MULTI-SCALE SUB-BAND CONSTANT-Q TRANSFORM DISCRIMINATOR FOR HIGH-FIDELITY VOCODER

Yicheng Gu *Xueyao Zhang* *Liumeng Xue* *Zhizheng Wu*

School of Data Science, The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen), China

ABSTRACT

Generative Adversarial Network (GAN) based vocoders are superior in inference speed and synthesis quality when reconstructing an audible waveform from an acoustic representation. This study focuses on improving the discriminator to promote GAN-based vocoders. Most existing time-frequency-representation-based discriminators are rooted in Short-Time Fourier Transform (STFT), whose time-frequency resolution in a spectrogram is fixed, making it incompatible with signals like singing voices that require flexible attention for different frequency bands. Motivated by that, our study utilizes the Constant-Q Transform (CQT), which owns dynamic resolution among frequencies, contributing to a better modeling ability in pitch accuracy and harmonic tracking. Specifically, we propose a Multi-Scale Sub-Band CQT (MS-SB-CQT) Discriminator, which operates on the CQT spectrogram at multiple scales and performs sub-band processing according to different octaves. Experiments conducted on both speech and singing voices confirm the effectiveness of our proposed method. Moreover, we also verified that the CQT-based and the STFT-based discriminators could be complementary under joint training. Specifically, enhanced by the proposed MS-SB-CQT and the existing MS-STFT Discriminators, the MOS of HiFi-GAN can be boosted from 3.27 to 3.87 for seen singers and from 3.40 to 3.78 for unseen singers.

***Index Terms*—** Neural vocoder, constant-Q transform, generative adversarial networks (GAN), discriminator

1. INTRODUCTION

A neural vocoder reconstructs an audible waveform from an acoustic representation. Deep generative models including autoregressive-based [1, 2], flow-based [3, 4], GAN-based [5–11], and diffusion-based [12, 13] models have been successful for this task. Because of the superior inference speed and synthesis quality, GAN-based vocoders are always attractive to researchers. However, to synthesize expressive speech or singing voice, current GAN-based vocoders still hold problems like spectral artifacts such as hissing noise [9] and loss of details in mid and low-frequency parts [10].

To pursue high-quality GAN-based vocoders, the existing studies aim to improve both the generator and the discriminator. For the generator, SingGAN [10] adopts a neural source filter [14] module to utilize the sine excitation. BigVGAN [11] introduces a new activation function with anti-aliasing modules. For the discriminator, MelGAN [6] employs a time-domain-based discriminator that successfully models waveform structures at different scales for the first time. HiFi-GAN [8] extends it with a Multi-Scale Discriminator and Multi-Period Discriminator, and Fre-GAN [9] further improves it by replacing the averaging pooling with discrete-wavelet-transform-based filters to preserve frequency information. UniversalMelGAN [7] introduces a Multi-Resolution Discriminator, followed by [15] emphasizing its significance. Encodec [16] extends it to the Multi-Scale STFT (MS-STFT) Discriminator.

This study focuses on improving the discriminator. Among the existing works, most time-frequency-representation-based discriminators are rooted in Short-Time Fourier Transform (STFT) [7, 15, 16], which could fast extract easy-to-handle STFT spectrograms for neural networks. However, it also has limitations. Specifically, an STFT spectrogram has a fixed time-frequency resolution across all frequency bins (Section 2.1). When encountering signals like singing voices, which require different attention for different frequency bands [17], only an STFT spectrogram will be insufficient.

Motivated by that, this paper proposes a Constant-Q Transform (CQT) [18] based discriminator. The reason is that CQT has a more flexible resolution for different frequency bands than STFT. In the low-frequency band, CQT has a higher *frequency resolution*, which can model the pitch information accurately. In the high-frequency band, CQT has a higher *time resolution*, which can track the fast-changing harmonic variations. In addition, the CQT has log-scale distributed center frequencies, which can bring better pitch-level information [19]. Specifically, we design a Multi-Scale Sub-Band CQT (MS-SB-CQT) Discriminator. The discriminator operates on CQT spectrograms at different scales and performs sub-band processing according to the octave information of the CQT spectrogram. Moreover, during the experiments, we find that the proposed MS-SB-CQT and MS-STFT [16] Discriminators can be jointly used to boost the generator further, which reveals the complementary role between the CQT-based and the STFT-based discriminators.

2. MULTI-SCALE SUB-BAND CONSTANT-Q TRANSFORM DISCRIMINATOR

The architecture of the proposed MS-SB-CQT Discriminator, which can be integrated into any GAN-based vocoders, is illustrated in Fig. 1. It consists of identically structured sub-discriminators operating on CQT spectrograms in different scales. Each sub-discriminator will first send the real and imaginary parts of CQT to our proposed Sub-Band Processing (SBP) module individually to get their latent representations. These two representations will then be concatenated and sent to convolutional layers to get the outputs for computing loss. The details of each module will be introduced as follows.

2.1. The Strengths of Constant-Q Transform

In this section, the strengths of CQT will be exhibited by introducing its design idea. We will see how CQT owns a flexible time-frequency resolution and why it can model pitch-level information better.

Following [18], the CQT $\boldsymbol{X}^{cq}(k,n)$ can be defined as:

$$\boldsymbol{X}^{cq}(k,n) = \sum_{j=n-\lfloor N_k/2 \rfloor}^{n+\lfloor N_k/2 \rfloor} x(j) a_k^*(j-n+N_k/2), \qquad (1)$$

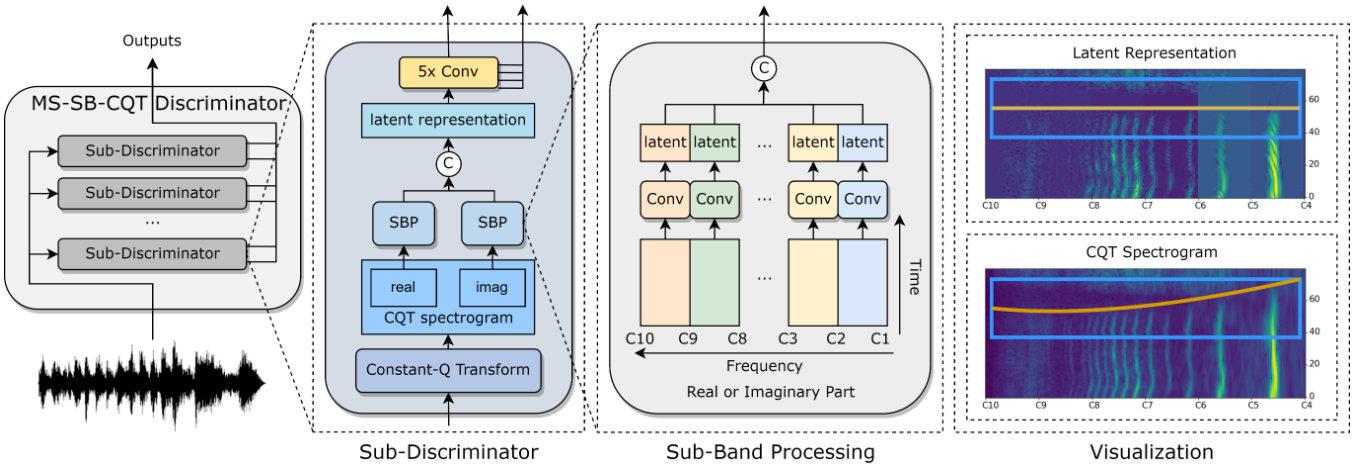


Fig. 1: Architecture of the proposed Multi-Scale Sub-Band Constant-Q Transform (MS-SB-CQT) Discriminator, which can be integrated with any GAN-based vocoder. Operator "C" denotes for concatenation. SBP means our proposed Sub-Band Processing module. It can be observed that the desynchronized CQT Spectrogram (bottom-right) has been synchronized (upper-right) after SBP.

where k is the index of frequency bin, $x(j)$ is the j-th sample point of the analyzed signal, N_k is the window length, $a_k(n)$ is a complex-valued kernel, and $a_k^*(n)$ is the complex conjugate of $a_k(n)$.

The kernels $a_k(n)$ can be obtained as:

$$a_k(n) = \frac{1}{N_k} w\left(\frac{n}{N_k}\right) e^{-i2\pi n \frac{Q_k}{N_k}}, \tag{2}$$

where $w(t)$ is the window function, and Q_k is the constant Q-factor:

$$Q_k \overset{ref.}{=} \frac{f_k}{\Delta f_k} = (2^{\frac{1}{B}} - 1)^{-1}, \tag{3}$$

where f_k is the center frequency, Δf_k is the bandwidth determining the resolution trade-off, and B is the number of bins per octave.

Notably, for STFT, the Δf_k is constant, meaning the time-frequency resolution is fixed for all frequencies. However, for CQT, its main idea is to keep Q_k constant. As a result, the low-frequency bands will have a smaller Δf_k, bringing a higher *frequency resolution*, which could model the pitch information better. Besides, the high-frequency bands will have a bigger Δf_k, bringing a higher *time resolution*, which could track fast-changing harmonics variations better.

In Eq. (2), the window length of the k-th frequency bin, N_k, can be obtained as:

$$N_k = \frac{f_s}{\Delta f_k} = \frac{f_s}{f_k} \cdot (2^{\frac{1}{B}} - 1)^{-1} \tag{4}$$

where f_s is the sampling rate, and f_k is defined as:

$$f_k = f_1 \cdot 2^{\frac{k-1}{B}}, \tag{5}$$

where f_1 is the lowest center frequency, which is set to 32.7 Hz (C1) in our study.

2.2. Multi-Scale Sub-Discriminators

To capture the information under more diverse time-frequency resolutions, we leverage the multi-scale idea [8, 15] and adopt sub-discriminators on CQTs with different overall resolution trade-offs.

Given Eq. (3) and (5), we can observe that the bandwidth Δf_k, which determines the resolution trade-off, is dependent on the number of bins per octave B. In other words, we can set the different B to obtain the different resolution distributions. Based on that, we follow [15, 16] to apply three sub-discriminators with B equals 24, 36, and 48, respectively.

2.3. Sub-Band Processing Module

As two sides of a coin, although the dynamic bandwidth Δf_k brings flexible time-frequency resolution, it also brings the unfixed window length N_k. As a result, the kernels $a_k(n)$ in different frequency bins are not temporally synchronized [20]. The CQT spectrogram with such artifacts has been visualized in the bottom right of Fig. 1.

To alleviate this problem, [20] designs a series of kernels that are temporally synchronized within an octave. This algorithm has also been used in toolkits like librosa [21] and nnAudio [22]. However, such an algorithm only makes the $a_k(n)$ of *intra-octave* temporally synchronized but leaves those of *inter-octave* unsolved. During experiments, we found that just using CQT spectrograms with such a bias could even hurt the quality of vocoders (Section 3.4).

Based on that, we utilize the philosophy of representation learning and design the Sub-Band Processing (SBP) module to address this problem further. In particular, the real or imaginary part of a CQT spectrogram will first be split into sub-bands according to octaves. Then, each band will be sent to its corresponding convolutional layer to get its representation. Finally, we concatenate the representations from all bands to obtain the latent representation of the CQT spectrogram. In the upper right of Fig. 1, it can be observed that our proposed SBP successfully learns the temporally synchronized representations among all the frequency bins.

2.4. Integration with GAN-based Vocoder

Our proposed discriminator can be easily integrated with existing GAN-based vocoders without interfering with the inference stage. We take HiFi-GAN [8] as an example. HiFi-GAN has a generator G and multiple discriminators D_m. The generation loss $\mathcal{L}_G$, and discrimination loss $\mathcal{L}_D$ are defined as, $\mathcal{L}_G = \sum_{m=1}^{M}[\mathcal{L}_{adv}(G; D_m) +$

Table 1: Analysis-synthesis results of different discriminators when being integrated into HiFi-GAN [8]. The best and the second best results of every column (except those from Ground Truth) in each domain (speech and singing voice) are **bold** and *italic*. "S" and "C" represent MS-STFT and MS-SB-CQT Discriminators respectively. The MOS scores are with 95% Confidence Interval (CI).

Domain	System	MCD (↓)		PESQ (↑)		FPC (↑)		F0RMSE (↓)		MOS (↑)	
		Seen	Unseen	Seen	Unseen	Seen	Unseen	Seen	Unseen	Seen	Unseen
Singing voice	Ground Truth	0.00	0.00	4.50	4.50	1.000	1.000	0.00	0.00	4.85 ± 0.06	4.73 ± 0.09
	HiFi-GAN	*2.82*	*3.17*	2.94	2.86	0.954	0.961	56.96	59.28	3.27 ± 0.16	3.40 ± 0.15
	HiFi-GAN (+S)	2.97	3.37	2.95	2.87	0.967	0.968	39.06	46.49	3.42 ± 0.16	3.56 ± 0.17
	HiFi-GAN (+C)	2.90	3.35	*3.03*	*2.95*	*0.970*	*0.971*	*35.57*	*41.09*	*3.66 ± 0.14*	*3.63 ± 0.16*
	HiFi-GAN (+S+C)	**2.54**	**3.08**	**3.09**	**2.98**	**0.971**	**0.973**	**35.45**	**39.90**	**3.87 ± 0.14**	**3.78 ± 0.12**
Speech	Ground Truth	0.00	0.00	4.50	4.50	1.000	1.000	0.00	0.00	4.62 ± 0.11	4.59 ± 0.11
	HiFi-GAN	*3.21*	2.10	3.01	3.14	*0.883*	*0.781*	186.19	*293.34*	3.91 ± 0.17	3.96 ± 0.16
	HiFi-GAN (+S)	3.47	2.10	2.97	3.09	0.869	0.772	195.05	298.53	**4.02 ± 0.15**	4.00 ± 0.17
	HiFi-GAN (+C)	3.26	*2.07*	*3.04*	**3.16**	**0.884**	0.768	**180.29**	301.83	4.01 ± 0.15	*4.13 ± 0.14*
	HiFi-GAN (+S+C)	**3.13**	**2.05**	**3.05**	*3.15*	*0.883*	**0.792**	*182.04*	**281.90**	**4.02 ± 0.17**	**4.14 ± 0.15**

$2\mathcal{L}_{fm}(G; D_m)] + 45\mathcal{L}_{mel}, \mathcal{L}_D = \sum_{m=1}^{M}[\mathcal{L}_{adv}(D_m; G)$, where M is the number of discriminators, D_m denotes the m-th discriminator, $\mathcal{L}_{adv}$ is the adversarial GAN loss, $\mathcal{L}_{fm}$ is the feature matching loss, and $\mathcal{L}_{mel}$ is the mel spectrogram reconstruction loss. Among these losses, only $\mathcal{L}_{fm}$ and $\mathcal{L}_{adv}$ are related to our discriminator. Thus, just adding $\mathcal{L}_{adv}(G; D_{\text{MS-SB-CQT}}) + 2\mathcal{L}_{fm}(G; D_{\text{MS-SB-CQT}})$ to $\mathcal{L}_G$ and $\mathcal{L}_{adv}(D_{\text{MS-SB-CQT}}; G)$ to $\mathcal{L}_D$ can integrate the proposed discriminator in the training process.

3. EXPERIMENTS

We conduct experiments to investigate the following four questions. **EQ1**: How effective is the proposed MS-SB-CQT Discriminator? **EQ2**: Could using MS-SB-CQT and MS-STFT Discriminators jointly improve the vocoder further? **EQ3**: How generalized is the MS-SB-CQT Discriminator under different GAN-based vocoders? **EQ4**: Is it necessary to adopt the proposed SBP module? Audio samples are available on our demo site[1].

3.1. Experimental setup

Dataset The experimental datasets contain both speech and singing voices. For the singing voice, we adopt M4Singer [23], PJS [24], and one internal dataset. We randomly sample 352 utterances from the three datasets to evaluate ***seen*** singers and leave the remaining for training (39 hours). 445 samples from Opencpop [25], PopCS [26], OpenSinger [27], and CSD [28] are chosen to evaluate ***unseen*** singers. For the speech, we use the train-clean-100 from LibriTTS [29] and LJSpeech [30]. We randomly sample 2316 utterances from the two datasets to evaluate ***seen*** speakers and leave the remaining for training (about 75 hours). 3054 samples from VCTK [31] are chosen to evaluate ***unseen*** speakers.

Implementation Details The CNN in SBP uses a Conv2D with kernel size of (3, 9). The CNNs in each Sub-Discriminator consist of a Conv2D with kernel size (3, 8) and 32 channels, three Conv2Ds with dilation rates of [1, 2, 4] in the time dimension and a stride of 2 over the frequency dimension, and a Conv2D with kernel size (3, 3) and stride (1, 1). For CQT, the global hop length is empirically set to 256, and the waveform will be upsampled from f_s to $2f_s$ before the computation to avoid the f_{max} of the top octave hitting the Nyquist Frequency.

[1] https://vocodexelysium.github.io/MS-SB-CQTD/

Evaluation Metrics For objective evaluation, we use Perceptual Evaluation of Speech Quality (PESQ) [32] and Mel Cepstral Distortion (MCD) [33] to evaluate the spectrogram reconstruction. We use F0 Root Mean Square Error (F0RMSE) and F0 Pearson Correlation Coefficient (FPC) for evaluating pitch stability. The Mean Opinion Score (MOS) and Preference Test are used for subjective evaluation. We invited 20 volunteers who are experienced in the audio generation area to attend the subjective evaluation. Each setting in the bellowing MOS test has been graded 200 times, and each pair in the preference test has been graded 120 times.

3.2. Effectiveness of MS-SB-CQT Discriminator (EQ1 & EQ2)

To verify the effectiveness of the proposed discriminator, we take HiFi-GAN as an example and enhance it with different discriminators. The results of the analysis-synthesis are illustrated in Table 1. Regarding singing voice, we can observe that: (1) both HiFi-GAN (+C) and HiFi-GAN (+S) perform better than HiFi-GAN, showing the importance of time-frequency-representation-based discriminators [15]; (2) HiFi-GAN (+C) performs better than HiFi-GAN (+S) with a significant boost in MOS, showing the superiority of our proposed MS-SB-CQT Discriminator; (3) HiFi-GAN (+S+C) performs best both objectively and subjectively, which shows that different discriminators will have complementary information for each other, confirming the effectiveness of jointly training. A similar conclusion can be drawn for the unseen speaker evaluation of speech data.

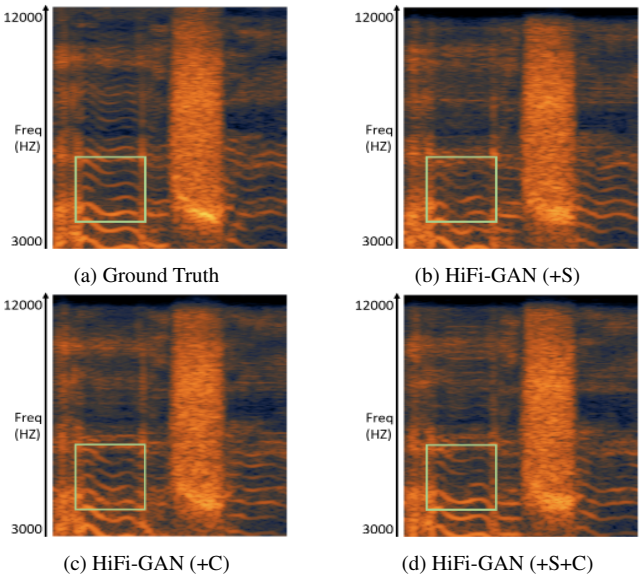
To further explore the specific benefits of using the CQT-based and the STFT-based discriminators jointly, we conducted a case study (Fig. 2). Notably, in the displayed high-frequency parts, STFT has a better frequency resolution, and CQT has a better time resolution. Regarding the frequency parts in the rectangle, it can be observed that: (1) HiFi-GAN with MS-STFT Discriminator (Fig. 2b) can reconstruct its frequency accurately but cannot track the changes due to insufficient time resolution; (2) HiFi-GAN with MS-SB-CQT Discriminator (Fig. 2c) can track the harmonics, but the frequency reconstruction is inaccurate due to the low-frequency resolution; (3) Integrating those two combines their strengths and thus achieve a better reconstruction quality (Fig. 2d).

3.3. Generalization Ability of MS-SB-CQT Discriminator (EQ3)

To verify the generalization ability of the proposed MS-SB-CQT Discriminator, besides HiFi-GAN, we also conduct experiments un-

Table 2: Analysis-synthesis results of our proposed MS-SB-CQT Discriminator when integrating in MelGAN [6] and NSF-HiFiGAN in singing voice datasets. The improvements are shown in **bold**. "S" and "C" represent MS-STFT and MS-SB-CQT Discriminators respectively. All the improvements in MCD, PESQ, and Preference are significant (p-value < 0.01).

System	MCD (↓)		PESQ (↑)		FPC (↑)		F0RMSE (↓)		Preference (↑)	
	Seen	Unseen	Seen	Unseen	Seen	Unseen	Seen	Unseen	Seen	Unseen
Ground Truth	0.00	0.00	4.50	4.50	1.000	1.000	0.00	0.00	/	/
MelGAN	4.44	5.21	2.23	2.15	0.968	0.964	46.80	51.73	8.47%	27.45%
MelGAN (+S+C)	**4.08**	**4.87**	**2.35**	**2.23**	0.960	0.962	51.78	**50.99**	**91.53%**	**72.55%**
NSF-HiFiGAN	1.73	2.04	3.95	3.88	0.985	0.980	25.62	31.17	41.67%	29.41%
NSF-HiFiGAN (+S+C)	**1.48**	**1.72**	**3.98**	**3.91**	0.979	**0.983**	**24.01**	31.19	**58.33%**	**70.59%**



(a) Ground Truth (b) HiFi-GAN (+S)

(c) HiFi-GAN (+C) (d) HiFi-GAN (+S+C)

Fig. 2: The comparison of mel spectrograms of HiFi-GANs enhanced by different discriminators. "S" and "C" represent MS-STFT and MS-SB-CQT Discriminators respectively. Integrating with both CQT- and STFT-based discriminators, HiFi-GAN could achieve a higher synthesis quality with more accurate harmonic tracking and frequency reconstruction.

der MelGAN[2] [6] and NSF-HiFiGAN[3]. Note that NSF-HiFiGAN is one of the state-of-the-art vocoders for singing voice [34]. It combines the neural source filter (NSF) [14] to enhance the generator of HiFi-GAN. The experimental results are presented in Table 2.

It is illustrated that: (1) In general, the performance of MelGAN and NSF-HiFiGAN can be improved significantly by jointly training with MS-SB-CQT and MS-STFT Discriminators, with both objective and subjective preference tests confirming the effectiveness; (2) In particular, MelGAN tends to overfit the low-frequency part and ignore mid and high-frequency components, resulting in audible metallic noise. After adding MS-STFT and MS-SB-CQT Discriminators, it could model the global information of spectrogram better[4], bringing in significantly better MCD and PESQ. Although the low-frequency-related metrics worsen, the preference test shows that the overall quality has remarkably increased; NSF-HiFiGAN can synthesize high-fidelity singing voices. However, it still lacks frequency details. Adding MS-STFT and MS-SB-CQT Discriminators tackles that problem[4], making synthesized samples closer to the ground truth. Subjective results with a higher preference percentage also demonstrate the effectiveness.

[2] https://github.com/descriptinc/melgan-neurips/

[3] https://github.com/nii-yamagishilab/project-NN-Pytorch-scripts

[4] We show the representative cases on the demo page.

3.4. Necessity of Sub-Band Processing (EQ4)

As introduced in Section 2.3, we propose the Sub-Band Processing module to obtain the temporally synchronized CQT latent representations. To verify the necessity of it, we conduct an ablation study that removes the SBP module from the proposed MS-SB-CQT Discriminator. We adopt Opencpop [25] as the experimental dataset. We randomly selected 221 utterances for evaluation and the remaining for training (about 5 hours).

Table 3: Analysis-synthesis results of HiFi-GAN enhanced by different CQT-based discriminators. MS-CQT Discriminator represents a discriminator that only removes the Sub-Band Processing module from our proposed MS-SB-CQT Discriminator.

System	MCD (↓)	PESQ (↑)	FPC (↑)	F0RMSE (↓)
HiFi-GAN	3.443	2.960	0.972	40.409
+ MS-CQT	3.502	2.932	0.964	50.918
+ MS-SB-CQT	**3.263**	**2.985**	**0.986**	**28.313**

In Tabel 3, we can see that HiFi-GAN can be enhanced successfully by our proposed MS-SB-CQT Discriminator. However, just applying the raw CQT to the discriminator (MS-CQT) would even harm the quality of HiFi-GAN. We speculate this is because the temporal desynchronization in inter-octaves of the raw CQT would burden the model learning. Therefore, it is necessary to adopt the proposed SBP module for designing a CQT-based discriminator.

4. CONCLUSION AND FUTURE WORKS

This study proposed a Multi-Scale Sub-Band Constant-Q Transform (MS-SB-CQT) Discriminator for GAN-based vocoder. The proposed discriminator outperforms the existing Multi-Scale Short-Time-Fourier-Transform (MS-STFT) Discriminator on both speech and singing voice. Besides, the proposed CQT-based discriminator can complement the existing STFT-based discriminator to improve the vocoder further. In future work, we will explore more time-frequency representations and other signal-processing approaches for better discriminators or generators.

5. REFERENCE

[1] Aäron van den Oord, et al., "Wavenet: A generative model for raw audio," in *SSW*. 2016, p. 125, ISCA.

[2] Nal Kalchbrenner, et al., "Efficient neural audio synthesis," in *ICML*. 2018, vol. 80, pp. 2415–2424, PMLR.

[3] Ryan Prenger, et al., "Waveglow: A flow-based generative network for speech synthesis," in *ICASSP*. 2019, pp. 3617–3621, IEEE.

[4] Wei Ping, et al., "Waveflow: A compact flow-based model for raw audio," in *ICML*. 2020, vol. 119, pp. 7706–7716, PMLR.

[5] Ryuichi Yamamoto, et al., "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *ICASSP*. 2020, pp. 6199–6203, IEEE.

[6] Kundan Kumar, et al., "Melgan: Generative adversarial networks for conditional waveform synthesis," in *NeurIPS*, 2019, pp. 14881–14892.

[7] Won Jang, et al., "Universal melgan: A robust neural vocoder for high-fidelity waveform generation in multiple domains," *arXiv*, vol. abs/2011.09631, 2020.

[8] Jiaqi Su, et al., "Hifi-gan: High-fidelity denoising and dereverberation based on speech deep features in adversarial networks," in *INTERSPEECH*. 2020, pp. 4506–4510, ISCA.

[9] Ji-Hoon Kim, et al., "Fre-gan: Adversarial frequency-consistent audio synthesis," in *INTERSPEECH*. 2021, pp. 2197–2201, ISCA.

[10] Rongjie Huang, et al., "Singgan: Generative adversarial network for high-fidelity singing voice generation," in *ACM Multimedia*. 2022, pp. 2525–2535, ACM.

[11] Sang-gil Lee, et al., "Bigvgan: A universal neural vocoder with large-scale training," in *ICLR*. 2023, OpenReview.net.

[12] Nanxin Chen, et al., "Wavegrad: Estimating gradients for waveform generation," in *ICLR*. 2021, OpenReview.net.

[13] Zhifeng Kong, et al., "Diffwave: A versatile diffusion model for audio synthesis," in *ICLR*. 2021, OpenReview.net.

[14] Xin Wang, et al., "Neural source-filter-based waveform model for statistical parametric speech synthesis," in *ICASSP*. 2019, pp. 5916–5920, IEEE.

[15] Jaeseong You, et al., "GAN vocoder: Multi-resolution discriminator is all you need," in *INTERSPEECH*. 2021, pp. 2177–2181, ISCA.

[16] Alexandre Défossez, et al., "High fidelity neural audio compression," *arXiv*, vol. abs/2210.13438, 2022.

[17] Rongjie Huang, et al., "Multi-singer: Fast multi-singer singing voice vocoder with A large-scale corpus," in *ACM Multimedia*. 2021, pp. 3945–3954, ACM.

[18] Judith C. Brown and Miller Puckette, "An efficient algorithm for the calculation of a constant q transform," *Journal of the Acoustical Society of America*, vol. 92, pp. 2698–2701, 1992.

[19] Yizhi Li, et al., "MERT: acoustic music understanding model with large-scale self-supervised training," *arXiv*, vol. abs/2306.00107, 2023.

[20] Christian Schörkhuber and Anssi Klapuri, "Constant-q transform toolbox for music processing," in *Sound and Music Computing Conference*, 2010, pp. 3–64.

[21] Brian McFee, et al., "librosa: Audio and music signal analysis in python," in *SciPy*. 2015, pp. 18–24, scipy.org.

[22] Kin Wai Cheuk, et al., "nnaudio: An on-the-fly GPU audio to spectrogram conversion toolbox using 1d convolutional neural networks," *IEEE Access*, vol. 8, pp. 161981–162003, 2020.

[23] Lichao Zhang, et al., "M4singer: A multi-style, multi-singer and musical score provided mandarin singing corpus," in *NeurIPS*, 2022.

[24] Junya Koguchi, et al., "PJS: phoneme-balanced japanese singing-voice corpus," in *APSIPA*. 2020, pp. 487–491, IEEE.

[25] Yu Wang, et al., "Opencpop: A high-quality open source chinese popular song corpus for singing voice synthesis," in *INTERSPEECH*. 2022, pp. 4242–4246, ISCA.

[26] Jinglin Liu, et al., "Diffsinger: Singing voice synthesis via shallow diffusion mechanism," in *AAAI*. 2022, pp. 11020–11028, AAAI Press.

[27] Rongjie Huang, et al., "Multi-singer: Fast multi-singer singing voice vocoder with A large-scale corpus," in *ACM Multimedia*. 2021, pp. 3945–3954, ACM.

[28] Soonbeom Choi, et al., "Children's song dataset for singing voice research," in *ISMIR*, 2020.

[29] Heiga Zen, et al., "Libritts: A corpus derived from librispeech for text-to-speech," in *INTERSPEECH*. 2019, pp. 1526–1530, ISCA.

[30] Keith Ito and Linda Johnson, "The lj speech dataset," `https://keithito.com/LJ-Speech-Dataset/`, 2017.

[31] Junichi Yamagishi, et al., "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92)," *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 2019.

[32] Antony W. Rix, et al., "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *ICASSP*. 2001, pp. 749–752, IEEE.

[33] Robert Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*. IEEE, 1993, vol. 1, pp. 125–128.

[34] Wen-Chin Huang, et al., "The singing voice conversion challenge 2023," *arXiv*, vol. abs/2306.14422, 2023.