VAE-SiFiGAN: Source-Filter HiFi-GAN Based on Variational Autoencoder Representations with Enhanced Pitch Controllability

Kenichi Ogita, Reo Yoneyama, Wen-Chin Huang, Tomoki Toda
Nagoya University, Nagoya, Japan
{ogita.kenichi, yoneyama.reo, wen.chinhuang}@g.sp.m.is.nagoya-u.ac.jp
tomoki@icts.nagoya-u.ac.jp

***Abstract*—Source-filter HiFi-GAN (SiFi-GAN) is a neural vocoder offering fast, high-quality voice synthesis with fundamental frequency (F_0) controllability. However, SiFi-GAN takes hand-crafted acoustic features from traditional signal processing as input, causing some limitations, such as sound quality degradation in F_0 extrapolation. This paper proposes VAE-SiFiGAN, which learns latent representations from Mel-spectrograms via a variational autoencoder (VAE). The latent representations learned through the probabilistic framework enable SiFi-GAN to better model the stochastic components in speech signals, achieving sound quality improvements in F_0 modification. Furthermore, to address the insufficient F_0 controllability caused by the entanglement of Mel-spectrograms and F_0 information, we propose to guide the latent representation learning process with hand-crafted features less affected by F_0 and used only during training. Experimental results show that VAE-SiFiGAN achieves superior F_0 controllability compared to SiFi-GAN.**

***Index Terms*—neural vocoder, variational autoencoder, source-filter model, pitch control**

I. Introduction

A neural vocoder [1]–[8] is a waveform generator based on deep neural networks (DNNs), achieving remarkably higher sound quality than conventional source-filter vocoders [9], [10]. HiFi-GAN [8] is one of the most popular neural vocoders due to its ability to balance sound quality with synthesis efficiency. On the other hand, for practical use, vocoders often need to also offer flexible control over the fundamental frequency (F_0), which is crucial for generating the desired intonation and pitch. The fully data-driven manner of HiFi-GAN tends to limit the controllability of F_0. To address this issue, Source-filter HiFi-GAN (SiFi-GAN) [11] incorporates an F_0-driven mechanism and source-filter theory into HiFi-GAN, aiming to simultaneously achieve high speech quality, fast synthesis, and F_0 controllability.

Nonetheless, SiFi-GAN still suffers from several issues in the context of practical applications. One key limitation is its inability to reproduce the stochastic aspects of speech, such as acoustic fluctuations and variances caused by the physical speech production process, including the natural variability of vocal-fold vibration and articulation where no two utterances are exactly alike. This variability is essential for synthesizing natural-sounding voices [12], [13]. Therefore, it is desirable to develop vocoders incorporating a stochastic mechanism to realize a one-to-many mapping from acoustic features to waveforms and model these fluctuations. Many generative adversarial networks (GAN)-based neural vocoders [6]–[8], [14]–[16], including SiFi-GAN, utilize the GAN's probabilistic generative framework, yet they practically learn an almost deterministic mapping from input features to waveforms, thus depending entirely on the acoustic features to capture these variations. Additionally, many F_0-controllable neural vocoders [11], [15]–[18], including SiFi-GAN, employ WORLD [10] features that are extracted deterministically using signal-processing algorithms. Consequently, the combination of deterministic features and near one-to-one waveform generation prevents the model from capturing the natural variability in the input speech, limiting the ability to synthesize expressive waveforms. Moreover, this feature extraction algorithm involves processing steps for which differentiable implementations are not readily available, making it difficult to incorporate the algorithm into end-to-end systems directly. Furthermore, acoustic feature extraction algorithms based on signal processing are generally noise-sensitive, which can lead to degraded synthesis quality in real-world applications.

In this work, we propose VAE-SiFiGAN, which adopts a variational autoencoder (VAE) [19] framework to learn probabilistic latent representations. In contrast to the original SiFi-GAN which takes WORLD features as input, VAE-SiFiGAN extracts stochastic latent representations from the input Mel-spectrogram, which offers flexibility to integrate with end-to-end systems, and is thus capable of modeling the uncertainty in speech, including fluctuations, while also improving robustness against background noise. Furthermore, to encourage F_0-independence in the learned representations, we propose an F_0-removal mechanism, where we align the posterior distribution of the VAE encoder with prior distributions defined by the WORLD features less affected by F_0. Experimental results show that our proposed VAE-SiFiGAN demonstrates superior F_0 control performance over SiFi-GAN.

II. Baseline Source-Filter HiFi-GAN

In this section, we describe the baseline model, SiFi-GAN [11]. In SiFi-GAN, the input features consist of Mel-generalized cepstral coefficients (MGC) and band-aperiodicity (BAP), extracted using WORLD analyzer [10] based on signal processing.

A. Source-filter networks

The SiFi-GAN generator is decomposed into the source-network and filter-network connected in series. The source-network is composed of upsampling and downsampling modules. Upsampling modules include transposed 1D convolutional neural networks (CNNs) and quasi-periodic residual blocks (QP-ResBlocks). Each QP-ResBlock comprises multiple iterations of Leaky ReLU [20], pitch-dependent dilated convolution neural networks (PDCNNs) [14], [21], and 1D CNN. Downsampling modules hierarchically receive an F_0-dependent sine wave, generated in the same manner as used in Neural Source-Filter (NSF) [22]. The input features are progressively upsampled through the transposed CNNs and QP-ResBlocks, while F_0-dependent sine waves are simultaneously fed to each upsampling layer via downsampling CNNs. In order to extract the source excitation signal, the output of the final QP-ResBlock is passed through Leaky ReLU and a 1D CNN. Benefiting from this F_0-driven architecture, the model can enhance F_0 controllability.

The filter-network is composed of transposed CNNs with multi-receptive field fusion (MRF) modules, closely resembling the HiFi-GAN [8] generator architecture. The key difference is that the final QP-ResBlock output from the source-network is fed to each block through downsampling CNNs. This cascade structure of the source network and the filter network is essential for effectively capturing high-frequency components of speech and improving F_0 controllability.

B. Training with source excitation regularization loss

The training criteria for the SiFi-GAN follow HiFi-GAN; however, in order to explicitly decompose the generator into the source-network and the filter-network, a regularization loss, as described in Equation (1) and [16], is applied to the output of the source-network:

$$L_{\text{reg}}(G) = \mathbb{E}_{\mathbf{x},\mathbf{c}}\left[\frac{1}{N}||\log\psi(\hat{S}) - \log\psi(S)||_1\right] \quad (1)$$

where $\mathbf{x}$ and $\mathbf{c}$ denote the ground truth speech and input features; ψ and N denote the function that converts an amplitude spectrogram to a Mel-spectrogram and the number of dimensions of the Mel-spectrogram; $\hat{S}$ and S denote the amplitude spectrum of the source excitation signal output by the source-network and the residual spectrogram, respectively. This residual spectrogram is obtained by extracting the spectral envelope using CheapTrick [23] and by normalizing the average power in each frame. The regularization loss ensures that the output source excitation signal has a flat spectral characteristic, like actual excitation signals that have not yet been colored by the vocal tract.

The final loss function for the generator is thus defined as a combination of an adversarial loss $\mathcal{L}_{G,\text{adv}}$, a Mel-spectral L1 loss $\mathcal{L}_{\text{mel}}$, a feature-matching loss $\mathcal{L}_{\text{fm}}$, and the regularization loss $\mathcal{L}_{\text{reg}}$, as shown in Equation (2):

$$\mathcal{L}_G = \mathcal{L}_{G,\text{adv}} + \lambda_{\text{mel}}\mathcal{L}_{\text{mel}} + \lambda_{\text{fm}}\mathcal{L}_{\text{fm}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}} \quad (2)$$

where λ_{mel}, λ_{fm}, and λ_{reg} are loss-balancing hyperparameters.

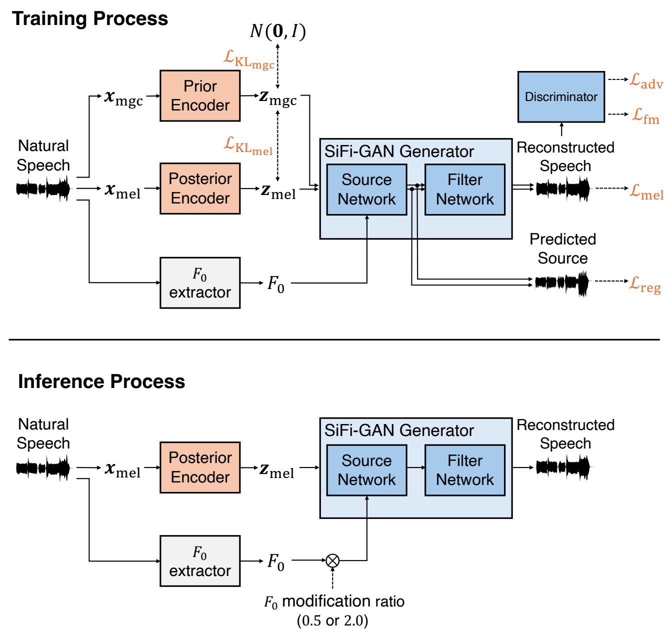


Fig. 1: Overview of VAE-SiFiGAN. mgc and mel denote Mel-generalized cepstral coefficients concatenated with band-aperiodicity and Mel-spectrogram, respectively.

III. Proposed Method: VAE-SiFiGAN

An overview of the proposed VAE-SiFiGAN is illustrated in Fig. 1. Instead of using the original features (*i.e.*, MGC and BAP) from SiFi-GAN [11], we introduce learnable latent representations extracted from Mel-spectrograms via the VAE encoder. In addition to the VAE encoder (*i.e.*, posterior encoder), we additionally incorporate a prior encoder as in VITS [24], a text-to-speech model based on variational inference, to control the posterior distribution.

A. Posterior encoder

We adopt the posterior encoder structure from VITS as the encoder that extracts latent representations from Mel-spectrograms. It consists of 1D CNNs and non-causal WaveNet residual blocks [5], [25], enabling it to capture the long-term dependencies of speech signals. Given the Mel-spectrogram $\mathbf{x}_{\text{mel}}$ as input, the posterior encoder yields a latent representation $\mathbf{z}_{\text{mel}}$. To enable flexible F_0 control as in conventional source–filter vocoders [9], [10], the F_0 series extracted by an F_0 estimator is externally provided to the SiFi-GAN generator in addition to $\mathbf{z}_{\text{mel}}$.

However, since Mel-spectrograms generally contain F_0 information, $\mathbf{z}_{\text{mel}}$ naturally retains some F_0 cues. If $\mathbf{z}_{\text{mel}}$ still carries F_0 cues that contradict the externally specified F_0 values, modifying the F_0 can lead to conflicts, ultimately degrading F_0 controllability [15], [17]. Therefore, an additional mechanism is needed to remove any residual F_0 content from $\mathbf{z}_{\text{mel}}$.

B. Prior encoder for F_0 removal

To address the potential conflict described in section III-A, we introduce the prior encoder that helps eliminate F_0 information from $\mathbf{z}_{\text{mel}}$. In line with SiFi-GAN, the prior encoder receives hand-crafted features with reduced F_0 influence, namely MGC and BAP, collectively denoted as $\mathbf{x}_{\text{mgc}}$. In contrast to Mel-spectrograms, MGC+BAP features are nearly independent of F_0 due to their extraction algorithms [23], [26], which are designed to remove F_0 information as part of the process. Therefore, the latent representation $\mathbf{z}_{\text{mgc}}$ produced by the prior encoder is expected to be almost free of F_0 information. The prior encoder shares the same architecture as the posterior encoder.

During training, the posterior encoder is regularized by the prior encoder through Kullback-Leibler (KL) divergence loss, which compels the Mel-based latent feature $\mathbf{z}_{\text{mel}}$ to discard F_0 information. The prior encoder, in turn, is regularized by a standard normal distribution $N(\mathbf{0}, I)$. Consequently, the training objectives for these encoders are formulated as Equations (3) and (4):

$$\mathcal{L}_{\text{kl}_{\text{mgc}}} = \text{KL}[q_\theta(\mathbf{z}_{\text{mgc}} \mid \mathbf{x}_{\text{mgc}}) \;||\; N(\mathbf{0}, I)] \tag{3}$$

$$\mathcal{L}_{\text{kl}_{\text{mel}}} = \text{KL}[q_\phi(\mathbf{z}_{\text{mel}} \mid \mathbf{x}_{\text{mel}}) \;||\; q_\theta(\mathbf{z}_{\text{mgc}} \mid \mathbf{x}_{\text{mgc}})] \tag{4}$$

where θ and ϕ denote the parameters of the prior and posterior encoders; $q_\theta(\mathbf{z}_{\text{mgc}} \mid \mathbf{x}_{\text{mgc}})$ and $q_\phi(\mathbf{z}_{\text{mel}} \mid \mathbf{x}_{\text{mel}})$ denote the posterior distributions of latent representations $\mathbf{z}_{\text{mgc}}$ and $\mathbf{z}_{\text{mel}}$, respectively.

The prior encoder is used only for guiding the posterior encoder's distribution to disentangle F_0 information. After training, inference relies solely on the posterior encoder (*i.e.*, the Mel-spectrogram $\mathbf{x}_{\text{mel}}$ and its corresponding latent representation $\mathbf{z}_{\text{mel}}$). This design allows inference to rely exclusively on Mel-spectrogram inputs, eliminating the need for hand-crafted acoustic features such as MGC and BAP, while maintaining robust F_0 controllability.

C. Training criteria

Finally, we extend the SiFi-GAN training objective by incorporating the KL divergence losses for both the posterior and prior encoders, as defined in Equation (5):

$$\begin{aligned}\mathcal{L}_G =& \lambda_{\text{kl}_{\text{mgc}}}\mathcal{L}_{\text{kl}_{\text{mgc}}} + \lambda_{\text{kl}_{\text{mel}}}\mathcal{L}_{\text{kl}_{\text{mel}}} + \lambda_{\text{mel}}\mathcal{L}_{\text{mel}} \\ &+ \lambda_{\text{fm}}\mathcal{L}_{\text{fm}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}} + \mathcal{L}_{G,\text{adv}}\end{aligned} \tag{5}$$

where $\lambda_{\text{kl}_{\text{mgc}}}$, $\lambda_{\text{kl}_{\text{mel}}}$, λ_{mel}, λ_{fm}, λ_{reg} are loss-balancing hyperparameters. In order to seamlessly leverage the complementary properties of both latent representations, our method employs a single generator that generates two separate speech reconstructions from $\mathbf{z}_{\text{mgc}}$ and $\mathbf{z}_{\text{mel}}$. Therefore, all loss terms except $\mathcal{L}_{\text{kl}_{\text{mgc}}}$ and $\mathcal{L}_{\text{kl}_{\text{mel}}}$ are calculated based on the average of these two outputs.

IV. Experimental evaluation

In this section, we demonstrate the performance of our proposed method. We generated singing voices in the scenarios of both copy-synthesis and F_0 transformation.

A. Data preparation

Following the previous work [11], we used Namine Ritsu's database [27], which contains a collection of Japanese vocal recordings from a single female singer. The dataset comprises 110 songs with a total duration of approximately 4.35 hours, and the annotated F_0 range spans from 100 to 1000 Hz. Each song was further segmented into shorter phrases based on rests indicated in the musical score.

For feature extraction, we used a fast Fourier transform (FFT) size of 1024 and a 5-ms frame shift for all computations. The spectral envelopes, extracted using the CheapTrick algorithm [23], were converted into 40-dimensional Mel-generalized cepstral coefficients (MGC), while 3-dimensional band-aperiodicity parameters (BAP) were obtained via the D4C algorithm [26]. A 1024-point FFT with a Hanning window was applied to extract 80-dimensional Mel-spectrograms (MEL), whose magnitudes were then converted to a logarithmic scale. All acoustic features were normalized to zero mean and unit variance before being fed into the model. For F_0 extraction, we applied the Harvest algorithm [28], followed by interpolation and smoothing to obtain a one-dimensional continuous F_0 (cF_0) [29]. The sine waves used in the SiFi-GAN [11] generator were then synthesized from cF_0 according to the generation method described in [22]. Note that no voiced/unvoiced flag is used either as an input feature or for sine wave generation.

B. Model details

We compared our proposed VAE-SiFiGAN with the following baseline and ablation models:

- **SiFi-GAN**: Baseline vanilla SiFi-GAN vocoder, conditioned on {MGC, BAP}. We set $\lambda_{\text{mel}} = 45.0$, $\lambda_{\text{fm}} = 2.0$, and $\lambda_{\text{reg}} = 1.0$.
- **VAE-SiFiGAN**: Proposed model with the posterior and prior encoders, conditioned on {MGC, BAP, MEL} during training but used only {MEL} for inference. We set $\lambda_{\text{kl}_{\text{mgc}}} = 1.0$, $\lambda_{\text{kl}_{\text{mel}}} = 1.0$, $\lambda_{\text{mel}} = 45.0$, $\lambda_{\text{fm}} = 2.0$, and $\lambda_{\text{reg}} = 1.0$.
- **w/o Prior**: Proposed model without the prior encoder. We set the prior distribution of the posterior encoder to a standard normal distribution $N(\mathbf{0}, I)$.

We adopted the original architecture and training configuration in [11] for SiFi-GAN. Both posterior and prior encoders have the same architecture, with a hidden layer dimension of 192, 16 WaveNet residual blocks [5], [25], and a kernel size of 7. Each encoder outputs 30-dimensional latent representations. All models used the same UnivNet multi-period and multi-resolution discriminators [30]. We trained all vocoders for 500k steps using the Adam [31] optimizer, with a mini-batch size of 16 and a mini-batch length of 8400.

C. Objective Evaluation

To evaluate the F_0 controllability of each model, we report the root mean squared error (RMSE [Hz]) of the log-F_0 and the voiced/unvoiced classification error (V/UV [%]). Each metric was evaluated using F_0 modification ratios, which are

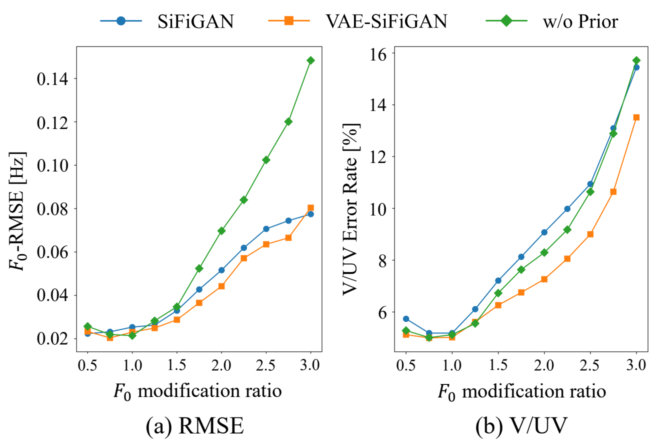


Fig. 2: Results of objective evaluation.

multiplied by the original F_0 values in Hz, from 0.5 to 3.0 in increments of 0.25.

The results of the objective evaluation are presented in Fig. 2. In the $F_0 \times 2.0$ or higher conditions, **w/o Prior** model exhibits a marked increase in RMSE. Additionally, as shown in Fig. 3, **w/o Prior** fails to accurately reconstruct the harmonic structure when F_0 is heavily extrapolated. These observations are presumably due to inconsistent F_0 cues remaining in the latent representations and the externally supplied F_0 series, thereby reducing its F_0 controllability.

Compared with **SiFi-GAN**, **VAE-SiFiGAN** achieves lower RMSE for most F_0 scaling conditions, except at $F_0 \times 0.5$ and 3.0, suggesting that disentangling F_0 information from the latent representations plays a pivotal role in effective F_0 control. Furthermore, **VAE-SiFiGAN** outperforms the other two models in terms of V/UV performance across the F_0 scaling conditions.

D. Subjective Evaluation

To evaluate the perceptual quality of the synthesized singing voices, we conducted a five-point Mean Opinion Score (MOS) test. In this test, we evaluated singing voices generated by copy synthesis and those generated under F_0 scaling conditions of 0.5 and 2.0. Twenty-two Japanese speakers participated in the test, and each of them assessed 12 samples per method under each F_0 scaling condition.

The results of the subjective evaluation are presented in Fig. 4. **VAE-SiFiGAN** and **w/o Prior** yield higher perceived quality than **SiFi-GAN** under $F_0 \times 0.5$ and 2.0. Notably, despite **SiFi-GAN** achieving a substantially lower RMSE than **w/o Prior** under $F_0 \times 2.0$, its MOS is lower than that of **w/o Prior**. We found that under the $F_0 \times 0.5$ and 2.0 conditions, **SiFi-GAN** occasionally produces buzzy voices, suggesting that relying solely on hand-crafted features extracted with signal processing algorithms fails to capture certain spectral characteristics. In contrast, **VAE-SiFiGAN** and **w/o Prior** randomly sample their latent representations from the estimated latent distribution on each occasion, and these latent representations are helpful for achieving more robust speech generation even under extrapolated F_0 conditions.

Despite **VAE-SiFiGAN** considerably outperforming **w/o Prior** on RMSE and V/UV, it only achieves a comparable MOS score at $F_0 \times 2.0$. One plausible explanation is that **w/o Prior** sacrifices some degree of F_0 control for a more natural-sounding output, reflecting a trade-off between F_0 controllability and acoustic fidelity.

In **VAE-SiFiGAN**, although MEL is used as input, the posterior encoder's alignment with MGC+BAP-based latent representations, which may contain feature extraction errors, effectively forces the model to rely on MGC+BAP for final speech reconstruction. This reliance can degrade overall sound quality, suggesting that reducing dependence on hand-crafted features while still preserving robust F_0 control is a key direction for future work.

V. Conclusion

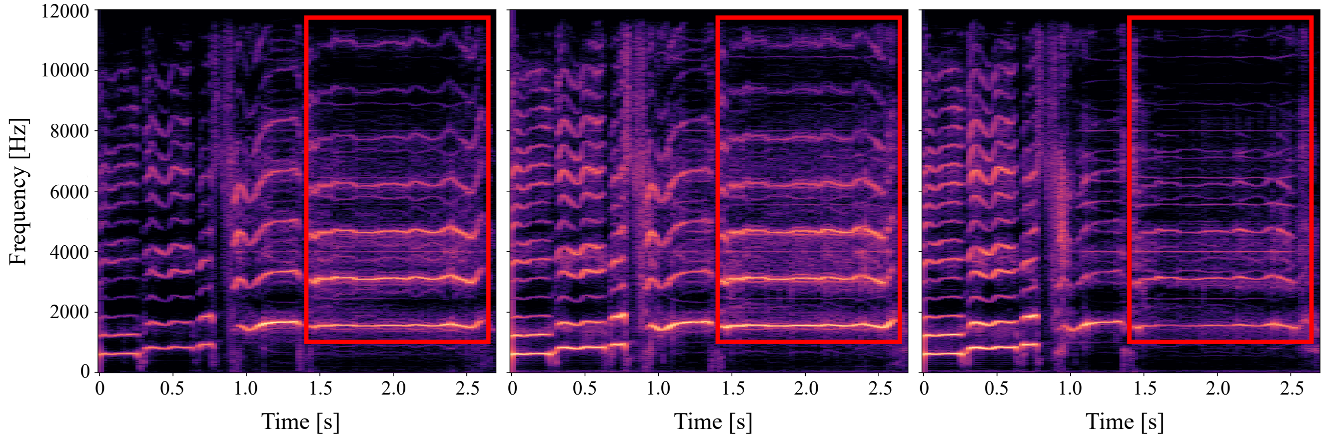
In this study, we propose VAE-SiFiGAN, designed to extend the applicability of SiFi-GAN by incorporating a learnable latent representation derived from Mel-spectrograms together with an F_0 removal mechanism. Experimental results demonstrate that VAE-SiFiGAN achieves superior F_0 controllability compared to conventional SiFi-GAN, which relies on hand-crafted acoustic features. Future work includes further refining the architecture to simultaneously ensure robust F_0 control and sound quality, verifying its robustness in noisy environments, investigating its integration into end-to-end applications such as SVS and TTS, and integrating a learnable F_0 predictor to enhance performance.

Acknowledgment

This work was partly supported by JST AIP Acceleration Research JPMJCR25U5, Japan.

References

[1] A. van den Oord, S. Dieleman, H. Zen, *et al.*, "WaveNet: A Generative Model for Raw Audio," in *Proc. SSW*, 2016, p. 125.
[2] A. van den Oord, Y. Li, I. Babuschkin, *et al.*, "Parallel WaveNet: Fast High-Fidelity Speech Synthesis," in *Proc. ICML*, 2018, pp. 3915–3923.
[3] N. Kalchbrenner, E. Elsen, K. Simonyan, *et al.*, "Efficient Neural Audio Synthesis," in *Proc. ICML*, 2018, pp. 2415–2424.
[4] J.-M. Valin and J. Skoglund, "LPCNet: Improving Neural Speech Synthesis Through Linear Prediction," in *Proc. ICASSP*, 2019, pp. 5891–5895.
[5] R. Prenger, R. Valle, and B. Catanzaro, "WaveGlow: A Flow-based Generative Network for Speech Synthesis," in *Proc. ICASSP*, 2019, pp. 3617–3621.
[6] 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. ICASSP*, 2020, pp. 6199–6203.
[7] K. Kumar, R. Kumar, T. de Boissiere, *et al.*, "MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis," in *Proc. NeurIPS*, 2019, pp. 14 910–14 921.
[8] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," in *Proc. NeurIPS*, 2020, pp. 17 022–17 033.
[9] H. Kawahara, "Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited.," in *Proc. ICASSP*, vol. 2, 1997, pp. 1303–1306.
[10] 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.



(a) SiFi-GAN (b) VAE-SiFiGAN (c) w/o Prior

Fig. 3: Spectrograms of generated singing voices.

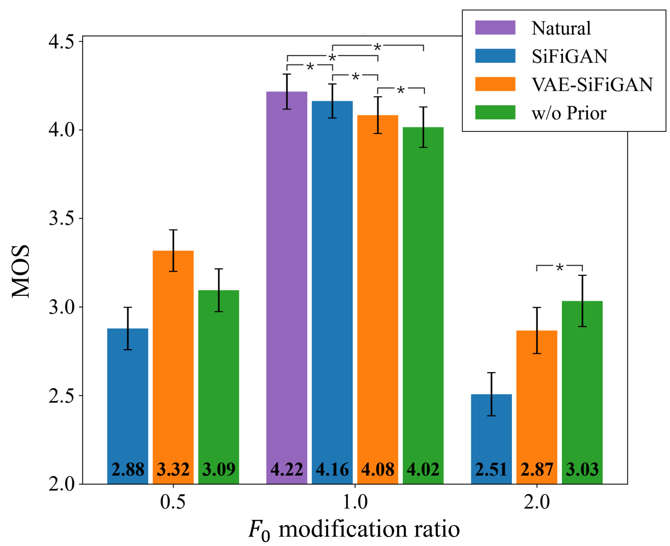


Fig. 4: Results of subjective evaluation. Error bars indicate 95% confidence intervals. There is no statistical difference ($p > 0.05$) between any pairs marked with an asterisk.

[11] R. Yoneyama, Y.-C. Wu, and T. Toda, "Source-Filter HiFiGAN: Fast and Pitch Controllable High-Fidelity Neural Vocoder," in *Proc. ICASSP*, 2023, pp. 1–5.

[12] T. Toda and K. Tokuda, "A Speech Parameter Generation Algorithm Considering Global Variance for HMM-Based Speech Synthesis," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 5, pp. 816–824, 2007.

[13] S. Takamichi, T. Toda, A. Black, G. Neubig, S. Sakti, and S. Nakamura, "Postfilters to Modify the Modulation Spectrum for Statistical Parametric Speech Synthesis," *IEEE/ACM TASLP*, vol. 24, no. 4, pp. 755–767, 2016.

[14] 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 TASLP*, vol. 29, pp. 792–806, 2021.

[15] R. Yoneyama, Y.-C. Wu, and T. Toda, "High-Fidelity and Pitch-Controllable Neural Vocoder Based on Unified Source-Filter Networks," *IEEE/ACM TASLP*, vol. 31, pp. 3717–3729, 2023.

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

[17] Y. Hono, K. Hashimoto, Y. Nankaku, and K. Tokuda, "PeriodGrad: Towards Pitch-Controllable Neural Vocoder Based on a Diffusion Probabilistic Model," in *Proc. ICASSP*, 2024, pp. 12 782–12 786.

[18] Y. Ohtani, T. Okamoto, T. Toda, and H. Kawai, "FIRNet: Fundamental Frequency Controllable Fast Neural Vocoder With Trainable Finite Impulse Response Filter," in *Proc. ICASSP*, 2024, pp. 10 871–10 875.

[19] D. P. Kingma and M. Welling, "Auto-encoding Variational Bayes," in *Proc. ICLR*, 2014, pp. 1–5.

[20] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier Nonlinearities Improve Neural Network Acoustic Models," in *Proc. ICML*, 2013, pp. 3–11.

[21] 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 TASLP*, vol. 29, pp. 1134–1148, 2021.

[22] X. Wang, S. Takaki, and J. Yamagishi, "Neural Source-filter-based Waveform Model for Statistical Parametric Speech Synthesis," *IEEE/ACM TASLP*, vol. 28, pp. 402–415, 2020.

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

[24] J. Kim, J. Kong, and J. Son, "Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech," in *Proc. ICML*, 2021, pp. 5530–5540.

[25] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search," in *Proc. NeurIPS*, 2020, pp. 8067–8077.

[26] M. Morise, "D4C, A band-aperiodicity estimator for high-quality speech synthesis," *Speech Communication*, vol. 84, pp. 57–65, 2015.

[27] Canon, *[NamineRitsu] Blue (YOASOBI) [ENUNU model Ver.2, Singing DBVer.2 release]*, https://www.youtube.com/watch?v=pKeo9IE_L1I, Accessed: 2024.1.30.

[28] M. Morise, "Harvest: A High-Performance Fundamental Frequency Estimator from Speech Signals," in *Proc. Interspeech*, 2017, pp. 2321–2325.

[29] K. Yu and S. Young, "Continuous F0 modeling for HMM based statistical parametric speech synthesis," *IEEE/ACM TASLP*, vol. 19, no. 5, pp. 1071–1079, 2010.

[30] W. Jang, D. Lim, J. Yoon, B. Kim, and J. Kim, "UnivNet: A Neural Vocoder with Multi-Resolution Spectrogram Discriminators for High-Fidelity Waveform Generation," in *Proc. Interspeech*, 2021, pp. 2207–2211.

[31] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *Proc. ICLR*, 2015, pp. 1–15.