An Investigation of Time-Frequency Representation Discriminators for High-Fidelity Vocoder

Yicheng Gu, *Student Member, IEEE*, Xueyao Zhang, Liumeng Xue, *Member, IEEE*
Haizhou Li, *Fellow, IEEE*, Zhizheng Wu, *Senior Member, IEEE*

***Abstract*—Generative Adversarial Network (GAN) based vocoders are superior in both inference speed and synthesis quality when reconstructing an audible waveform from an acoustic representation. This study focuses on improving the discriminator for GAN-based vocoders. Most existing Time-Frequency Representation (TFR)-based discriminators are rooted in Short-Time Fourier Transform (STFT), which owns a constant Time-Frequency (TF) resolution, linearly scaled center frequencies, and a fixed decomposition basis, making it incompatible with signals like singing voices that require dynamic attention for different frequency bands and different time intervals. Motivated by that, we propose a Multi-Scale Sub-Band Constant-Q Transform CQT (MS-SB-CQT) discriminator and a Multi-Scale Temporal-Compressed Continuous Wavelet Transform CWT (MS-TC-CWT) discriminator. Both CQT and CWT have a dynamic TF resolution for different frequency bands. In contrast, CQT has a better modeling ability in pitch information, and CWT has a better modeling ability in short-time transients. Experiments conducted on both speech and singing voices confirm the effectiveness of our proposed discriminators. Moreover, the STFT, CQT, and CWT-based discriminators can be used jointly for better performance. The proposed discriminators can boost the synthesis quality of various state-of-the-art GAN-based vocoders, including HiFi-GAN, BigVGAN, and APNet.**

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

I. Introduction

A vocoder has always been important for various audio generation tasks (e.g., Singing Voice Synthesis (SVS) [1, 2], Text-To-Speech (TTS) [3, 4]). It reconstructs an audible waveform from an acoustic feature [3, 5, 6] outputted by the acoustic model, directly affecting the resulting audio quality and generation speed. Among different types of vocoders, the neural network-based ones are essential due to their superior synthesis quality compared to the DSP-based ones [7, 8].

The synthesis quality and the inference speed are two primary considerations when designing a neural vocoder. Initially, autoregressive-based vocoders, represented by WaveNet [9] and WaveRNN [10], first successfully model the speech signal via deep neural networks. Although the autoregressive-based vocoders achieved breakthroughs regarding synthesis quality, their inference speed cannot meet the need for real-world applications due to the degradation brought by the sample-by-sample generation scheme. Subsequently, distilling-based models (e.g., ParallelWaveNet [11], ClariNet [12]), flow-based models (e.g., WaveFlow [13], WaveGlow [14]), glottis-based models (e.g., GlotNet [15], LPCNet [16]t), diffusion-based models (e.g., DiffWave [17], FreGrad [18]), and differentialble digital signal processing (DDSP)-based models (e.g., NSF [19], GOLF [20]) were proposed. Although the inference efficiency for these models is significantly boosted, the synthesis quality was relatively degraded, making them incompatible with applications that require a high synthesis quality. Regarding this matter, GAN-based vocoders [21–28] are proposed and currently widely used. Specifically, the parallelized generation scheme via noncausal convolution layers ensures the inference speed, and the adversarial training, which makes it possible to utilize audio-level losses in different perspectives, ensures the synthesis quality. However, to synthesize expressive speech or singing voice, current GAN-based vocoders still hold problems like spectral artifacts such as ***hissing noise*** [25], ***glitches*** [29], and ***the loss of details in mid and low-frequency regions*** [26].

To pursue high-quality GAN-based vocoders, the existing studies aim to improve both the generator and the discriminator. For the generator, SingGAN [26] adopts a neural source filter (NSF) [19] module to utilize the phase-continuous sine excitation to alleviate the glitches [29]. BigVGAN [27] introduces a new activation function [30] with anti-aliasing modules to improve the generalization ability and the reconstruction quality in high-frequency bands. SnakeGAN [31] utilizes a fast DDSP vocoder [29] to generate raw audio for providing prior knowledge. For the discriminator, MelGAN [22] employs a time-domain-based discriminator that successfully models waveform structures at different scales for the first time. HiFi-GAN [24] extends it with a Multi-Scale Discriminator (MSD) and Multi-Period Discriminator (MPD), which operates on the reshaped audio signal obtained from average pooling and periodical sampling individually. FreGAN [25] further improves them by replacing the pooling and the sampling process with Discrete Wavelet Transform (DWT) to avoid aliasing effects. UniversalMelGAN [23] introduces a Multi-Resolution Discriminator that operates on the amplitude spectrogram of an STFT matrix, followed by [32] emphasizing its significance in boosting the synthesis quality. Furthermore, Encodec [5] extends it to the Multi-Scale STFT (MS-STFT) Discriminator by utilizing the phase information together.

Among the existing works, most Time-Frequency Representation (TFR)-based discriminators [5, 23, 32] are rooted in the Short-Time Fourier Transform (STFT) [33], which could fast extract easy-to-handle STFT spectrograms for neural networks. However, an STFT spectrogram has a constant Time-Frequency (TF) resolution, linearly scaled center frequencies, and a fixed decomposition basis. *When encountering signals like singing voices, which require dynamic attention for different frequency bands and different time intervals [26], only an STFT spectrogram will be insufficient.*

TABLE I: A comparison of the Short-Time Fourier Transform (STFT), Constant-Q Transform (CQT), and Continuous Wavelet Transform (CWT) in terms of decomposition basis, frequency distribution, and TF resolution.

Transform	TF Resolution	Basis	Frequency Distribution
STFT [33]	Constant	Fourier	Arithmetic Series
CQT [35]	Dynamic	Fourier	Geometric Series
CWT [36]	Dynamic	Wavelet	Harmonic Series

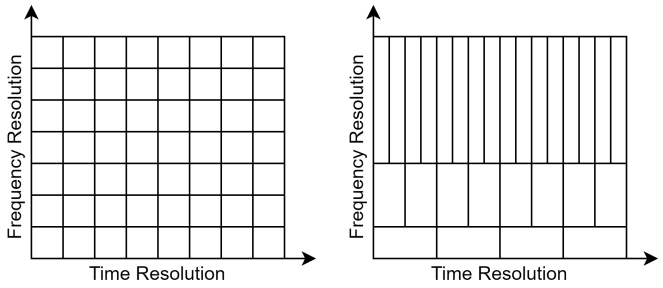
This study focuses on improving the discriminator part to improve the synthesis quality. In particular, we utilize the Constant-Q Transform (CQT) [34, 35] and the Continuous Wavelet Transform (CWT) [36] to design discriminators. Both CQT and CWT have a flexible resolution for different frequency bands, which brings a better modeling ability in F0 accuracy and harmonic tracking. Moreover, CQT has a better modeling ability in pitch-level information, and CWT has a better modeling ability in short-time transients. To make CQT and CWT spectrograms feasible with the GAN-based framework, we propose the Sub-Band Processor module and the Temporal Compressor module. Based on them, we design a Multi-Scale Sub-Band CQT (MS-SB-CQT) Discriminator [37] and a Multi-Scale Temporal-Compressed CWT (MS-TC-CWT) Discriminator that operate on CQT and CWT spectrograms in different scales individually. The main contributions of this paper are summarized as follows:

- To make the CQT feasible with the discriminator, we propose a Sub-Band Processor for CQT to tackle the temporal desynchronization issue in the CQT spectrogram.
- To make the CWT feasible with the GAN-based framework, we propose a Temporal Compressor for CWT to compress the high-dimensional CWT spectrogram into the low-dimensional latent representation.
- To increase the diversity of the discriminated features, we modify the Multi-Resolution Processing for CQT and CWT and propose a Multi-Basis Processing technique to integrate CWT into the same Multi-Scale Processing framework based on multiple sub-discriminators.
- To utilize the complementary role between different TFRs, we present a joint training strategy to use multiple discriminators based on STFT, CQT, and CWT.

To the best of our knowledge, this is the first study that employs multiple TF analysis techniques in a single GAN-based vocoder framework. Our proposed framework can improve the synthesis quality in the spectrogram and pitch stability without impacting the inference stage of the generator.

II. Background: Time-Frequency Analysis

Time-Frequency (TF) analysis aims to convert a time-domain signal into a TFR over both time and frequency domains. In this section, we will briefly introduce three classical TF analysis techniques: the Short-Time Fourier Transform (STFT), Constant-Q Transform (CQT), and Continuous Wavelet Transform (CWT). A comparison of the three transforms is presented in Table I,



(a) STFT (b) CQT and CWT

Fig. 1: Illustration of the TF resolution of the STFT, CQT, and CWT. A *thinner* chunk in the time/frequency axis means a better time/frequency resolution. It can be observed that the CQT and CWT spectrogram have a higher frequency resolution in low-frequency bands and a higher time resolution in high-frequency bands, while the STFT spectrogram has a fixed TF resolution across all frequency bands.

A. Short-Time Fourier Transform (STFT)

Short-Time Fourier Transform (STFT) [33] converts a time-domain signal x into its TFR, as:

$$\boldsymbol{X}^{\text{STFT}}(k,n) = \sum_{j=0}^{N_k} x(j+n-\lfloor N_k/2 \rfloor) w(j)\, e^{-i2\pi j \frac{Q_k}{N_k}}, \quad (1)$$

where k is the index of frequency bin, N_k is the window length, $x(t)$ denotes the t-th sample point, $w(t)$ is the window function, and Q_k is the Q-factor, which is defined as,

$$Q_k \overset{ref.}{=} \frac{f_k}{\Delta f_k}, \quad (2)$$

where f_k is the center frequency, Δf_k is the bandwidth. The center frequency f_k can be obtained as:

$$f_k = \frac{k f_s}{N_{fft}}, \quad (3)$$

where N_{fft} is the number of FFT bins. The bandwidth Δf_k, which determines the TF resolution trade-off, is defined as:

$$\Delta f_k = \frac{f_s}{N_k}, \quad (4)$$

where f_s is the sampling rate.

Although STFT can be easily implemented to model a speech signal, it also has drawbacks when encountering expressive speech or singing voice due to its characteristics (Table I), which are listed as follows:

- **Fixed TF resolution:** In STFT, the window length N_k is fixed, bringing the Δf_k a constant (Eq. 4). This means the TF resolution is fixed for all frequency bins. Thus, STFT is not a good candidate to model signals that require a dynamic resolution for different frequency bands.
- **Limitation in harmonics modeling:** The center frequency f_ks are linearly distributed in STFT (Eq. 3), which is incompatible with signals made up of harmonic frequency components that require geometrically distributed center frequencies for accurate modeling.

- **Inaccurate modeling of short-time transients:** According to the Gibbs phenomenon [38], when converging a square wave using the decomposition basis $e^{-i2\pi nk}$ in STFT, no matter how many harmonics are used, a non-neglectable approximation error will always exist, meaning a poor modeling ability in short-time transients.

B. Constant-Q Transform (CQT)

Constant-Q Transform (CQT) [34] converts a time-domain signal x into its TFR with a constant Q-factor, as:

$$\boldsymbol{X}^{CQT}(k,n) = \sum_{j=n-\lfloor N_k/2 \rfloor}^{n+\lfloor N_k/2 \rfloor} x(j) a_k^*(j-n+N_k/2), \quad (5)$$

where $a_k(n)$ is a complex-valued convolutional kernel, and $a_k^*(n)$ is the complex conjugate of $a_k(n)$. $\lfloor \cdot \rfloor$ denotes rounding down. The kernels $a_k(n)$ can be obtained as:

$$a_k(n) = \frac{1}{N_k} w\left(\frac{n}{N_k}\right) e^{-i2\pi n \frac{Q_k}{N_k}}, \quad (6)$$

where the quality factor Q_ks are defined constantly to conform with the human auditory system [39, 40]:

$$Q_k \overset{ref.}{=} \frac{f_k}{\Delta f_k} = (2^{\frac{1}{B}} - 1)^{-1}, \quad (7)$$

where B is the number of bins per octave. The center frequencies f_k in CQT are defined as:

$$f_k = f_1 \cdot 2^{\frac{k-1}{B}}, \quad (8)$$

where f_1 is the lowest center frequency.

Compared with STFT, CQT overcomes the fixed TF resolution and the linearly-distributed center frequencies (Table I), resulting in the following advantages:

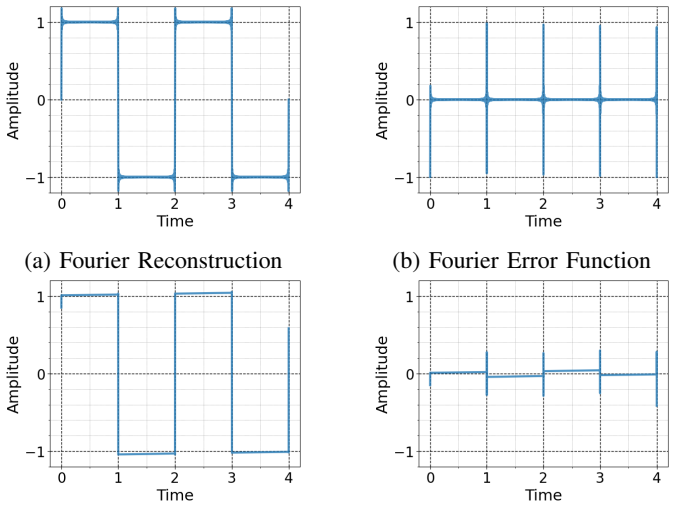
- **Dynamic TF resolution:** CQT utilizes a constant Q-factor (Eq. 7) to obtain a dynamic TF resolution across different frequency bins. Thus, as illustrated in Fig. 1, the low-frequency bands will have a smaller Δf_k, bringing a higher *frequency resolution*, which could model the F0 more accurately; The high-frequency bands will have a bigger Δf_k, bringing a higher *time resolution*, which could track fast-changing harmonics variations better.
- **Enhanced capability in harmonic modeling:** For a better modeling ability with the sound that is made up of harmonic components, CQT utilizes a series of geometrically distributed center frequencies (Eq. 8). In our study, we set f_1 to 32.7 Hz (C1) to ensure the center frequencies conform with the notes in Western Music.

C. Continuous Wavelet Transform (CWT)

Continuous Wavelet Transform (CWT) [36] converts a time-domain signal x into its TFR with shifting and scaling, as:

$$\boldsymbol{X}^{CWT}(k,n) = \frac{1}{|a_k|^{\frac{1}{2}}} \sum_{j=1}^{N} x(j) \psi^*\left(\frac{j-n}{a_k}\right), \quad (9)$$

where a_k is the scaling factor, $\psi(n)$ is the mother wavelet, and $\frac{1}{|a_k|^{\frac{1}{2}}} \psi(\frac{j-n}{a_k})$ is the scaled child wavelet.



(a) Fourier Reconstruction (b) Fourier Error Function

(c) Wavelet Reconstruction (d) Wavelet Error Function

Fig. 2: The visualization of the reconstructed square wave and the associated error function with different decomposition basis. It can be observed that the wavelet basis can reconstruct the signal with a smaller error regarding the step transient.

By taking the FT between the mother wavelet and the child wavelet, the bandwidth Δf_k can be obtained as:

$$\Delta f_k = \frac{\Delta f}{a_k}, \quad (10)$$

where Δf is the bandwidth of the mother wavelet. The relationship between the scaling factor a_k and the center frequency f_k is defined as:

$$f_k = \frac{f_s}{a_k}, \quad (11)$$

thus also obtaining a constant Q-factor:

$$Q_k \overset{ref.}{=} \frac{f_k}{\Delta f_k} = \frac{f_s}{\Delta f} \quad (12)$$

Compared with STFT, CWT tackles the issue brought by the constant TF resolution and the fixed decomposition basis (Table I), bringing the following benefits:

- **Dynamic TF resolution:** CWT owns a constant Q-factor (Eq. 12), which brings a dynamic TF resolution (Fig. 1).
- **More diverse TFRs:** In CWT, the decomposition basis $\psi(n)$ is a variable. Hence, it is possible to decompose a signal with different $\psi(n)$s for more diverse TFRs. Specifically, to model phase information, this study employs the Complex Morlet Wavelet (CMOR) [41] and the Complex Gaussian Wavelet Family (CGAU) [42].
- **Enhanced capability in modeling of short-time transients:** The energy-centralized characteristic of wavelet basis also guarantees CWT a better modeling ability in short-time transients. Fig. 2 compares the decomposition basis in STFT and CWT when modeling a square wave. Wavelet basis can achieve a reconstructed signal with the same number of components but a significantly smaller error in the places where the "step transients" occur due to its "harsher shape," hence showing a better modeling ability in short-time transients.

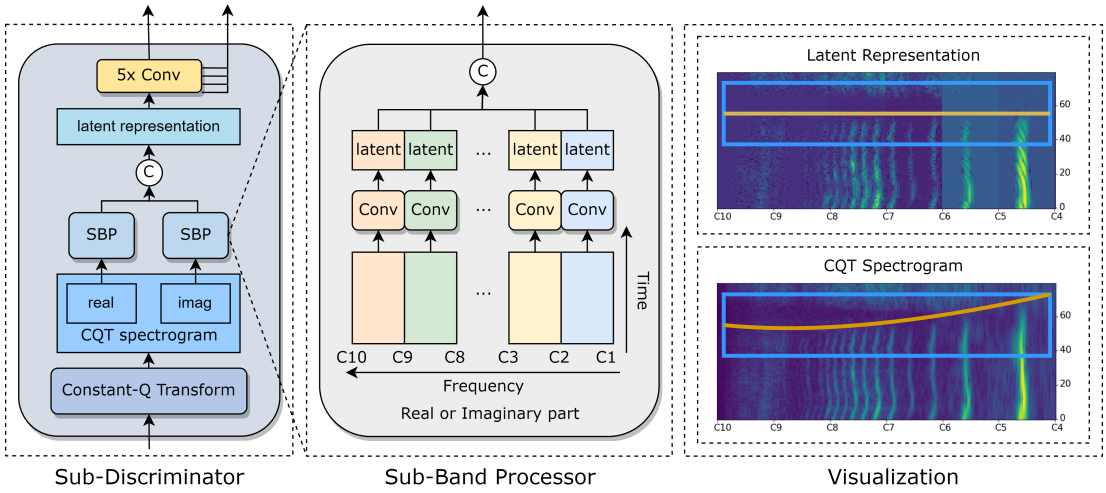


Fig. 3: Architecture of the Sub-Discriminator in MS-SB-CQT Discriminator. Operator "C" denotes for concatenation. SBP means our proposed Sub-Band Processor module. It can be observed that the desynchronized CQT Spectrogram (bottom-right) has been synchronized (upper-right) after SBP.

III. Methodology

As discussed in Section II, compared with STFT, CQT and CWT hold better modeling ability in expressive speech and singing voice. To utilize them, we propose the MS-SB-CQT Discriminator and MS-TC-CWT Discriminator and introduce a joint training strategy to use discriminators based on STFT, CQT, and CWT in the same framework. Specifically, we modify Multi-Resolution Processing for CQT and CWT and introduce Multi-Basis Processing for CWT in Section III-A, propose a Sub-Band Processor for CQT in Section III-B, propose a Temporal Compressor for CWT in Section III-C, and elaborate the integration with the generator in Section III-D.

A. Multi-Scale Processing for CQT and CWT Discriminators

Multi-Scale Processing applies Sub-Discriminators operating on diverse TFRs to reduce the bias brought by the fixed pattern of a specific TF analysis technique, which has been widely used [5, 23–26]. Motivated by the concept of Multi-Scale Processing, we apply Multi-Resolution Processing to both CQT- and CWT-based discriminators and propose Multi-Basis Processing for CWT-based discriminators only.

1) Multi-Resolution Processing: Multi-Resolution Processing utilizes TFRs with different TF resolution distributions for discriminating to alleviate the TF resolution trade-off bias due to the Uncertainty Principle [43].

For the CQT-based discriminator, Given Eq. (7) and (8):

$$\Delta f_k = \frac{f_1 \cdot 2^{\frac{k-1}{B}}}{(2^{\frac{1}{B}} - 1)^{-1}} \tag{13}$$

Thus, the bandwidth Δf_k, which determines the TF resolution trade-off of the k-th frequency bin, depends on the number of bins per octave B. We use different Bs to obtain spectrograms with different TF resolution distributions.

For the CWT-based discriminator, as illustrated in Eq. 10, the bandwidth Δf_k is dependent on the scaling factor a_k. Thus, we utilize different scaling factor series to obtain spectrograms with different TF resolution distributions.

2) Multi-Basis Processing: We propose the Multi-Basis Processing as an extra boost for Multi-Scale Processing on the CWT-base discriminator since CWT has a flexible decomposition basis $\psi(n)$. Specifically, the wavelet basis does not conform to the physical property that the sound comprises a series of sinusoidal components; thus, decomposing signals into such a domain will typically bring biases. To alleviate that, we utilize different wavelet bases to obtain CWT spectrograms in different decomposition domains. Since preliminary experiments verify an independent role between scaling factor a and decomposition basis $\psi(n)$, we implement Multi-Basis Processing in parallel with Multi-Scale Processing for saving up memory, employing the CMOR, the 1-st and 8-st derivatives of CGAU as three distinct complex wavelet bases.

B. Sub-Band Processor for CQT Discriminator

As two sides of a coin, although the dynamic Δf_k brings flexible TF resolution, it also brings the unfixed N_k, which can be obtained from Eq. (2), (4), and (7):

$$N_k = \frac{f_s}{f_k} \cdot (2^{\frac{1}{B}} - 1)^{-1} \tag{14}$$

Thus, in a fixed time step t, the kernels $a_k(t)$ in different frequency bins will convolute different amounts of sample points of the original signal as shown in Eq. 5. This means the convolutional kernels are not temporally synchronized [35] and will cause artifacts in the resulting spectrogram visualized in the bottom right of Fig. 3.

To alleviate this problem, [35] designs a series of temporally synchronized kernels within each octave. This algorithm has also been widely used in toolkits like librosa [44] and nnAudio [45]. However, such an algorithm only makes the $a_k(t)$ of *intra-octave* temporally synchronized but leaves those of *inter-octave* unsolved. Training our neural vocoder using features with such a bias will cause a burden to the adversarial training process and harm the resulting synthesis quality.

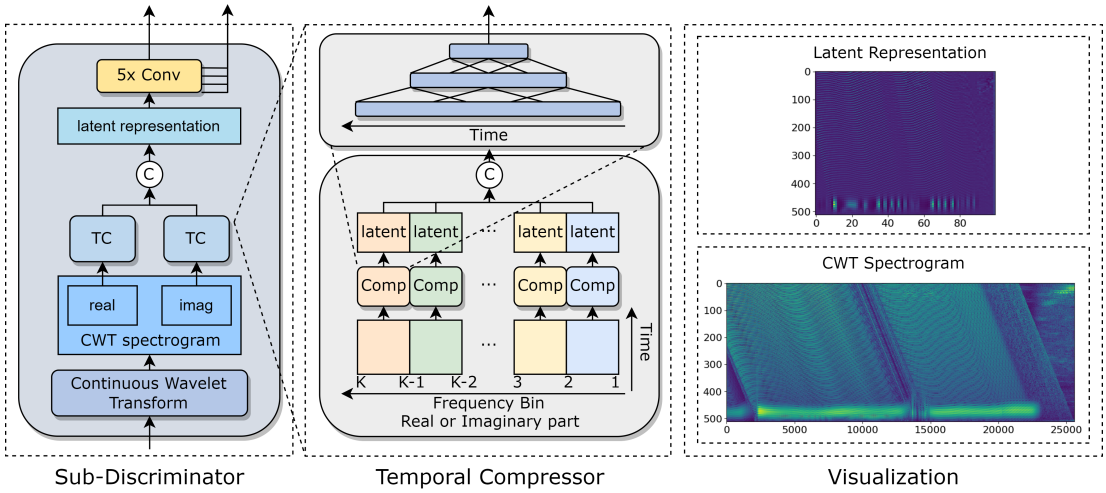


Fig. 4: Architecture of the Sub-Discriminator in MS-TC-CWT Discriminator. Operator "C" denotes for concatenation. TC means our proposed Temporal Compressor module. Comp is a series of temporal-overlapped convolution layers. K is the total number of frequency bins. It can be observed that the CWT Spectrogram (bottom-right) can be compressed while maintaining the overall energy distribution over different frequency bins (upper-right).

Based on that, we utilize the philosophy of representation learning and design the Sub-Band Processor (SBP) module to address this problem further (Fig. 3). In particular, the real or imaginary part of a CQT spectrogram will first be split into sub-bands according to different octaves. Then, each band will be sent to a 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. 3, it can be observed that the SBP module successfully learns the desired representation that is temporally synchronized among all the frequency bins.

C. Temporal Compressor for CWT Discriminator

Although the variable wavelet basis $\psi(n)$ brings better TFR diversity and modeling ability in short-time transients, it also has drawbacks. Specifically, the characteristics of wavelet basis bring the requirement of "unit hop length" [46] for alleviating information loss and maintaining good invertibility. Thus, the resulting CWT spectrogram from an audio signal of shape (1, T) will have a shape of (Scales, T). Such a large spectrogram will be incompatible with deep learning applications with limited GPU memory.

To reduce the memory complexity, instead of directly utilizing the raw CWT spectrogram, we use the learnable compressed latent representation obtained via our designed Temporal Compressor (TC) module (Fig. 4). Specifically, the real or imaginary part of the CWT spectrogram will first be sent scale-wise to a series of Conv2D layers, which use temporal-overlapped convolutional windows to maintain a good continuity between frames, to get its temporal-compressed representation. Then, we concatenate the representations from all scales to obtain the resulting latent representation of the CWT spectrogram. In the upper right of Fig. 4, it can be observed that our proposed TC module successfully compressed the spectrogram while maintaining the overall energy distribution.

D. Joint Training and Integration with Generators

Our proposed discriminators can be easily integrated with any GAN-based vocoders without interfering with the inference stage. We take HiFi-GAN [24] 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 as follows:

$$\mathcal{L}_G = \sum_{m=1}^{M} [\mathcal{L}_{adv}(G; D_m) + 2\mathcal{L}_{fm}(G; D_m)] + 45\mathcal{L}_{mel}, \quad (15)$$

$$\mathcal{L}_D = \sum_{m=1}^{M} [\mathcal{L}_{adv}(D_m; G), \quad (16)$$

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. Suppose we want to integrate a new discriminator D_{new} in the training process, just adding $\mathcal{L}_{adv}(G; D_{\text{new}}) + 2\mathcal{L}_{fm}(G; D_{\text{new}})$ to $\mathcal{L}_G$ and $\mathcal{L}_{adv}(D_{\text{new}}; G)$ to $\mathcal{L}_D$ can finish the integration process.

IV. Experiments

We conduct experiments to explore the following questions:

- **EQ1**: Effectiveness of the proposed MS-SB-CQT and MS-TC-CWT Discriminators.
- **EQ2**: Effectiveness of using MS-SB-CQT, MS-TC-CWT and MS-STFT Discriminators jointly.
- **EQ3**: Generalization of the proposed discriminators to various GAN-based vocoders.
- **EQ4**: Effectiveness of the proposed Sub-Band Processor module and the Multi-Basis Processing technique.
- **EQ5**: What do the latent representations in MS-SB-CQT and MS-TC-CWT Discriminators exactly learn?

TABLE II: Statistics of the datasets used for training and evaluating the vocoders. LibriTTS, LJSpeech, and VCTK are three speech datasets, while the others are singing voice datasets. EN, CN, KR, and JP means English, Chinese, Korean, and Japanese individually.

Dataset	Language	#Hours	#Utterances	#Speakers
LJSpeech [53]	EN	23.9	13,100	1
LibriTTS [52]	EN	585.8	375,086	2,456
VCTK [54]	EN	82.9	33,971	109
Internal Dataset	CN	5.1	3,561	1
M4Singer [47]	CN	29.7	20,896	19
PJS [48]	JP	0.4	291	1
CSD [51]	EN, KR	4.1	2,864	1
OpenSinger [50]	CN	51.9	43,075	74
PopCS [1]	CN	5.9	1,651	1
Opencpop [49]	CN	5.2	3,756	1

A. Experiment Setup

1) Datasets: The experimental datasets for Section IV-C and Section IV-D contain both speech and singing voices. For the singing voice, we adopt M4Singer [47], PJS [48], 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 [49], PopCS [1], OpenSinger [50], and CSD [51] are chosen to evaluate ***unseen*** singers. For the speech, we use the train-clean-100 from LibriTTS [52] and LJSpeech [53]. We randomly sample 2,316 utterances from the two datasets to evaluate ***seen*** speakers and leave the remaining for training (about 75 hours). 3,054 utterances from VCTK [54] are chosen to evaluate ***unseen*** speakers. As for Section IV-E, we adopt Opencpop [49]. We randomly selected 221 utterances for evaluation and the remaining for training (about 5 hours). The detail of each dataset is listed in Table II.

2) Preprocessing: We resampled all the training data to 24kHz, excluding the LibriTTS [52] and OpenSinger [50] datasets, which have a sampling rate of 24kHz originally. Each utterance will first be converted to an STFT matrix with an fft size of 1024, hop length of 256, window length of 1024, fmin of 0, and fmax of 12000, which will then be transformed into a mel-spectrogram with 100 mel-filters. The mel-spectrogram is normalized in log-scale with values ≤ 0.00001 clipped.

3) Training: All the models are trained using the AdamW [55] optimizer with $\beta_1 = 0.8$, $\beta_2 = 0.99$, and a initial learning rate of 0.0002. Exponential decay Scheduler is used with a factor $\gamma = 0.999$. All the experiments are conducted on four NVIDIA RTX4090, V100, or A100 GPUs with the batch size of 16 for around 1,500k steps.

4) Configurations of Generators and Discriminators: We use HiFi-GAN [24], NSF-HiFiGAN [1], BigVGAN [27] and APNet [56] as the experimental generators and Multi-Period Discriminator [24], Multi-Scale Discriminator [24], Multi-Scale STFT Discriminator [5], Multi-Scale Sub-Band CQT Discriminator and Multi-Scale Temporal-Compressed CWT Discriminator as the experimental discriminators. The implementation codes are available in Amphion [57].

The implementation details for the generators are:

- **HiFi-GAN** - The v1 version of the HiFi-GAN [24]. We reimplemented it using [1] with the same hyperparameters.
- **NSF-HiFiGAN** - The integration of NSF and HiFi-GAN. It is one of the SOTA vocoders for singing voice [58]. We reimplemented it using [2] with the same hyperparameters.
- **BigVGAN** - The base version of the BigVGAN [27]. It is one of the SOTA vocoders for speech synthesis. We reimplemented it using [3] with the same hyperparameters.
- **APNet** - The original version of the APNet [56]. It has a fast inference speed with high synthesis quality. We reimplemented it using [4] with the same hyperparameters.

The implementation details for the discriminators are:

- **MSD** - Multi-Scale Discriminator: The discriminator proposed by HiFi-GAN [24]. We re-implemented it using [1] with the same hyperparameters.
- **MPD** - Multi-Period Discriminator: The discriminator proposed by HiFi-GAN [24]. We re-implemented it using [3]. We made modifications to the number of periods ([2, 3, 5, 7, 11, 17, 23, 37]) to make it more effecitve.
- **MS-STFTD** - Multi-Scale STFT Discriminator: The discriminator proposed by Encodec [5]. We re-implemented it using [5] with the same hyperparameters.
- **MS-SB-CQTD** - Multi-Scale Sub-Band CQT Discriminator: One of the proposed discriminators. The CNN in SBP uses a Conv2D with a kernel size of (3, 9) and a channel of 2 covering both the real and imaginary parts. The CNNs in each Sub-Discriminator consist of a Conv2D with kernel size (3, 8) and 32 output channels, three Conv2Ds with dilation rates of [1, 2, 4] in the time dimension, a stride of 2 over the frequency dimension, and a fixed channel of 32, and a Conv2D with kernel size (3, 3), stride (1, 1) and an output channel of 1. LeakyReLU is used as the activation function after each CNN block in the Sub-Discriminator, and Weight Norm [59] is applied to all the CNN blocks. For CQT computation, the global hop length is 256, and the Bs set for three sub-discriminators are [24, 36, 48]. To cover all the frequency bands given the $f_1 = 32.7$, 9 octaves are computed. 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.
- **MS-TC-CWTD** - Multi-Scale Temporal-Compressed CWT Discriminator: One of the proposed discriminators. The CNNs in the TC module have the kernel sizes of [(16, 1), (16, 1), (8, 1)], strides of [(8, 1), (8, 1), (4, 1)], paddings of [(8, 0), (8, 0), (4, 0)] and a channel of 2 covering both the real and the imaginary parts, which is equivalent to a window length of 2048 and a hop length of 256. The CNNs in each Sub-Discriminator are the same as the ones in the MS-SB-CQT Discriminator with the same activation function and weight normalization. Regarding CWT computation, the scale factor a equals linearly interpolated numbers of 1 to [512, 256, 128].

[1] https://github.com/jik876/hifi-gan
[2] https://github.com/MoonInTheRiver/DiffSinger
[3] https://github.com/NVIDIA/BigVGAN
[4] https://github.com/yangai520/APNet
[5] https://github.com/facebookresearch/encodec

TABLE III: Analysis-synthesis results of different discriminators when being integrated into HiFi-GAN [24]. 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", "C" and "W" represent MS-STFT, MS-SB-CQT and MS-TC-CWT Discriminators respectively. The MOS scores are within 95% Confidence Interval (CI).

Domain	System	PESQ (↑)		FPC (↑)		F0RMSE (↓)		Periodicity (↓)		MOS (↑)	
		Seen	Unseen	Seen	Unseen	Seen	Unseen	Seen	Unseen	Seen	Unseen
Singing voice	Ground Truth	4.500	4.500	1.000	1.000	0.000	0.000	0.0000	0.0000	4.84 ± 0.07	4.86 ± 0.07
	HiFi-GAN	2.938	2.863	0.954	0.962	56.502	60.773	0.0675	0.0804	3.30 ± 0.17	3.61 ± 0.15
	HiFi-GAN (+S)	2.954	2.867	0.966	0.968	39.408	47.793	0.0636	*0.0734*	3.44 ± 0.16	3.72 ± 0.16
	HiFi-GAN (+C)	*3.031*	*2.947*	*0.968*	0.971	**36.098**	*43.172*	*0.0620*	0.0735	3.65 ± 0.15	3.83 ± 0.16
	HiFi-GAN (+W)	3.006	**2.967**	0.965	*0.975*	42.096	47.161	0.0644	0.0757	*3.77 ± 0.14*	**3.93 ± 0.15**
	HiFi-GAN (+S+C+W)	**3.040**	*2.947*	**0.972**	**0.977**	*36.137*	**42.022**	**0.0580**	**0.0717**	**3.80 ± 0.16**	*3.91 ± 0.17*
Speech	Ground Truth	4.500	4.500	1.000	1.000	0.000	0.000	0.0000	0.0000	4.83 ± 0.08	4.83 ± 0.09
	HiFi-GAN	3.014	3.141	*0.881*	0.773	184.590	295.428	**0.0062**	0.0103	4.00 ± 0.18	4.07 ± 0.22
	HiFi-GAN (+S)	2.927	3.090	0.866	0.765	195.881	300.368	0.0067	**0.0088**	4.01 ± 0.18	4.10 ± 0.21
	HiFi-GAN (+C)	*3.041*	*3.159*	*0.881*	0.765	*180.673*	305.598	**0.0062**	0.0099	*4.07 ± 0.18*	3.98 ± 0.20
	HiFi-GAN (+W)	2.950	3.099	0.880	*0.784*	187.033	*289.233*	*0.0064*	0.0102	**4.08 ± 0.19**	*4.10 ± 0.19*
	HiFi-GAN (+S+C+W)	**3.102**	**3.257**	**0.882**	**0.799**	**178.665**	**264.935**	0.0068	*0.0095*	4.03 ± 0.20	**4.19 ± 0.17**

B. Evaluation Metrics

1) Objective Evaluation: We investigate objective metrics focusing on spectrogram reconstruction, F0 accuracy, and phase distortion. The details are listed below:

- **PESQ** (Perceptual Evaluation of Speech Quality) [60]: A full-reference algorithm that predicts synthesis quality. We employ the Wide-band raw PESQ score from [6].
- **F0RMSE** (F0 Root Mean Square Error): The Root Mean Square Error (RMSE) of the log-scale F0 (in cent).
- **FPC** (F0 Pearson Correlation Coefficient): The Pearson Correlation Coefficient of F0 trajectories.
- **Periodicity** distortion [61]: The RMSE of the periodicity, which reflects the glitch artifacts that are caused by phase distortion according to previous works [27, 31, 61].

2) Subjective Evaluation: We use the Mean Opinion Score (MOS) and ABX Preference Test for subjective evaluation. In each MOS test, a total of 20 utterances (10 in-distribution utterances and 10 out-of-distribution utterances) will be evaluated. Listeners were asked to give a naturalness score between 1 and 5 with an interval of 0.5 for each utterance synthesized by generators trained with different setups as well as the ground truth audio. In each ABX test, a total of 30 utterances (6 comparative pairs from Section IV-D and 2 comparative pairs from Section IV-E), while each comparative pairs have 6 utterances for evaluating) will be evaluated. Listeners were asked to judge which utterance in each pair had better synthesis quality with the help of the ground truth audio. We invited 20 volunteers who are experienced in the audio generation area to attend the subjective evaluation. Thus, each system in the bellowing MOS test has been graded 200 times, and each pair in the preference test has been graded 120 times. The audio samples are available on our demo page[7].

[6]https://github.com/vBaiCai/python-pesq

[7]https://vocodexelysium.github.io/TFR-Discriminators/

C. Effectiveness of the Proposed Discriminators and Using Them Jointly (EQ1 & EQ2)

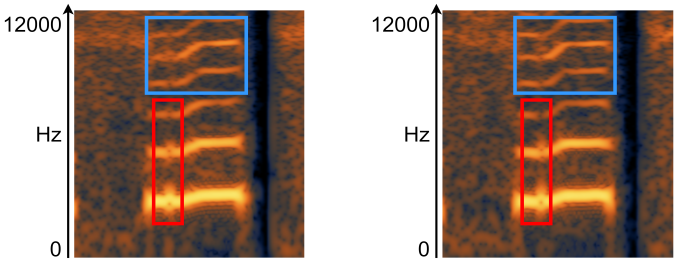
To verify the effectiveness of the proposed discriminators, we take HiFi-GAN with MPD and MSD as the baseline model and enhance it with different discriminators. The results of the analysis-synthesis are illustrated in Table III.

Regarding singing voice, we can observe that: (1) HiFi-GAN (+S), HiFi-GAN (+C), and HiFi-GAN (+W) all outperform HiFi-GAN both subjectively and objectively, confirming the importance of the extra adversarial losses in the frequency domain [32]; (2) Both HiFi-GAN (+C) and HiFi-GAN (+W) outperform the HiFi-GAN (+S) objectively and subjectively, illustrating the effectiveness of utilizing TFRs with dynamic TF resolution; (3) HiFi-GAN (+C) outperforms HiFi-GAN (+W) objectively especially on F0-related metrics, showing the effectiveness of the pitch-aware center frequency distribution. HiFi-GAN (+W) outperforms HiFi-GAN (+C) subjectively, showing the effectiveness of the diverse energy-centered wavelet basis; (4) HiFi-GAN (+S+C+W) outperforms both objectively and subjectively on seen singers while holding better objective results with a similar subjective score on unseen singers, confirming the effectiveness of joint training.

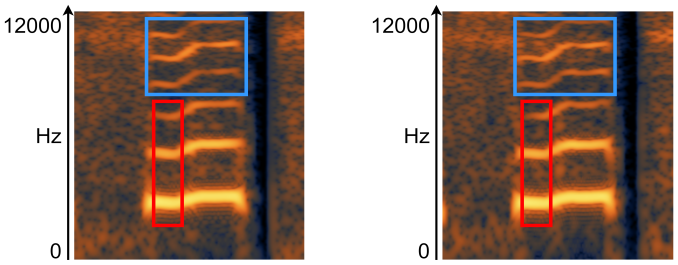
Regarding speech, we can observe that: (1) For seen speakers, HiFi-GAN (+S+C+W) performs best objectively with a similar MOS score with HiFi-GAN (+W) (best) and HiFi-GAN (+C) (second best), illustrating the effectiveness of our proposed methods. (2) For unseen speakers, HiFi-GAN (+S+C+W) performs both objectively and subjectively best, indicating the enhanced generalization ability via utilizing different TFR-based discriminators. (3) Specifically, HiFi-GAN (+C) holds a comparatively higher MOS score in seen speakers but a lower MOS score in unseen speakers. We speculate that although CQT has a dynamic TF resolution which brings a better modeling ability, its generalization ability is degraded due to the incompatibility between the pitch-aware center frequency and speech (non-musical) signal.

TABLE IV: Analysis-synthesis results of our proposed Discriminators when applying on NSF-HiFiGAN [1], BigVGAN [27], and APNet [56] in *singing voice* datasets. The improvements are shown in **bold**. "S", "C" and "W" represent MS-STFT, MS-SB-CQT and MS-TC-CWT Discriminators respectively.

System	PESQ (↑)		FPC (↑)		F0RMSE (↓)		Periodicity (↓)		Preference (↑)	
	Seen	Unseen	Seen	Unseen	Seen	Unseen	Seen	Unseen	Seen	Unseen
Ground Truth	4.500	4.500	1.000	1.000	0.000	0.000	0.0000	0.0000	/	/
NSF-HiFiGAN	3.945	3.876	0.985	0.981	27.197	34.012	0.0377	0.0451	36.84%	39.29%
NSF-HiFiGAN (+S+C+W)	**3.980**	**3.907**	0.981	0.980	**25.816**	**32.100**	**0.0314**	**0.0380**	**63.16%**	**60.71%**
BigVGAN	3.526	3.464	0.982	0.986	22.894	26.338	0.0772	0.0820	15.79%	4.26%
BigVGAN (+S+C+W)	**3.696**	**3.626**	0.982	0.977	28.505	37.075	**0.0387**	**0.0449**	**84.21%**	**95.74%**
APNet	3.254	3.117	0.816	0.832	193.642	191.684	0.0925	0.1086	15.79%	16.07%
APNet (+S+C+W)	3.210	**3.119**	**0.956**	**0.973**	**47.365**	**43.624**	0.0985	**0.1056**	**84.21%**	**83.93%**



(a) HiFi-GAN (+S) (b) HiFi-GAN (+C)



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

Fig. 5: The comparison of mel spectrograms from HiFi-GANs enhanced by different discriminators. "S", "C" and "W" represent MS-STFT, MS-SB-CQT and MS-TC-CWT Discriminators respectively. Integrated with three discriminators, HiFi-GAN could achieve a higher synthesis quality with more accurate harmonic tracking, fundamental frequency reconstruction, and fewer glitches.

To further explore the specific benefits of using the STFT-based, CQT-based, and CWT-based discriminators jointly, we conducted a case study as illustrated in Fig. 5. Notably, in the displayed low-frequency parts, STFT has a better time resolution, while CQT and CWT have a better frequency resolution. It can be observed that: (1) Regarding TF resolution (*upper* rectangles), HiFi-GAN with MS-SB-CQT Discriminator or with MS-TC-CWT Discriminator (Fig. 5b, Fig. 5c) can reconstruct its frequency accurately but "flattened" harmonic component due to the insufficient time resolution, while HiFi-GAN with MS-STFT Discriminator (Fig. 5a) can model the transients in the harmonic component but at the expanse of inaccurate pitch information due to the insufficient frequency resolution; (2) Additionally (bottom rectangles), HiFi-GAN with MS-TC-CWT Discriminator (Fig. 5c) can achieve a glitch-free spectrogram, showing its effectiveness in modeling short-time transients. (3) Integrating those three discriminators combines their strengths and thus achieves a better-reconstructed spectrogram (Fig. 5d), showing the effectiveness of joint training.

D. Effectiveness of Proposed Training Strategy (EQ3)

To verify the generalization ability of the complementary role between the STFT-, CQT-, and CWT-based discriminators, we also conduct experiments under NSF-HiFiGAN, BigVGAN, and APNet, as presented in Table IV.

It is illustrated that: (1) In general, the performance of NSF-HiFiGAN, BigVGAN, and APNet can be improved significantly by jointly training with MS-SB-CQT, MS-TC-CWT, and MS-STFT Discriminators, with both improved objective evaluation scores and subjective preference tests confirming the effectiveness; (2) Regarding NSF-HiFiGAN, although it can synthesis high-fidelity singing voices, it still lacks the modeling abilities of high-frequency band details. Adding adversarial losses based on diverse TFRs alleviates that problem, making synthesized samples closer to the ground truth. Subjective results with a higher preference percentage demonstrate the effectiveness; (3) For BigVGAN, although using the Snake activation function enhanced the generalization ability, it also brings artifacts in phase modeling. Since all MS-STFT, MS-SB-CQT, and MS-TC-CWT Discriminators utilize the phase information, the extra adversarial losses alleviate this problem. Improved subjective scores with significantly higher preference percentages illustrated the effectiveness; (4) As for APNet, although maintaining all the operations in the frame level significantly reduces the memory and computational costs, the difficulty in modeling phase information also brings quality degradation with metallic sound, especially regarding those Aperiodic Parts. Adding MS-STFT, MS-SB-CQT, and MS-TC-CWT Discriminators alleviates that problem, bringing in significantly better FPC and F0-RMSE. We believe this is because the reduction of the metallic sound greatly boosts the performance of the F0-detection algorithm. The subjective evaluation with a higher preference score also confirms its effectiveness. Representative cases regarding these findings can be found on our demo page[7].

TABLE V: Analysis-synthesis results of HiFi-GAN enhanced by discriminators with varied modules and processing techniques. The improvements are shown in **bold**. "C" and "W" represent MS-SB-CQT and MS-TC-CWT Discriminators respectively.

System	PESQ (↑)	FPC (↑)	F0RMSE (↓)	Periodicity (↓)	Preference (↑)
Ground Truth	4.500	1.000	0.000	0.0000	/
HiFi-GAN	2.960	0.972	43.139	0.0611	/
HiFi-GAN (+W)	2.880	**0.978**	**40.338**	0.0705	**54.67%**
w/o Multi-Basis Processing technique	2.898	0.969	42.306	0.0647	45.33%
HiFiGAN (+C)	**2.985**	**0.985**	**29.374**	**0.0634**	**59.46%**
w/o Sub-Band Processor module	2.932	0.963	51.162	0.0612	40.54%

E. Ablation Studies (EQ4)

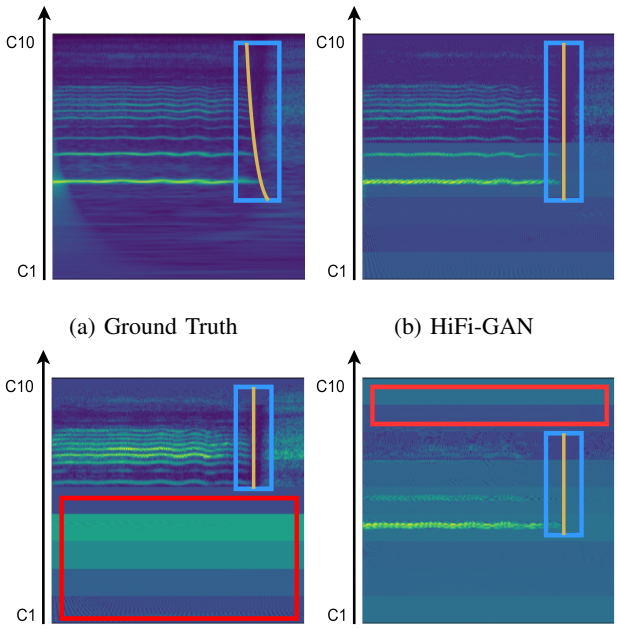
We propose the Sub-Band Processor module to obtain the temporally synchronized CQT latent representations, and the Multi-Basis Processing technique reduces the bias and thus improves the effectiveness further. We conduct an ablation study of those two methods to verify their necessity. The results of the analysis-synthesis are illustrated in Table V.

Regarding the CQT-based discriminators, we can see that our proposed MS-SB-CQT Discriminator significantly outperforms the one without the Sub-Band Processor. We speculate this is because the convolutional kernel in the Conv2D layer cannot handle properly the temporal desynchronization in the *inter-octaves* parts of the CQT spectrogram in the initial stage, which would cause a burden and eventually harm the overall audio quality. Regarding the CWT-based discriminators, we can see that our proposed MS-TC-CWT Discriminator performs better than the one that utilizes only one wavelet (we use the CMOR wavelet in the experiment), showing the effectiveness of utilizing different wavelet bases.

F. Analysis on Learned Representation (EQ5)

We also conducted a case study on the learned representation in the MS-SB-CQT and MS-TC-CWT Discriminators. Regarding the representation in MS-TC-CWT Discriminator, only compression is learned. Regarding the representation in MS-SB-CQT Discriminator, however, apart from the ability to obtain a temporally synchronized spectrogram, it also learns to apply dynamic attention to different frequency bands (Fig. 6).

Notably, among these three vocoders, HiFi-GAN performs averagely across all frequency bands; NSF-HiFiGAN performs better in the low-frequency bands due to its glitch-free ability and accurate pitch modeling but worse in the high-frequency bands due to the aliasing effects; BigVGAN performs better in the high-frequency bands due to its aliasing-free ability but worse in the low-frequency bands due to glitches. It can be observed that: (1) All the SBP modules trained with different generators learned to obtain the temporally synchronized representation (blue rectangle) compared with the ground truth CQT spectrogram; (2) The SBP module trained with NSF-HiFiGAN and BigVGAN learn to mask the low-frequency band and high-frequency band individually(red rectangle). We speculate it is to ignore the frequency band where the generator already has a good synthesis quality and to focus on the frequency bands with a poor reconstruction quality to discriminate whether the audio is generated or not, which indicates the SBP module learns applying dynamic attention to different frequency bands.



(a) Ground Truth (b) HiFi-GAN

(c) NSF-HiFiGAN (d) BigVGAN

Fig. 6: The comparison of the latent representations from the SBP module trained with different generators. All the SBP modules learn to obtain the temporally synchronized representation, while the ones trained with NSF-HiFiGAN and BigVGAN additionally apply dynamic attention to different frequency bands, masking the frequency band where the generator already holds a high reconstruction quality.

V. Conclusion

This study proposed a Multi-Scale Sub-Band Constant-Q Transform (MS-SB-CQT) Discriminator and a Multi-Scale Temporal-Compressed Continuous Wavelet Transform (MS-TC-CWT) Discriminator for GAN-based Vocoders. Experiments conducted on both speech and singing voices confirm the effectiveness of our proposed methods, with strengths conforming to their design ideas. Moreover, with joint training, the proposed two discriminators can also be complementary with the existing MS-STFT Discriminator to improve the neural vocoder further.

REFERENCES

[1] J. Liu, C. Li, Y. Ren, F. Chen, and Z. Zhao, "DiffSinger: Singing Voice Synthesis via Shallow Diffusion Mechanism," in *AAAI*, 2022, pp. 11 020–11 028.

[2] J. Hwang, S. Lee, and S. Lee, "HiddenSinger: High-Quality Singing Voice Synthesis via Neural Audio Codec and Latent Diffusion Models," *CoRR*, vol. abs/2306.06814, 2023.

[3] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Liu, "FastSpeech 2: Fast and High-Quality End-to-End Text to Speech," in *ICLR*, 2021.

[4] K. Shen, Z. Ju, X. Tan, Y. Liu, Y. Leng, L. He, T. Qin, S. Zhao, and J. Bian, "NaturalSpeech 2: Latent Diffusion Models are Natural and Zero-Shot Speech and Singing Synthesizers," *CoRR*, vol. abs/2304.09116, 2023.

[5] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High Fidelity Neural Audio Compression," *arXiv*, vol. abs/2210.13438, 2022.

[6] Z. Du, S. Zhang, K. Hu, and S. Zheng, "FunCodec: A Fundamental, Reproducible and Integrable Open-source Toolkit for Neural Speech Codec," *CoRR*, vol. abs/2309.07405, 2023.

[7] H. Kawahara, "STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds," *AST*, pp. 349–353, 2006.

[8] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans Inf Syst*, vol. 99, no. 7, pp. 1877–1884, 2016.

[9] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," in *SSW*. ISCA, 2016, p. 125.

[10] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. van den Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient Neural Audio Synthesis," in *ICML*, 2018, pp. 2415–2424.

[11] A. Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. Driessche, E. Lockhart, L. Cobo, F. Stimberg *et al.*, "Parallel wavenet: Fast high-fidelity speech synthesis," in *ICML*, 2018, pp. 3918–3926.

[12] W. Ping, K. Peng, and J. Chen, "ClariNet: Parallel Wave Generation in End-to-End Text-to-Speech," in *ICLR*, 2019.

[13] W. Ping, K. Peng, K. Zhao, and Z. Song, "WaveFlow: A Compact Flow-based Model for Raw Audio," in *ICML*, vol. 119, 2020, pp. 7706–7716.

[14] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A Flow-based Generative Network for Speech Synthesis," in *ICASSP*, 2019, pp. 3617–3621.

[15] L. Juvela, B. Bollepalli, V. Tsiaras, and P. Alku, "Glotnet—a raw waveform model for the glottal excitation in statistical parametric speech synthesis," *TASLP*, vol. 27, no. 6, pp. 1019–1030, 2019.

[16] J.-M. Valin and J. Skoglund, "LPCNet: Improving neural speech synthesis through linear prediction," in *ICASSP*, 2019, pp. 5891–5895.

[17] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "DiffWave: A Versatile Diffusion Model for Audio Synthesis," in *ICLR*, 2021.

[18] T. D. Nguyen, J.-H. Kim, Y. Jang, J. Kim, and J. S. Chung, "FreGrad: Lightweight and Fast Frequency-aware Diffusion Vocoder," *arXiv:2401.10032*, 2024.

[19] X. Wang, S. Takaki, and J. Yamagishi, "Neural Source-filter-based Waveform Model for Statistical Parametric Speech Synthesis," in *ICASSP*, 2019, pp. 5916–5920.

[20] C. Yu and G. Fazekas, "Singing Voice Synthesis Using Differentiable LPC and Glottal-Flow-Inspired Wavetables," in *ISMIR*, 2023, pp. 667–675.

[21] R. Yamamoto, E. Song, and J. Kim, "Parallel Wavegan: A Fast Waveform Generation Model Based on Generative Adversarial Networks with Multi-Resolution Spectrogram," in *ICASSP*, 2020, pp. 6199–6203.

[22] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. C. Courville, "MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis," in *NeurIPS*, 2019.

[23] W. Jang, D. Lim, and J. Yoon, "Universal MelGAN: A Robust Neural Vocoder for High-Fidelity Waveform Generation in Multiple Domains," *arXiv*, vol. abs/2011.09631, 2020.

[24] J. Su, Z. Jin, and A. Finkelstein, "HiFi-GAN: High-Fidelity Denoising and Dereverberation Based on Speech Deep Features in Adversarial Networks," in *INTERSPEECH*, 2020, pp. 4506–4510.

[25] J. Kim, S. Lee, J. Lee, and S. Lee, "Fre-GAN: Adversarial Frequency-Consistent Audio Synthesis," in *INTERSPEECH*, 2021, pp. 2197–2201.

[26] R. Huang, C. Cui, F. Chen, Y. Ren, J. Liu, Z. Zhao, B. Huai, and Z. Wang, "SingGAN: Generative Adversarial Network For High-Fidelity Singing Voice Generation," in *ACM MM*, 2022, pp. 2525–2535.

[27] S. Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S. Yoon, "BigVGAN: A Universal Neural Vocoder with Large-Scale Training," in *ICLR*, 2023.

[28] T. Shibuya, Y. Takida, and Y. Mitsufuji, "BigVSAN: Enhancing GAN-based Neural Vocoders with Slicing Adversarial Network," *CoRR*, vol. abs/2309.02836, 2023.

[29] D. Wu, W. Hsiao, F. Yang, O. Friedman, W. Jackson, S. Bruzenak, Y. Liu, and Y. Yang, "DDSP-based Singing Vocoders: A New Subtractive-based Synthesizer and A Comprehensive Evaluation," in *ISMIR*, 2022, pp. 76–83.

[30] Z. Liu, T. Hartwig, and M. Ueda, "Neural networks fail to learn periodic functions and how to fix it," *NeurIPS*, vol. 33, pp. 1583–1594, 2020.

[31] S. Li, S. Liu, L. Zhang, X. Li, Y. Bian, C. Weng, Z. Wu, and H. Meng, "SnakeGAN: A Universal Vocoder Leveraging DDSP Prior Knowledge and Periodic Inductive Bias," in *ICME*, 2023, pp. 1703–1708.

[32] J. You, D. Kim, G. Nam, G. Hwang, and G. Chae, "GAN Vocoder: Multi-Resolution Discriminator Is All You Need," in *INTERSPEECH*, 2021, pp. 2177–2181.

[33] J. Allen, "Short term spectral analysis, synthesis, and modification by discrete Fourier transform," *TASLP*, vol. 25, no. 3, pp. 235–238, 1977.

[34] J. C. Brown and M. Puckette, "An efficient algorithm for the calculation of a constant Q transform," *JASA*, vol. 92, pp. 2698–2701, 1992.

[35] C. Schörkhuber and A. Klapuri, "Constant-Q transform toolbox for music processing," in *SMCC*, 2010, pp. 3–64.

[36] O. Rioul and P. Duhamel, "Fast algorithms for discrete and continuous wavelet transforms," *IEEE Trans. Inf. Theory*, vol. 38, no. 2, pp. 569–586, 1992.

[37] Y. Gu, X. Zhang, L. Xue, and Z. Wu, "Multi-scale sub-band constant-q transform discriminator for high-fidelity vocoder," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 10 616–10 620.

[38] E. Hewitt and R. Hewitt, "The Gibbs-Wilbraham phenomenon: An episode in fourier analysis," *Arch. Hist. Exact Sci. 21, 129–160*, 1979.

[39] S. Seneff, "Pitch and spectral analysis of speech based on an auditory synchrony model," Ph.D. dissertation, Massachusetts Institute of Technology, Research Laboratory of Electronics, 1985.

[40] J. P. Stautner, "Analysis and synthesis of music using the auditory transform," Ph.D. dissertation, Massachusetts Institute of Technology, 1983.

[41] M. Taner, "Joint time/frequency analysis, Q quality factor and dispersion computation using Gabor-Morlet wavelets or the Gabor-Morlet transform," *RSI*, pp. 1–5, 1983.

[42] R. Navarro and A. Tabernero, "Gaussian wavelet transform: two alternative fast implementations for images," *MSSP*, vol. 2, pp. 421–436, 1991.

[43] G. B. Folland and A. Sitaram, "The uncertainty principle: a mathematical survey," *JFAA*, vol. 3, pp. 207–238, 1997.

[44] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and Music Signal Analysis in Python," in *SciPy*, 2015.

[45] K. W. Cheuk, H. Anderson, K. Agres, and D. Herremans, "nnAudio: An on-the-Fly GPU Audio to Spectrogram Conversion Toolbox Using 1D Convolutional Neural Networks," *IEEE Access*, pp. 161 981–162 003, 2020.

[46] G. R. Lee, R. Gommers, F. Waselewski, K. Wohlfahrt, and A. OLeary, "PyWavelets: A Python package for wavelet analysis," *JOSS*, vol. 4, no. 36, p. 1237, 2019.

[47] L. Zhang, R. Li, S. Wang, L. Deng, J. Liu, Y. Ren, J. He, R. Huang, J. Zhu, X. Chen, and Z. Zhao, "M4Singer: A Multi-Style, Multi-Singer and Musical Score Provided Mandarin Singing Corpus," in *NeurIPS*, 2022.

[48] J. Koguchi, S. Takamichi, and M. Morise, "PJS: phoneme-balanced Japanese singing-voice corpus," in *APSIPA*, 2020, pp. 487–491.

[49] Y. Wang, X. Wang, P. Zhu, J. Wu, H. Li, H. Xue, Y. Zhang, L. Xie, and M. Bi, "Opencpop: A High-Quality Open Source Chinese Popular Song Corpus for Singing Voice Synthesis," in *INTERSPEECH*, 2022.

[50] R. Huang, F. Chen, Y. Ren, J. Liu, C. Cui, and Z. Zhao, "Multi-Singer: Fast Multi-Singer Singing Voice Vocoder With A Large-Scale Corpus," in *ACM MM*, 2021.

[51] S. Choi, W. Kim, S. Park, S. Yong, and J. Nam, "Children's song dataset for singing voice research," in *ISMIR*, 2020.

[52] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech," in *INTERSPEECH*, 2019, pp. 1526–1530.

[53] K. Ito and L. Johnson, "The LJ Speech Dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[54] J. Yamagishi, C. Veaux, K. MacDonald *et al.*, "CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," *CSTR*, 2019.

[55] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *ICLR*, 2019.

[56] Y. Ai and Z. Ling, "APNet: An All-Frame-Level Neural Vocoder Incorporating Direct Prediction of Amplitude and Phase Spectra," *TASLP*, pp. 2145–2157, 2023.

[57] X. Zhang, L. Xue, Y. Wang, Y. Gu, X. Chen, Z. Fang, H. Chen, L. Zou, C. Wang, J. Han, K. Chen, H. Li, and Z. Wu, "Amphion: An Open-Source Audio, Music and Speech Generation Toolkit," *arXiv*, vol. abs/2312.09911, 2023.

[58] W.-C. Huang, L. P. Violeta, S. Liu, J. Shi, Y. Yasuda, and T. Toda, "The Singing Voice Conversion Challenge 2023," *arXiv*, vol. abs/2306.14422, 2023.

[59] T. Salimans and D. P. Kingma, "Weight Normalization: A Simple Reparameterization to Accelerate Training of Deep Neural Networks," in *NeurIPS*, 2016, p. 901.

[60] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *ICASSP*, 2001, pp. 749–752.

[61] M. Morrison, R. Kumar, K. Kumar, P. Seetharaman, A. C. Courville, and Y. Bengio, "Chunked Autoregressive GAN for Conditional Waveform Synthesis," in *ICLR*, 2022.