional source-filter vocoders is challenging for these vocoders because of their purely data-driven training.

To tackle speech controllability, the WaveNet-based QP-Net [42], [43] and PWG-based QP-PWG [25], [26] vocoders with pitch-dependent dilated convolution neural networks (PDCNNs) have been proposed. PDCNNs effectively improve F_0 controllability by dynamically changing the CNN dilation size in accordance with the input F_0. Although QP-PWG outperforms PWG in F_0 controllability, there is still room for improvement in F_0 controllability.

C. Neural Vocoders Based on Source-Filter Modeling

Many neural vocoders based on the source-filter architecture are proposed to combine the high modeling capacity of deep neural networks with the merits of conventional source-filter modeling. For example, NSF [20], [21], [44] models the source excitation generation part by adding multiple sinusoidal signals and the resonance filtering part through multistage dilated convolution neural networks. The neural homomorphic vocoder (NHV) [22] generates waveforms on the basis of partly trainable digital signal processing modules with adversarial training. As another example, LPCNet [16] adopts linear filtering based on linear predictive coding [45], [46], [47], [48], and the neural network predicts the residual signal in an autoregressive manner, whereas GlotGAN [17], [18] and GELP [19] generate it in a non-autoregressive manner.

Despite their practical approach to introducing the source-filter architecture in deep neural networks, using signal-processing-based modules under the ad hoc assumptions usually results in sound quality and F_0 controllability degradations. We hypothesize that the reason for the degradation is the massive burden on neural networks and the lack of ability to control F_0. For instance, in NSF, the insufficient capacity of the source excitation generation part, which outputs the source excitation signal by adding a fixed number of sinusoidal signals, forces the spectral filtering part based on multistage dilated convolution neural networks to compensate for the missing information of the source excitation signal. As another example, LPCNet, GlotGAN, and GELP, whose neural-network-based source excitation generation parts are trained in a data-driven manner, suffer from significant degradation of their performance when there are unseen acoustic features such as F_0 values that are outside of the F_0 range of training data.

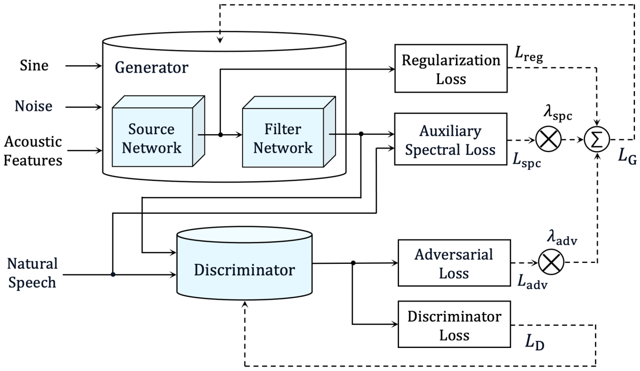

Fig. 2. Overall architecture of proposed uSFGAN.

III. Unified Source-Filter GAN

To develop a high-fidelity and F_0-controllable neural vocoder, we propose uSFGAN, which represents the source-filter architecture with a single neural network based on GAN. The generator network is factorized into a source excitation generation network (source network) and a resonance filtering network (filter network) using a regularization loss to make the source network output reasonable source excitation signals. To further improve the F_0 controllability, we introduce F_0-driven mechanisms designed on the basis of QP-PWG and NSF into the source network. Moreover, inspired by the recent successes of the neural vocoders that adopt harmonic-plus-noise (HN) speech modeling [13], [14], [21], [22], we introduce HN source excitation generation to obtain better sound quality. The overall architecture of uSFGAN is shown in Fig. 2, and the generator architectures are shown in Fig. 3.

A. Factorization of Generator Network

To make the proposed generator function like a source-filter model for achieving high acoustic controllability, two novel regularization losses are applied to the output of the source network to achieve reasonable source excitation signal generation.

1) Spectral Envelope Flattening Regularization Loss: The first regularization loss is designed on the assumption that the spectral envelopes of the source excitation signal are flat and their amplitude is constant. To match the constraints for the output signal of the source network, we take the L1 norm of the log amplitude spectral envelopes of the output source excitation signal calculated by using a simplified version [27] of CheapTrick [49]. The spectral envelope flattening regularization loss is formulated as

$$L_{\mathrm{reg}}(G) = \mathbb{E}_{\boldsymbol{z}}\left[\frac{1}{N}||\,\log \hat{E}_{\boldsymbol{z}}\,||_1\right], \tag{1}$$

where $||\cdot||_1$, $\hat{E}_{\boldsymbol{z}}$, N, and $\boldsymbol{z}$ denote the L1 norm, the magnitude of the source spectral envelopes of the output source excitation

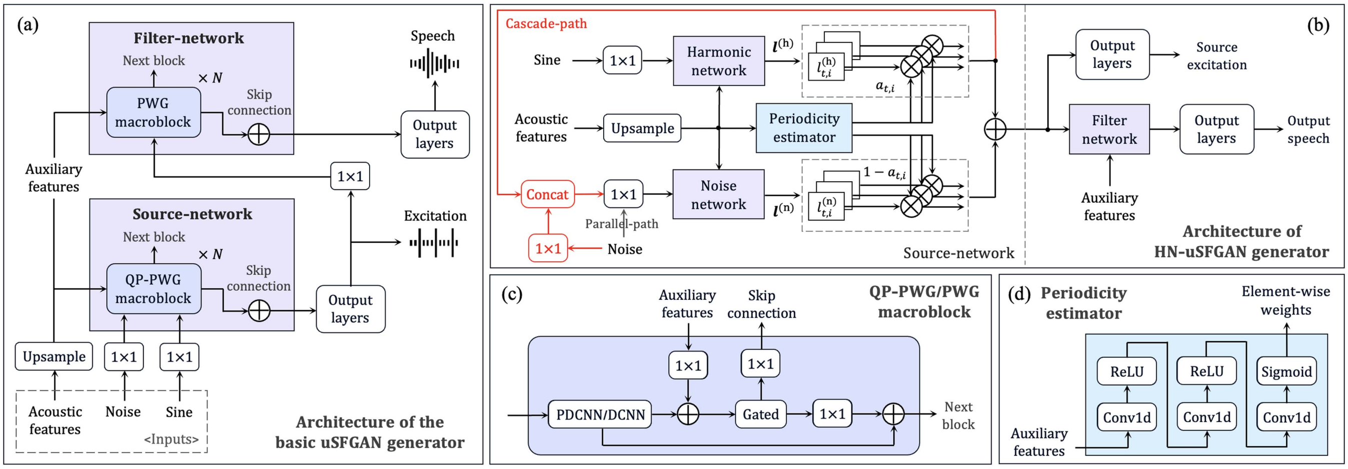

Fig. 3. Details of generator architectures. (a) Primary uSFGAN generator. (b) Harmonic-plus-noise uSFGAN generator. (c) QP-PWG macroblock. (d) Periodicity estimator. The red lines and blocks are used only for cascade harmonic-plus-noise source excitation generation. The output layers consist of two pairs of ReLU activation and one-by-one (1 × 1) convolution layers.

signal, the number of elements in the magnitude, and Gaussian noise signal, respectively. Note that when this loss reaches zero, the linear amplitude values $\hat{E}$ are one over all frequency and time frames. The initial uSFGAN paper [27] employed the L2 norm was employed for this loss function, but we use the L1 norm because of the improvement in several objective evaluation indices, such as F_0 reconstruction accuracy and voiced or unvoiced decision error rate.

2) Residual Spectra Targeting Regularization Loss: The second loss is designed to utilize the residual spectra calculated from the target speech and spectral envelopes extracted in the same way as above. However, to minimize the effects of the estimation error of F_0 and phases between generated and ground-truth speeches, we apply the mel-filter-bank to the amplitude spectrogram. The residual spectra regularization loss is formulated as

$$L_{\mathrm{reg}}(G) = \mathbb{E}_{\boldsymbol{x},\boldsymbol{z}}\left[\frac{1}{N}||\,\log\psi(S_{\boldsymbol{x}}) - \log\psi(\hat{S}_{\boldsymbol{z}})\,||_1\right], \qquad (2)$$

where $\boldsymbol{x}$, ψ, and N denote the ground-truth speech, function that transforms a spectral magnitude into the corresponding mel-spectrogram and the number of elements in the mel-spectrogram, respectively; $S_{\boldsymbol{x}}$ denotes the magnitude of residual spectra that have the same frame-wise average power as that of the ground-truth speech, and $\hat{S}_{\boldsymbol{z}}$ denotes the spectral magnitude of the output source excitation signal. Unlike the spectral envelope flattening regularization loss, this loss leaves the power estimation to the source network, similarly to an actual human speech production process where the power is controlled during the sound generation.

B. F_0-Driven Source Excitation Generation

Source excitation signals have high periodicity owing to their generation process that is based on vocal folds vibrations. Inspired by NSF [20], [21], [44], we input a sinusoidal-based signal to the generator generated by the same formula as that of NSF. The signal retains the input F_0 as the fundamental frequency but with an additional random noise signal. Moreover, we apply PDCNNs, which effectively enlarge the receptive fields in accordance with the input F_0 by dynamically changing the DCNN dilation factors. We found that using both the sinusoidal input and PDCNNs significantly improves F_0 controllability. However, the PDCNNs also tend to introduce undesired periodic components to the unvoiced segments. This tendency prevents the proper generation of other aperiodic source components, such as frication, aspiration, and transient sources, which adversely affect sound quality and naturalness.

To improve the source excitation signal modeling, especially for the unvoiced parts, we introduce a harmonic-plus-noise excitation generation mechanism inspired by the current successful works [13], [14], [21], [22] based on [50]. To explicitly model the periodic and aperiodic components, previous works [13], [14], [21], [22], [50] prepared two networks for generating each component and devised the architecture and input features for each. We adopt two harmonic-plus-noise modeling schemes, the cascade and parallel model structures, referring to PeriodNet [13]. Hono et al. represent the dependence of the periodic and aperiodic speech signals with the model structure. The cascade model structure combines the periodic and aperiodic speech generators in series so that the latter generator can predict the aperiodic component taking into account the dependence of the periodic component. On the other hand, the parallel model structure assumes their independence. To ascertain whether the cascade or parallel structure scheme is superior in modeling the source excitation signal, we propose the two approaches following PeriodNet. Moreover, the periodicity estimation is crucial for the naturalness of generated speech. Regarding NHV [22] and HN parallel waveGAN (HN-PWG) [14], we prepare a network to estimate periodicity-related weights from acoustic features and mix periodic and aperiodic source components on the basis of the weights.

The HN source excitation generation module consists of three networks: the harmonic network, noise network, and periodicity estimator, as shown in Fig. 3(b). Note that the red lines and red blocks are used only in the cascade approach. The harmonic network outputs latent features $l^{(\mathrm{h})}$ that correspond to the periodic components of the source excitation signal from a sinusoidal signal and auxiliary features. On the other hand, the noise network outputs latent features $l^{(\mathrm{n})}$ that correspond to the aperiodic components of the source excitation signal from a random noise signal and auxiliary features. In the cascade approach, the noise network also receives the output of the harmonic network. We use the QP-PWG macroblock shown in Fig. 3(c) in the harmonic network, while the PWG macroblock is used in the noise network. We adopt the harmonicity estimator of HN-PWG as the periodicity estimator shown in Fig. 3(d). Conditioned on the auxiliary features, the periodicity estimator outputs the channel-wise and sample-wise weights $\boldsymbol{a}$ within $[0, 1]$ corresponding to the speech periodicity. The two generated representations are summed element-wise using the estimated weights. The source excitation latent feature $\boldsymbol{l}$ is formulated as

$$l_{t,i} = a_{t,i} \cdot l^{(\mathrm{h})}_{t,i} + (1 - a_{t,i}) \cdot l^{(\mathrm{n})}_{t,i} \tag{3}$$

where the subscripts indicate the i^{th} channel of the t^{th} sample of each latent feature or weight. Since periodicity is estimated from auxiliary features, the input sinusoidal signal is generated using the continuous F_0 values obtained by interpolating the discontinuous F_0 values.

The cascade approach comprises three steps, as shown in Fig. 3(b). First, the harmonic network outputs the periodic source excitation representation, which is modulated using the channel-wise weights predicted by the periodicity estimator. Second, a random noise signal is mapped to a latent representation and mixed with a periodic source representation using a 1x1 convolution layer and the noise network. Finally, the output latent feature of the noise network is modulated using the weights and summed up with the modulated periodic source excitation representation to output the final source excitation representation. On the other hand, in the parallel approach, the aperiodic source representation is generated without the output periodic source representation of the harmonic network.

C. Adversarial Training

The training procedure of uSFGAN is common for GAN-based training plus auxiliary regularization losses. The discriminator D is trained to identify natural samples as *real* and generated samples as *fake* by minimizing the following optimization criterion:

$$L_D(G, D) = \mathbb{E}_{\boldsymbol{x}}\left[(1 - D(\boldsymbol{x}))^2\right] + \mathbb{E}_{\boldsymbol{z}}\left[D(G(\boldsymbol{z}))^2\right], \tag{4}$$

where $\boldsymbol{x}$ denotes the natural samples distributed from the data distribution of the natural samples, and $\boldsymbol{z}$ is random noise distributed from the Gaussian distribution $\mathcal{N}(\mathbf{0}, I)$. On the other hand, the generator G is trained to deceive the discriminator by minimizing the following adversarial loss:

$$L_{\mathrm{adv}}(G, D) = \mathbb{E}_{\boldsymbol{z}}\left[(1 - D(G(\boldsymbol{z})))^2\right]. \tag{5}$$

The final loss function of the generator can be written as the sum of the regularization loss L_{reg}, the auxiliary spectral loss L_{spc}, and the adversarial loss L_{adv}:

$$L_G(G, D) = L_{\mathrm{reg}}(G) + \lambda_{\mathrm{spc}} L_{\mathrm{spc}}(G) + \lambda_{\mathrm{adv}} L_{\mathrm{adv}}(G, D), \tag{6}$$

where λ_{spc} and λ_{adv} are loss balancing hyperparameters.

GAN-based adversarial training is effective for neural vocoders to implicitly learn perceptual aspects, such as phases, required for generating high-quality samples. GAN-based vocoders usually adopt auxiliary losses in the spectral domain and the feature matching loss to avoid mode collapse and improve the training stability. PWG [35] adopts the multi-resolution short-time Fourier transform (STFT) loss that can partly capture information about distance in phases in addition to the spectral structure between the natural speech $\boldsymbol{x}$ and the generated speech $\hat{\boldsymbol{x}} = G(\boldsymbol{z})$. It is formulated as the sum of the spectral convergence losses (L_{sc}) and the log STFT magnitude losses (L_{mag}) as follows:

$$L_{\mathrm{spc}}(G) = \frac{1}{M} \sum_{m=1}^{M} L_{\mathrm{s}}^{(m)}(G) \tag{7}$$

$$L_{\mathrm{s}}(G) = \mathbb{E}_{\boldsymbol{x},\boldsymbol{z}}\left[L_{\mathrm{sc}}(\boldsymbol{x}, G(\boldsymbol{z})) + L_{\mathrm{mag}}(\boldsymbol{x}, G(\boldsymbol{z}))\right] \tag{8}$$

$$L_{\mathrm{sc}}(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \frac{||\, |\mathrm{STFT}(\boldsymbol{x})| - |\mathrm{STFT}(\hat{\boldsymbol{x}})| \,||_{\mathrm{F}}}{||\, |\mathrm{STFT}(\boldsymbol{x})| \,||_{\mathrm{F}}} \tag{9}$$

$$L_{\mathrm{mag}}(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \frac{1}{N} ||\, \log|\mathrm{STFT}(\boldsymbol{x})| - \log|\mathrm{STFT}(\hat{\boldsymbol{x}})| \,||_1, \tag{10}$$

where M and $L_{\mathrm{s}}^{(m)}$ denotes the number of sets of analysis parameters for STFT and the spectral loss defined as (8) calculated with the m^{th} set. Moreover, $||\cdot||_{\mathrm{F}}$, $|\mathrm{STFT}(\cdot)|$, and N denote the Frobenius norm, the STFT magnitudes, and the number of elements in the magnitude, respectively.

On the other hand, HiFi-GAN [12] adopts the L1 loss in the mel-spectrogram domain because of the more human perception-related advantage. It is formulated as follows:

$$L_{\mathrm{spc}}(G) = \mathbb{E}_{\boldsymbol{x},\boldsymbol{s}}\left[\frac{1}{N} ||\, \phi(\boldsymbol{x}) - \phi(G(\boldsymbol{z})) \,||_1\right], \tag{11}$$

where ϕ and N denote the function of converting a speech signal to the corresponding mel-spectrogram and the number of elements in the mel-spectrogram, respectively.

We aim to develop a vocoder capable of synthesizing speech that faithfully reflects the input acoustic features. Since uSFGAN is conditioned on vocoder features, such as F_0, spectral envelopes, and aperiodicity, the estimation error is inevitable. Therefore, we argue that a looser constraint in the auxiliary spectral loss eases the mismatch between input and real features, especially F_0 and phases. Although the multi-resolution STFT loss and feature matching loss facilitate fine matches between the ground-truth and generated speeches, it is difficult for uSFGAN to satisfy them fully. In fact, for our best-proposed model, the mel-spectral L1 loss is used with the exact formulation as that of HiFi-GAN. The application of the mel-filter-bank eases the effect of F_0 and phase mismatch, making the optimization more straightforward and reasonable. Moreover, with adversarial training with sufficiently strong sophisticated discriminators, the generator can learn reasonable phases from the adversarial loss. Furthermore, although HiFi-GAN and MelGAN [36] adopt the feature matching loss to obtain the deep classification information provided by the discriminator, uSFGAN does not adopt the feature matching loss because of the mismatching problem of phase and F_0 between generated and ground-truth speeches.

IV. Experimental Evaluations

A. Data Preparation

We used the VCTK corpus [51], which contains 109 English speakers. We used only mic2 samples, and p315 was unavailable owing to a technical problem. The sampling rate was set to 24 kHz using the sox[1] downsampling function. No preprocessing, such as normalization or low-cut filtering, was applied to the audio. We divided the dataset following a specific rule to evaluate robustness against unseen F_0 values. The minimum and maximum F_0 values of the VCTK corpus were respectively found to be about 50 Hz and 400 Hz through careful investigation of each speaker. We limited the F_0 range of the training data from 70 Hz to 340 Hz and excluded two speakers (p271 and p300) from the training data to evaluate the robustness of unseen speakers. Thanks to this limitation, we can evaluate the methods using various conditions of seen or unseen speakers and F_0 ranges.

B. Model Details

1) Baseline Models: As the baselines, we used the following four models.

- *HiFi-GAN:* A high-fidelity GAN-based neural vocoder with four multi-period discriminators and four multi-scale discriminators. HiFi-GAN has no clue for controlling F_0, so we used it as the baseline for the evaluation of speech reconstruction. To train the HiFi-GAN model, we adopted the HiFi-GAN V1 [12] configuration and used an unofficial open-source implementation[2] for training the model.
- *WORLD:* A conventional source-filter model. This model achieves flexible controllability of acoustic features with reasonable sound quality. We used a Python wrapper[3] of the original WORLD implementation[4].
- *HN-NSF:* Harmonic-plus-noise neural source-filter with time-variant and trainable sinc filters that predict their cut-off frequency from the input acoustic features. We reimplemented the model on the basis of the official open-source code[5] without changing the model configuration except for increasing the training iterations.
- *QP-PWG:* A F_0-controllable neural vocoder based on GAN without the source-filter separation. It controls F_0 via the PDCNNs and input auxiliary F_0. We increased the number of residual blocks from the original configuration: PDCNNs 10 → 30 and DCNNs 10 → 30. The capacities of the QP-PWG model and the basic uSFGAN model detailed below are the same regarding the number of residual blocks.

We conditioned the *HiFi-GAN* model by using the mel-spectrogram as the original model with 80 mel-filter-banks, 1024 fast Fourier transform (FFT) points, 1024 points of the Hanning window, and the hop size was set to 120 (5 [ms]). We trained it for 2500 k iterations as the original model with the batch size set to 16, and the batch length set to 18000 (0.75 [s]), using the original setting of the Adam [52] optimizer. The loss weights followed the original setting. The weights of the adversarial loss, the feature matching loss, and the mel-spectral loss were set to 1.0, 2.0, and 45.0, respectively. *HN-NSF* was conditioned using discrete F_0, the mel-generalized cepstrum (MGC), and mel-cepstral aperiodicity (MAP). We trained it for 600 k steps with the batch size set to 1 as the original model, and the batch length was set to 24000 (1.0 [s]) using the original setting of the Adam optimizer. This model was trained using only the L2 loss on the log power spectrogram. *QP-PWG* was conditioned using almost the same features as those for *HN-NSF*, but continuous F_0 and a binary sequence representing voiced or unvoiced (V/UV) segments were used instead of the discrete F_0. We trained it for 600 k steps with the batch size set to 5 and the batch length set to 18000 (0.75 [s]) using the original setting of the RAdam [53] optimizer. The loss weights followed the original setting. The weights of the adversarial loss and the multi-resolution STFT loss were set to 4.0 and 1.0, respectively.

We extracted F_0 using the Harvest algorithm [54] with carefully set F_0 search range for each speaker. Then we extracted the log power spectral envelope using the CheapTrick algorithm [49] and coded it into the corresponding 41-dimensional MGC with the all-pass-constant set to 0.466. Also, we extracted aperiodicity using D4C algorithm [55] and coded them into the corresponding 21-dimensional MAP. These features were calculated with a shift period set to 5 ms. The mel-spectrogram was calculated using the librosa [56] function with the FFT size and window length set to 1024, and the hop length to 120 (5 [ms]) with a Hanning window.

2) Proposed Models: We used the following three uSFGAN-based models in the comparison experiments.

- *uSFGAN:* This model was based on our method proposed in [27]. The source network comprises 30 PDCNN blocks with six cycles, the filter network comprises 30 DCNN blocks with three cycles, and the PWG discriminator and PWG-based training procedure were used. The modifications are that the regularization loss became the L1 norm, and the input signal became a one-channel sinusoidal-based signal generated by the formula of NSF instead of a two-channel signal (a random noise signal and a sinusoidal-based signal without randomness). The updated loss leads to better performance of the objective metrics, and the input signal was for simplification of the comparison.
- *C-uSFGAN* (Cascade HN-uSFGAN): The first proposed model with the cascade harmonic-plus-noise excitation

[1] [Online]. Available: http://sox.sourceforge.net/

[2] An unofficial code of HiFi-GAN: https://github.com/kan-bayashi/ParallelWaveGAN

[3] A Python wrapper of WORLD vocoder: https://github.com/JeremyCCHsu/Python-Wrapper-for-World-Vocoder

[4] WORLD official implementation https://github.com/mmorise/World

[5] NSF official Pytorch implementation: https://github.com/nii-yamagishilab/project-NN-Pytorch-scripts

TABLE I
NUMBER OF MODEL PARAMETERS AND REAL-TIME FACTORS (RTF) CALCULATED ON A SINGLE GPU (TITAN RTX 3090) AND CPU WITH FOUR THREADS (AMD EPYC 7302)

Model	Parameters	RTF (GPU) ↓	RTF (CPU) ↓
HiFi-GAN	12.9 M	7×10^{-3}	0.83
HN-NSF	0.7 M	15×10^{-3}	1.49
QP-PWG	2.5 M	44×10^{-3}	4.26
uSFGAN	2.4 M	44×10^{-3}	4.38
C-uSFGAN	2.4 M	37×10^{-3}	3.99
P-uSFGAN	2.3 M	36×10^{-3}	4.19

generation, the residual spectra targeting loss, the mel-spectral loss, and the HiFi-GAN discriminator. The harmonic network had 20 PDCNN blocks with four cycles, the noise network was composed of five CNN blocks without cycles, and the filter network was the same as that of the basic *uSFGAN*.
- *P-uSFGAN* (Parallel HN-uSFGAN): The second proposed model. The network architecture was the same as that of *C-uSFGAN* except for the parallel or cascade architecture. This model is based on that in [28], but we made several improvements to it. Specifically, continuous F_0 was removed from the auxiliary features, and the two-dimensional BAP was changed to the corresponding 21-dimensional MAP. More details about the feature choices are described in Appendix A. Moreover, empirically better loss weighting hyperparameters were used in this article.

To enable the model to access the F_0 information only from the input sine waves, the auxiliary features included only MGC and MAP in all models. According to our preliminary experiments, this information restriction mechanism is essential for the proposed models to deal with excessively deviated F_0 such as the $2.0 \times F_0$ of female speakers with higher average F_0. The extractions of these acoustic features followed the same process as the baselines. The batch size and batch length of all the proposed models were set to 5 and 18000 (750 [ms]), respectively, as in the *QP-PWG*. The uSFGAN was trained with only the auxiliary losses for the first 100 k iterations and with the discriminator in the remaining 500 k steps using the RAdam optimizer with the same setting as that in *QP-PWG*. The loss weights were set on the basis of that of *QP-PWG:* $\lambda_{\text{adv}} = 4.0, \lambda_{\text{spc}} = 1.0, \lambda_{\text{reg}} = 1.0$. On the other hand, *C-uSFGAN* and *P-uSFGAN* followed the *HiFi-GAN* training procedure of simultaneously training the generator and the discriminators from scratch for 600 k iterations using the Adam optimizer with the same setting as that in *HiFi-GAN*. The loss weights were set based on those of *HiFi-GAN:* $\lambda_{\text{adv}} = 1.0, \lambda_{\text{spc}} = 45.0, \lambda_{\text{reg}} = 1.0$.

The model sizes of the baselines and proposed models are shown in Table I. Their inference speeds are also detailed with the real-time factor (RTF) in the same table. As shown in the table, the proposed models are much smaller than *HiFi-GAN* with the *V1* configuration, whereas *HiFi-GAN* achieves a much higher inference speed on a single GPU and CPU than the proposed models. HiFi-GAN adopts a configuration based on upsampling, where the preceding layers have lower temporal resolutions, resulting in higher computational efficiency and enabling fast waveform generation. On the other hand, the other models operate at a fixed temporal resolution consistent with the output waveform from the input. Since the computational complexity is proportional to the temporal resolution, these models tend to have slower speeds than the upsampling-based approach.

3) Ablation Models: To investigate the effectiveness of each component in our best-proposed *P-uSFGAN* described above, we prepared the following four ablation models for the comparison experiments. The input features and the training procedure of the ablation models followed those of *P-uSFGAN*.
- *Reg-Loss: P-uSFGAN* trained with the spectral envelope flattening loss instead of the residual spectra targeting loss.
- *HN-SN: P-uSFGAN* without the parallel harmonic-plus-noise source network but with the generator of the basic uSFGAN (30 layers of PDCNNs).
- *HiFi-D: P-uSFGAN* without the multi-period or multi-scale discriminator of HiFi-GAN but with the discriminator of PWG. We set $\lambda_{\text{adv}} = 8.0$ to match the reduced number of discriminators.
- *Mel-Loss: P-uSFGAN* trained with the multi-resolution STFT loss of PWG instead of the mel-spectral L1 loss. We set $\lambda_{\text{spc}} = 20.0$ so that the loss values before and after the change have roughly the same magnitude.

C. Evaluation of Speech Reconstruction

To evaluate the robustness of the proposed models for unseen acoustic features, both objective and subjective tests were conducted for the speech reconstruction performances. That is, three evaluation sets, including natural acoustic features within, beyond, and below the F_0 training range were adopted.

1) Objective Evaluation: As the objective evaluation measurements, the root mean square error of log F_0 [Hz] (RMSE), the voiced or unvoiced decision error [%] (V/UV), and mel-cepstral distortion [dB] (MCD) were used. The results are shown in Table II where results are divided on the basis of F_0 range. Each group included 200 utterances containing equal numbers of utterances by seen and unseen speakers. Since the primary purpose of our experiment was to investigate the F_0 robustness of the neural vocoders, and we confirmed that the proposed method did not cause significant degradation for unknown speakers [28], we only report the evaluation results for all speakers together.

Conventional parametric vocoders such as *WORLD* usually achieve higher objective acoustic controllability than neural vocoders [26], and the results of objective evaluation also demonstrate the same tendency. Specifically, baseline neural vocoders suffer from degradation when unseen F_0 was given, even though they partly outperform *WORLD* in the case of the seen F_0 range. In particular, *QP-PWG* shows large degradation in the V/UV error rate for the F_0 range below the training range. On the other hand, *uSFGAN*, whose difference from *QP-PWG* is the explicit decomposition of the source and filter network and the input sinusoidal-based signal, does not show significant degradation in any case. This implies the benefit provided by the source-filter modeling, that is, an inductive bias for the

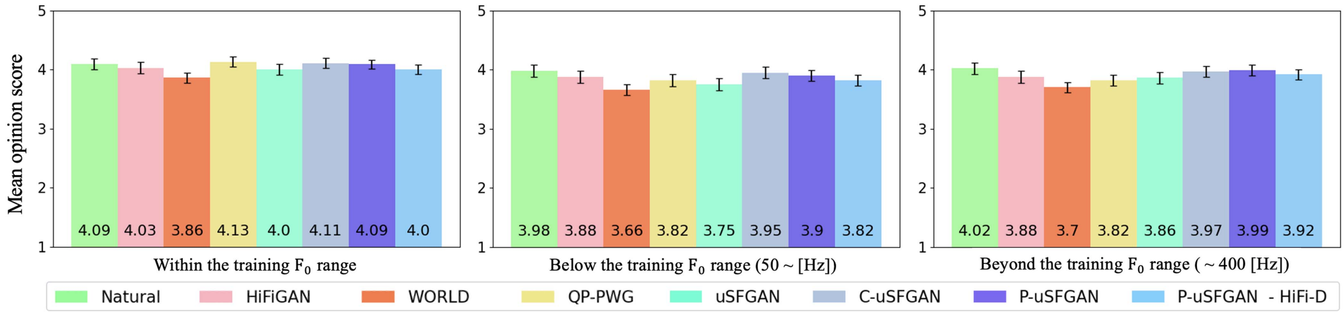

Fig. 4. Evaluation results of the MOS test. The average scores of all ranges are natural: 4.02, HiFi-GAN: 3.88, WORLD: 3.70, QP-PWG: 3.82, uSFGAN: 3.86, C-uSFGAN: 3.97, P-uSFGAN: 3.99, and P-uSFGAN - HiFi-D: 3.92.

TABLE II
RESULTS OF OBJECTIVE EVALUATIONS OF SPEECH RECONSTRUCTION. THE BEST SCORES ARE IN BOLD

method	RMSE ↓	V/UV ↓	MCD ↓
Within the training F_0 range (70 – 340 [Hz])			
WORLD	**0.05**	13	3.23
HiFi-GAN	0.07	10	3.37
HN-NSF	**0.05**	13	4.50
QP-PWG	0.07	11	**2.84**
uSFGAN	0.06	12	3.07
C-uSFGAN	**0.05**	**8**	2.86
P-uSFGAN	**0.05**	**8**	2.88
Below the training F_0 range (> 50 [Hz])			
WORLD	0.10	20	3.28
HiFi-GAN	0.11	24	3.11
HN-NSF	0.10	27	4.45
QP-PWG	0.12	25	2.92
uSFGAN	0.10	22	2.98
C-uSFGAN	**0.09**	**18**	**2.69**
P-uSFGAN	**0.09**	20	2.85
Beyond the training F_0 range (< 400 [Hz])			
WORLD	**0.06**	9	3.35
HiFi-GAN	0.08	8	3.60
HN-NSF	0.07	8	4.81
QP-PWG	0.09	9	3.07
uSFGAN	0.07	9	3.17
C-uSFGAN	**0.06**	**7**	2.91
P-uSFGAN	**0.06**	**7**	**2.87**

The bold values indicate the best scores.

speech production process leading to robustness to unseen F_0. Note that *C-uSFGAN* and *P-uSFGAN* show the best results in V/UV error rate, greatly outperforming WORLD, indicating the effectiveness of the harmonic-plus-noise architecture and of updating the loss functions. In conclusion, the proposed methods attain acoustic controllability similar to or better than those of conventional parametric vocoders.

2) Subjective Evaluation: For the subjective evaluation, we conducted an opinion test on sound quality using seven models and natural speech with ten subjects. Each subject evaluated 20 utterances per method. We recruited English-speaking evaluators through Amazon Mechanical Turk and instructed them to listen to the audio in a quiet room with headphones or earphones. Also, we filtered out scores from evaluators with unreasonable answers, such as where almost all scores were the same or the score of natural speech was lower than any system.

The results are shown in Fig. 4 where results are divided on the basis of F_0 range. *HN-NSF* was clearly inferior to the other models in sound quality, so we excluded it in the subjective evaluation experiment because of the possibility of undesired bias that the other samples would be highly evaluated. *HN-NSF* is a very basic baseline, and we speculate that the degradation was due to the simplicity of the model architecture and its low capacity to adapt to the large number of speakers in the VCTK corpus. However, we did not conduct any hyperparameter tuning on *HN-NSF* and note that there is a possibility that its performance can be improved by increasing the number of layers or introducing adversarial training.

We can see that all models except for *WORLD* achieve comparable scores for natural speech. Interestingly, *QP-PWG*, which uses the discriminator of PWG, achieves the best score, outperforming HiFi-GAN. The reason for the improvement of *QP-PWG* from the original model would be the increase in the number of the generator layers (20 → 60 residual blocks). However, for the unseen F_0 ranges, the proposed *C-uSFGAN* and *P-uSFGAN* achieve the best results, whereas *QP-PWG* is considerably degraded. Moreover, the differences between *HiFi-GAN* and natural speech become more prominent than in the case within the training F_0 range. On the other hand, there are no significant differences between *C-uSFGAN* and *P-uSFGAN* and natural speech in all cases. These results indicate that *HiFi-GAN* is data-driven and *QP-PWG* is highly data-driven. However, our proposed *C-uSFGAN* and *P-uSFGAN* complement the shortcomings of a data-driven approach.

D. Evaluation of F_0 Transformation

Next, we evaluated the performances of F_0 transformation with factors within $[2^{-1.0}, 2^{1.0}]$. The magnifications were taken equally on the logarithmic axis with the base at 2. The ground-truth F_0 was determined by multiplying the F_0 extracted from natural speech with the scale factors, and they were also adopted as the input F_0 of the models.

1) Objective Evaluation Settings: We extracted F_0 using the WORLD analyzer by the following procedure. When F_0 was multiplied by a scale factor greater than one, only the upper

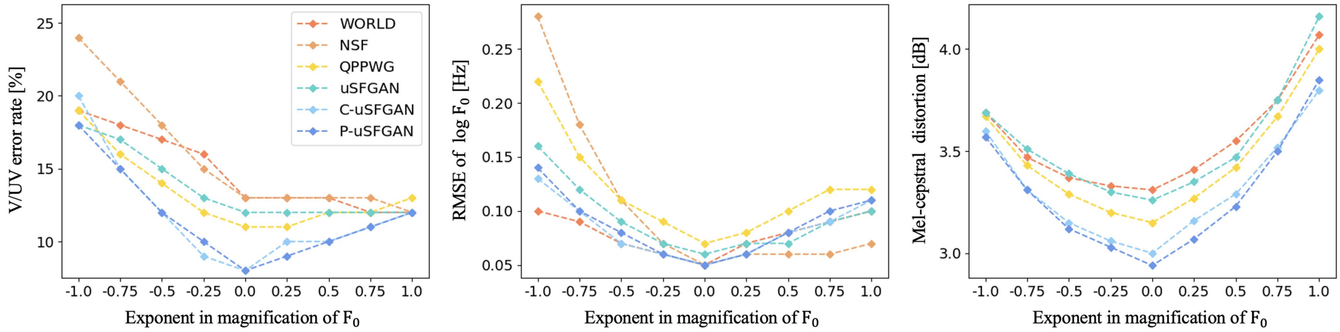

Fig. 5. Objective evaluation results of F_0 transformation for the comparison with baseline models. The MCD values of *HN-NSF* are excluded because it deviates from the range of the y-axis where the results of the other models are gathered.

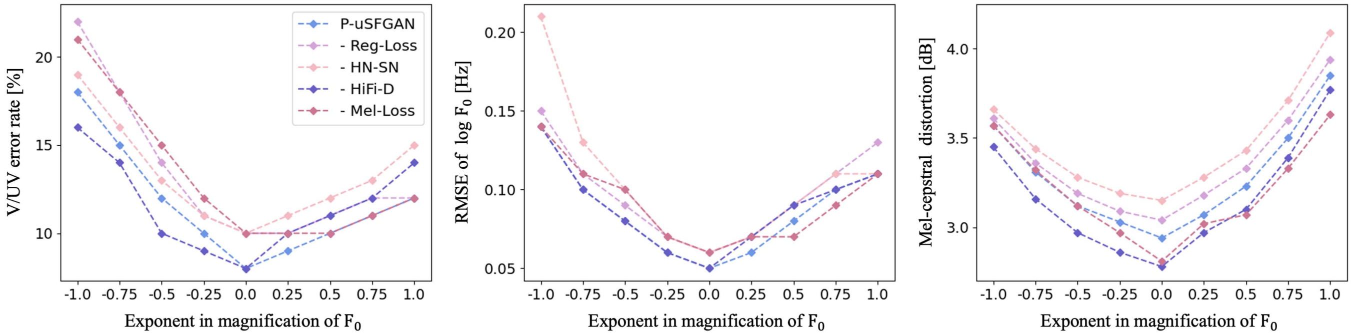

Fig. 6. Objective evaluation results of F_0 transformation for the ablation study.

bound of the F_0 search range was multiplied and transformed; otherwise, only the lower bound of the range was multiplied and transformed. MCD was calculated using the CheapTrick [49] algorithm provided by the WORLD analyzer, and the extracted F_0 was used for the calculation. However, we downsampled audio signals to 16000 [Hz] before estimating spectral envelopes because the CheapTrick algorithm sometimes fails in the estimation when the F_0 adaptive window size is larger than the FFT size. We made the available fixed FFT size sufficiently large by reducing the size of the F_0 adaptive window through downsampling and calculated MCD more accurately. The evaluations were conducted using the evaluation data whose F_0 range was within the training F_0 range (i.e., $70 - 340$ [Hz]).

2) Objective Evaluation: The objective evaluation results of comparison with baseline models are shown in Fig. 5. The result of log F_0 RMSE shows that although other models suffer from degradation in extreme cases ($\mathrm{F}_0 \times \{2^{-1.0}, 2^{1.0}\}$), the proposed *C-uSFGAN* and *P-uSFGAN* models achieve stable values close to that of *WORLD*. However, the two models achieve much lower V/UV error rates than all baseline models, which we found to have more impact on sound quality in our preliminary experiments. Moreover, we can see that all proposed models achieve better MCDs than *WORLD*. Again, the V/UV error rate and the RMSE of log F_0 in *QP-PWG* degrade as the scale factor increases or decreases, respectively. In contrast, *uSFGAN* does not significantly degrade for any factor, indicating the benefit of the source-filter decomposition.

3) Ablation Study: The objective evaluation results of the ablation study are shown in Fig. 6. From the results for *P-uSFGAN* and *P-uSFGAN - HN-SN*, we can see that the harmonic-plus-noise source network is very effective in improving the V/UV error rate and RMSE of log F_0. Moreover, the residual spectra targeting loss (*P-uSFGAN* vs *P-uSFGAN - Reg-Loss*) and mel-spectral loss (*P-uSFGAN* vs *P-uSFGAN - Mel-Loss*) effectively improve the V/UV error rate. *P-uSFGAN - HiFi-D* shows relatively good results in these objective metrics, but it is inferior to *P-uSFGAN* in sound quality, at least in speech reconstruction.

4) Subjective Evaluation: For the subjective evaluation, we conducted preference tests on sound quality using *WORLD*, *C-uSFGAN*, and *P-uSFGAN* for four F_0 scaling factors $\{2^{-1.0}, 2^{-0.5}, 2^{0.5}, 2^{1.0}\}$. Twenty subjects participated, and each subject evaluated ten pairs per F_0 scaling factor per method pair. The results are shown in Fig. 7. From the figures, both *C-uSFGAN* and *P-uSFGAN* outperform *WORLD* for all given F_0 scale factors, and *P-uSFGAN* is superior to *C-uSFGAN* in 3/4 of the items.

E. Visualization of Output Source Excitation Signals

To investigate the behavior of cascade and parallel HN-uSFGAN models (*C-uSFGAN* and *P-uSFGAN*), we visualized their output periodic and aperiodic source excitation signals in Fig. 8 with the spectrograms. These signals were obtained from the output latent representations of $\boldsymbol{l}$, $\boldsymbol{l}^{(\mathrm{h})}$, and $\boldsymbol{l}^{(\mathrm{n})}$ using the

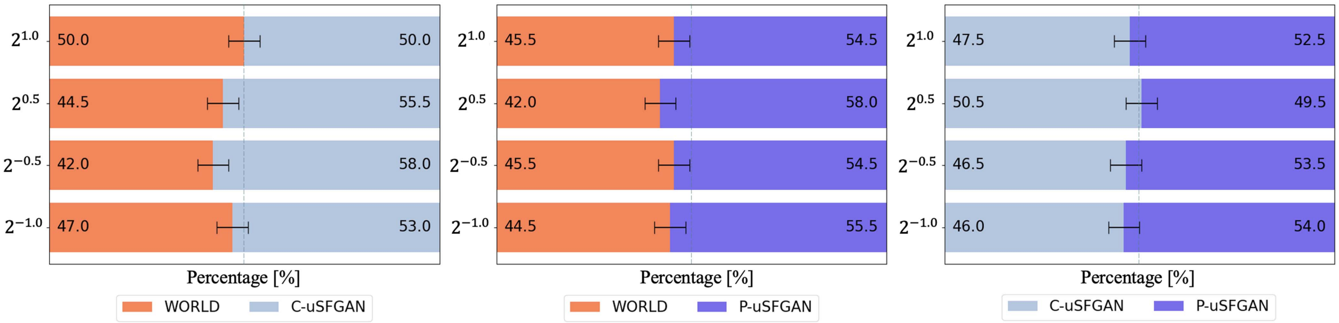

Fig. 7. Evaluation results of the preference test for F_0 transformation with the baseline *WORLD* and proposed *C-uSFGAN* and *P-uSFGAN*.

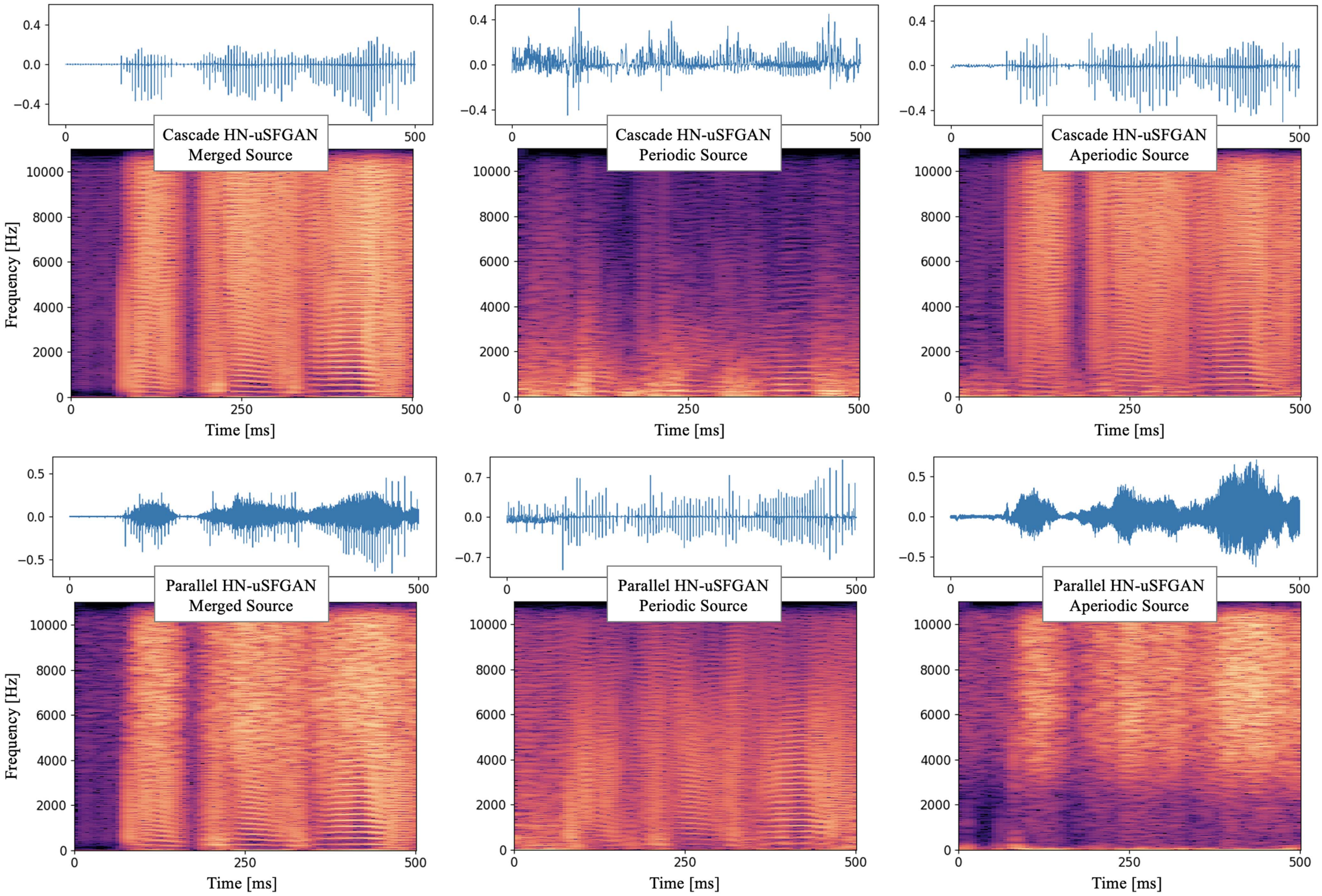

Fig. 8. Plots of output source excitation signals and spectrograms of *C-uSFGAN* (upper row) and *P-uSFGAN* (lower row) for 500 [ms]. The left column indicates the final source excitation signal, the middle column indicates the periodic source excitation signal, and the right column indicates the aperiodic source excitation signal.

output layers of the filter network and normalization of the signal power.

In Fig. 8, the output source excitation signals of *C-uSFGAN* seem to include fewer aperiodic components than in *P-uSFGAN*. Moreover, whereas *P-uSFGAN* well models the periodic and aperiodic components by the corresponding networks, *C-uSFGAN* does not seem to be able to disentangle these components. This indicates that the input aperiodic components are ignored as they pass through some networks. *C-uSFGAN* and *P-uSFGAN* achieve almost the same performance in speech reconstruction evaluation, as shown in Section IV-C. However, *P-uSFGAN* significantly outperforms *C-uSFGAN* in the evaluation of F_0 transformation, as shown in Section IV-D4. From the results, we can conclude that the disentanglement of periodic and aperiodic components has a good effect on the sound quality in F_0 transformation scenarios. Thus, we choose *P-uSFGAN* as our best-proposed model in this work.

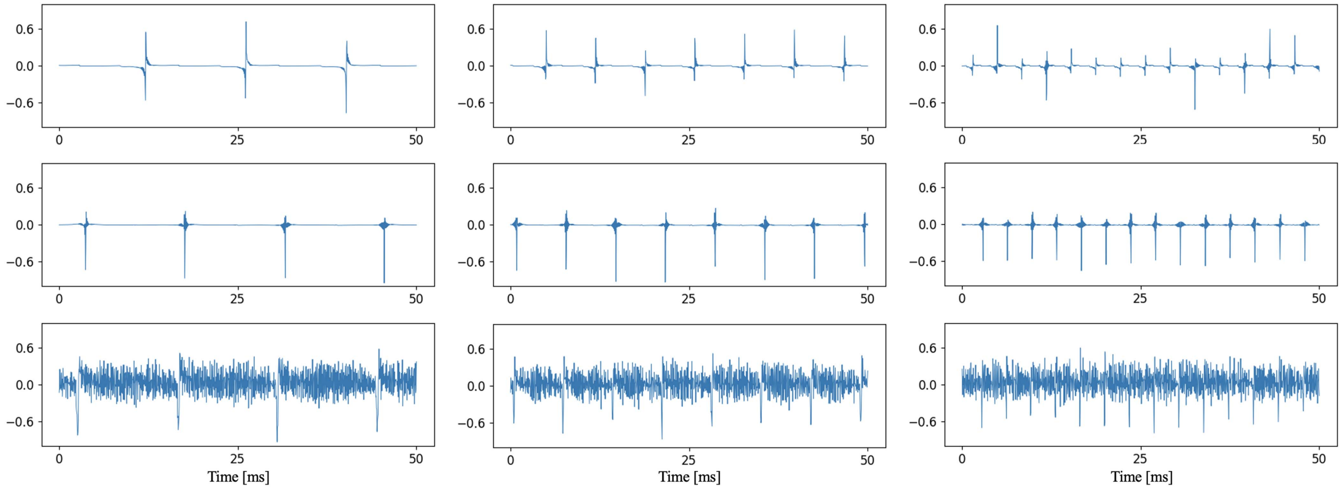

Fig. 9. Plots of output source excitation signals and spectrograms of *uSFGAN*, *C-uSFGAN*, and *P-uSFGAN* (from top to bottom row) with three F_0 scaling factors: 0.5, 1.0, and 2.0 (left to right column), for 50 [ms]. All of them were clipped from the same segment of the same utterance. The original F_0 values in this segment were around 140 [Hz].

Furthermore, source excitation signals of *uSFGAN*, *C-uSFGAN*, and *P-uSFGAN* for several F_0 scaling factors are plotted in Fig. 9. The figure shows that all proposed models can generate reasonable source excitation signals in accordance with the input F_0.

V. Conclusion

In this article, we proposed a novel source-filter modeling strategy that decomposes a single neural network using the regularization loss on the intermediate output. Thanks to the unified optimization of the source excitation and resonance filtering networks, our best-proposed method has been demonstrated to achieve equal or higher sound quality than the high-fidelity neural vocoder while attaining a similar or advanced F_0 controllability compared with a conventional parametric vocoder in the analysis-synthesis scenario. More experiments on the practical applications of the proposed neural vocoder and controllability over other acoustic features, such as spectral envelopes and aperiodicity, are left to future research.

Appendix A
Investigation of Input Acoustic Features

To further investigate the impact of different conditional acoustic features, we evaluated several models with different types of conditioning features with the same model architecture. In our proposed methods used in the experimental evaluations (Section IV), we chose the set of {MGC, MAP} as the best combination for the auxiliary features whose total number of dimensions is 62. Here, we compare *P-uSFGAN* with the following three models with different auxiliary features.

- *MEL:* This model adopts a full-band 80-dimensional log mel-spectrogram calculated in the setting described in Section IV-B instead of the vocoder features.
- *AuxF*$_0$*:* This model includes one-dimensional continuous F_0 in the default set of the auxiliary feature. The total number of dimensions of the auxiliary feature is 63.
- *BAP:* This model adopts the three-dimensional band-aperiodicity extracted using WORLD instead of 21-dimensional MAP. Coding is performed by one-dimensional interpolation on the frequency axis, which compresses the half-FFT size to three. The total number of dimensions of the auxiliary feature is 44.

All the ablation models were trained in the same setting as that in the *P-uSFGAN* model except for their auxiliary features. Note that all subnetworks (i.e., harmonic network, noise network, filter network, and periodicity estimator) are conditioned using the same auxiliary features.

The objective evaluation results are shown in Fig. 10. The *WORLD* results are provided as references. We found that differences between the models become apparent when F_0 is significantly high, so the study was conducted with F_0 increased by a factor of five. First, we can see that the *MEL* model degrades even with a small F_0 change. Since the mel-spectrogram already contains the F_0 information, we speculate that it is difficult for the model to manipulate F_0 by merely changing the sinusoidal inputs. The *AuxF*$_0$ model shows significant degradation with F_0 increased by a factor of two or more in its V/UV error rate, which is more critical for sound quality than the RMSE of log F_0. We confirmed that the generated speech is hardly voiced, resulting in significant degradation. We assume that this tendency is due to the fact that the inductive bias for speech production provided by the source-filter modeling is not obtained owing to the leakage of F_0 information to the filter network. The total degradation in the *MEL* model can be considered to have the same cause. From these experiences, we concluded that disentanglement of the input acoustic features and the restriction of F_0 information leakage to the filter network is essential to gaining the benefit from source-filter modeling. The *BAP* model, which gives periodicity information with fewer dimensions, shows minimal degradation

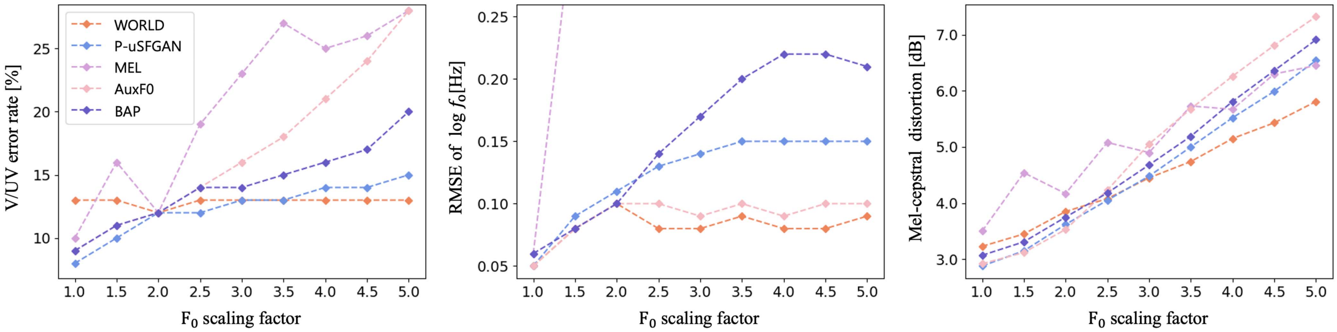

Fig. 10. Objective evaluation results of F_0 transformation for the ablation study on auxiliary features.

in both V/UV error rate and RMSE of log F_0. We assume that the degradation is because the neural network can ignore a fewer dimensional input feature (i.e., BAP) when the network can reconstruct the target waveform from the other input features in training. Moreover, this result suggests the importance of information about periodicity information in neural vocoders based on periodic and aperiodic component decomposition.

References

[1] 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. 3/4, pp. 187–207, 1999.

[2] 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.

[3] H. W. Dudley, "Remaking speech," *J. Acoustical Soc. Amer.*, vol. 11, no. 2, pp. 169–177, 1939.

[4] M. R. Schroeder, "Vocoders: Analysis and synthesis of speech," *Proc. IEEE*, vol. 54, no. 5, pp. 720–734, May 1966.

[5] 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.

[6] A. van den Oord et al., "WaveNet: A generative model for raw audio," in *Proc. 9th ISCA Speech Synth. Workshop*, 2016, Art. no. 125.

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

[8] 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.

[9] 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. [Online]. Available: https://openreview.net/forum?id=a-xFK8Ymz5J

[10] 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. [Online]. Available: https://openreview.net/forum?id=NsMLjcFaO8O

[11] A. Mustafa, N. Pia, and G. Fuchs, "StyleMelGAN: An efficient high-fidelity adversarial vocoder with temporal adaptive normalization," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 6034–6038.

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

[13] 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. Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 6049–6053.

[14] M.-J. Hwang, R. Yamamoto, E. Song, and J.-M. Kim, "High-fidelity parallel WaveGAN with multi-band harmonic-plus-noise model," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2021, pp. 2227–2231.

[15] 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. [Online]. Available: https://openreview.net/forum?id=v3aeIsY_vVX

[16] J.-M. Valin and J. Skoglund, "LPCNet: Improving neural speech synthesis through linear prediction," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 5891–5895.

[17] B. Bollepalli, L. Juvela, and P. Alku, "Generative adversarial network-based glottal waveform model for statistical parametric speech synthesis," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2017, pp. 3394–3398.

[18] 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.*, 2019, pp. 6915–6919.

[19] L. Juvela, B. Bollepalli, J. Yamagishi, and P. Alku, "GELP: GAN-excited linear prediction for speech synthesis from mel-spectrogram," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 694–698, doi: 10.21437/Interspeech.2019-2008.

[20] X. Wang, S. Takaki, and J. Yamagishi, "Neural source-filter-based waveform model for statistical parametric speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 5916–5920.

[21] X. Wang and J. Yamagishi, "Neural harmonic-plus-noise waveform model with trainable maximum voice frequency for text-to-speech synthesis," in *Proc. Speech Synth. Workshop*, 2019, pp. 1–6.

[22] Z. Liu, K. Chen, and K. Yu, "Neural homomorphic vocoder," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 240–244.

[23] M. Morrison, Z. Jin, N. J. Bryan, J.-P. Caceres, and B. Pardo, "Neural pitch-shifting and time-stretching with controllable LPCNet," 2021, *arXiv:2110.02360*.

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

[25] Y.-C. Wu, T. Hayashi, T. Okamoto, H. Kawai, and T. Toda, "Quasi-periodic parallel WaveGAN vocoder: A non-autoregressive pitch dependent dilated convolution model for parametric speech generation," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 3535–3539.

[26] 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.

[27] 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. Annu. Conf. Int. Speech Commun. Assoc.*, 2021, pp. 2187–2191.

[28] R. Yoneyama, Y.-C. Wu, and T. Toda, "Unified source-filter GAN with harmonic-plus-noise source excitation generation," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2022, pp. 848–852.

[29] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker-dependent WaveNet vocoder," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2017, pp. 1118–1122.

[30] N. Kalchbrenner et al., "Efficient neural audio synthesis," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2415–2424.

[31] K. Arora, L. El Asri, H. Bahuleyan, and J. Cheung, "Why exposure bias matters: An imitation learning perspective of error accumulation in language generation," in *Proc. Conf. Assoc. Comput. Linguistics*, 2022, pp. 700–710.

[32] A. van den Oord et al., "Parallel WaveNet: Fast high-fidelity speech synthesis," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 3918–3926.

[33] 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. [Online]. Available: https://openreview.net/forum?id=HklY120cYm

[34] R. Yamamoto, E. Song, and J.-M. Kim, “Probability density distillation with generative adversarial networks for high-quality parallel waveform generation,” in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 699–703.

[35] 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.

[36] K. Kumar et al., “Melgan: Generative adversarial networks for conditional waveform synthesis,” in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 14910–1421.

[37] R. Yamamoto, E. Song, M.-J. Hwang, and J.-M. Kim, “Parallel waveform synthesis based on generative adversarial networks with voicing-aware conditional discriminators,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 6039–6043.

[38] 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. Spoken Lang. Technol. Workshop*, 2021, pp. 492–498.

[39] J. Yang, J. Lee, Y. Kim, H. Cho, and I. Kim, “VocGAN: A high-fidelity real-time vocoder with a hierarchically-nested adversarial network,” in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 200–204.

[40] M. Bińkowski et al., “High fidelity speech synthesis with adversarial networks,” in *Proc. Int. Conf. Learn. Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=r1gfQgSFDr

[41] J. You et al., “GAN vocoder: Multi-resolution discriminator is all you need,” in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2021, pp. 2177–2181.

[42] Y.-C. Wu, T. Hayashi, P. L. Tobing, K. Kobayashi, and T. Toda, “Quasi-periodic WaveNet vocoder: A pitch dependent dilated convolution model for parametric speech generation,” in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 196–200.

[43] 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.

[44] 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.

[45] F. Itakura and S. Saito, “Analysis synthesis telephony based on the maximum likelihood method,” in *Proc. 6th Int. Congr. Acoust.*, 1968, pp. C17–C20. [Online]. Available: https://cir.nii.ac.jp/crid/1573950400351247616

[46] B. S. Atal and M. R. Schroeder, “Predictive coding of speech signals,” in *Proc. Speech Commun. Process.*, 1967, pp. 360–361.

[47] D. Wong, B.-H. Juang, and A. Gray, “An 800 bit/s vector quantization LPC vocoder,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 30, no. 5, pp. 770–780, Oct. 1982.

[48] A. V. McCree and T. P. Barnwell, “A mixed excitation LPC vocoder model for low bit rate speech coding,” *IEEE Speech Audio Process.*, vol. 3, no. 4, pp. 242–250, Jul. 1995.

[49] M. Morise, “CheapTrick, a spectral envelope estimator for high-quality speech synthesis,” *Speech Commun.*, vol. 67, pp. 1–7, 2015.

[50] Y. Stylianou, “Applying the harmonic plus noise model in concatenative speech synthesis,” *IEEE Speech Audio Process.*, vol. 9, no. 1, pp. 21–29, Jan. 2001.

[51] J. Yamagishi, C. Veaux, and K. MacDonald, “CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92),” [sound]. Univ. Edinburgh. The Centre for Speech Technol. Res. (CSTR), 2019, doi: 10.7488/ds/2645.

[52] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Proc. Int. Conf. Learn. Representations*, 2015. [Online]. Available: https://arxiv.org/abs/1412.6980

[53] L. Liu et al., “On the variance of the adaptive learning rate and beyond,” in *Proc. Int. Conf. Learn. Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=rkgz2aEKDr

[54] M. Morise, “Harvest: A high-performance fundamental frequency estimator from speech signals,” in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2017, pp. 2321–2325.

[55] M. Morise, “D4C, a band-aperiodicity estimator for high-quality speech synthesis,” *Speech Commun.*, vol. 84, pp. 57–65, 2016.

[56] B. McFee et al., “librosa/librosa: 0.9.2,” 2022, doi: 10.5281/zenodo.6759664.