修 士 論 文 概 要 書

Master's Thesis Summary

Date of submission: 07/22/2024 (MM/DD/YYYY)

専攻名 (専門分野) Department	Computer Science and Communications Engineering	氏 名 Name	Yenan Zhang	指 導 教 員 Advisor	Hiroshi Watanabe 印 Seal
研究指導名 Research guidance	Research on Audiovisual Information Processing	学籍番号 Student ID number	CD 5122FG28-7		
研究題目 Title	Music Super-Resolution Using Deep Neural Networks				

1. Introduction

Audio Super-Resolution (SR) is an important research topic because low-resolution recordings are ubiquitous in daily life. In this paper, we explore the music SR task through solo piano music, which is challenging due to the wide frequency response and dynamic range of music. In recent years, with the development of deep learning, audio SR methods based on it have become mainstream, but there are few SR methods focused on the field of music. Hence, it remains several challenges in the field of music SR, which require thorough investigation.

In this thesis, we propose two methods: Time-Domain Phase Repair (TD-PR) and BigWavGAN, for exploring different challenges. In the discussion of TD-PR, we thorough investigate the common annoying artifacts in Time-Domain Convolution Neural Networks (TD-CNNs) and identify the cause of the annoying artifacts via a subjective experiment. We further propose TD-PR, which uses a neural vocoder pre-trained on the wide-band data to repair the phase components in the waveform outputs of TD-CNNs. In the discussion of BigWavGAN, we propose BigWavGAN, which incorporates Demucs, a large-scale wave-to-wave model, with the state-of-the-art discriminators and adversarial training strategies to unleash the potential of large Deep Neural Network (DNN) models in music SR and achieve the optimal perceptual quality.

2. Related Work

2.1 Time-Domain Convolutional Neural Network Approaches

Various works have delved into the deep learning based approaches for audio SR. Some of them work in frequency domain. Frequency-domain approaches aim to directly recover the high-resolution components in the magnitude spectrogram, and generally require additional signal processing to estimate the corresponding phase information, such as Griffin-Lim algorithms [1] or neural vocoders [2]. Compared with frequency-domain approaches, TD-CNNs that directly learn a wave-to-wave mapping, are considered being able to avoid the phase problem in audio SR due to the direct waveform processing. AudioUNet is one of the pioneers of tackling audio SR by a TD-CNN [3]. Tagliasacchi et al. proposed SEANet [4], a GAN-based model for speech SR, of which the generator is a light-weight but effective TD-CNN. Defossez et al. proposed a TD-CNN model referred to as Demucs, which is a large model with over 130M parameters and is initially designed to address music source separation [5]. Considering the fact that Demucs has shown strong performance in tasks besides source separation [6], we utilize the Demucs model in the SR task as one of the TD-CNN baselines in this thesis. Although lots of efforts have been made to improve perceptual quality of TD-CNNs, none of them succeeds in removing the artifacts according to their open-available audio samples. Therefore, the cause of the annoying artifacts which TD-CNNs tend to produce in their waveform output, is yet to be identified.

2.2 Generative Adversarial Network Approaches in Audio Super-Resolution

Recent publications have delved into Generative Adversarial Network (GAN) based models in audio SR. Compared to models trained with standard mean square error losses, GAN-based models exhibit a superior capability to generate results with better perceptual quality. BEHMGAN is the state-of-the-art of GAN-based music SR model [7]. Notably, a state-of-the-art neural vocoder referred to as BigVGAN, which characterized by a large-size generator with an unprecedented scale of up to 112M parameters, is proposed by Lee et al. [8]. BigVGAN can synthesize high-fidelity audio and shows its superior zero-shot performance across various out-of-distribution scenarios. However, in the task of audio SR, there is no wave-to-wave GAN-based model in such a large model size. This inspired us to explore the large-scale wave-to-wave GAN model in music SR with high performance and superior generalization ability.

3. Methodology

3.1 TD-PR: Time-Domain Phase Repair

In order to alleviate the artifacts caused by distorted phase components, we propose Time-Domain Phase Repair (TD-PR). The TD-PR framework consists of two separately pretrained DNN modules and a phase replacement operation.

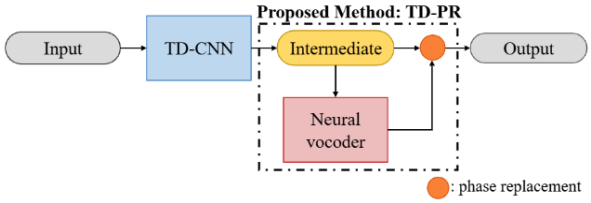

Fig. 1. Overview of the proposed method TD-PR.

The overview of the proposed method is shown in Fig. 1. Specifically, TD-CNN is trained to perform super-resolution for various narrow-band inputs. The neural vocoder takes only the magnitude of the TD-CNN's output as input, and re-synthesizes another waveform that contains repaired phase components. Then, the distorted phase components in TD-CNN's output are replaced by that from the vocoder.

3.2 BigWavGAN

The overview of BigWavGAN's architecture is shown in Fig. 2. The generator of BigWavGAN has the identical architecture with Demcus from [5]. It is a wave domain U-net model leveraging a Long Short-Term Memory (LSTM) recurrent neural network layer as the bottleneck. BigWavGAN benefits from the two types of discriminators: MSD and MRD, which works in the time domain and frequency domain separately.

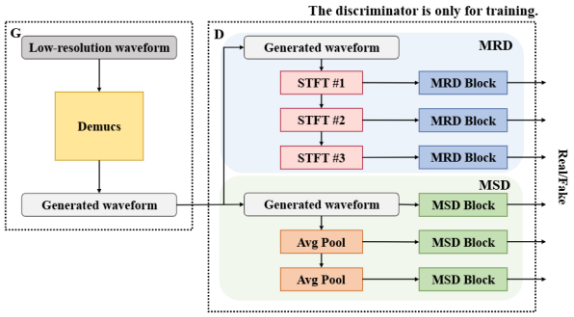

Fig. 2. Overview of architecture the proposed method BigWavGAN.

4. Experiment

We trained our model in this thesis by using the MAESTRO dataset [9]. It is composed of about 200 hours of high-quality classical piano recordings in waveform. Although these recordings have the sampling rate of 44.1 kHz or 48 kHz, we empirically found that 16 kHz is high enough for the piano solo. Hence, we performed music SR with the target bandwidth of 8 kHz. We used the official split of the MAESTRO dataset for training, validation and test. We cut all of the waveform into 30-second short clips for efficient training.

5. Evaluation

5.1 Evaluation on TD-PR

In addition to objective evaluations, we conducted the listening tests by collecting Mean Opinion Score (MOS) for subjective evaluations. The box plot of the MOS test results and the corresponding average for each method are shown in Fig. 3.

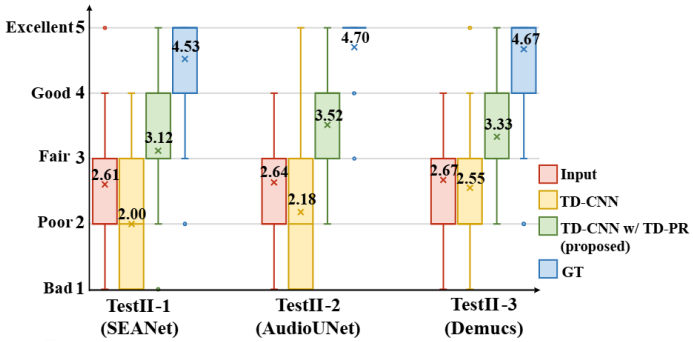

Fig. 3. Results of MOS listening test: The box plot of the MOS scores across input, TD-CNN, TD-PR and GT (Ground Truth). TD-PR is applied to 3 different TD-CNN baselines.

TD-PR obtained better MOS scores than all three TD-CNN baselines by a large margin, which indicate that TD-PR significantly improved the perceptual quality of TD-CNN baselines. Since the proposed TD-PR only repairs the phase components of the waveforms, the improved perceptual quality in turn indicates that phase distortion has been the cause of the annoying artifacts in TD-CNNs.

5.2 Evaluation on BigWavGAN

Besides Objective evaluations, we conducted a set of subjective evaluations to identify the advantage of the proposed BigWavGAN. The subjective evaluations are in the style of A/B test, rather than MOS collection, because A/B test can better measure tiny differences between two models. Since A/B test cannot handle multiple models at once, we conduct multiple A/B tests (e.g., BigWavGAN vs Demucs, BigWavGAN vs BEHMGAN) for a more comprehensive analysis.

The results of subjective evaluations are illustrated in Fig. 4. In all thress datasets, BigWavGAN significantly improved Demucs in terms of perceptual quality by a large margin. This also reveals that BigWavGAN achieved superior generalization to out-of-distribution data. Similar advantages of BigWavGAN are observed when it is compared with BEHMGAN, the state-of-the-art music SR model.

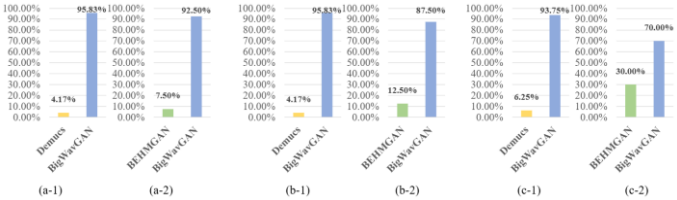

Fig. 4. Results of A/B listening tests: (a) is tested on MAESTRO; (b) is tested on MusicNet; (c) is tested on denoised real historical recordings.

6. Conclusion

In this thesis, we explore music Super-Resolution (SR) through solo piano music. First, we investigated into Time-Domain Convolutional Neural Networks (TD-CNNs), trying to identify the cause of the annoying artifacts and improve TD-CNNs' perceptual quality by alleviating the artifacts. To the best of our knowledge, this work is the first to demonstrate the artifacts in TD-CNNs are caused by the phase distortion via a subjective experiment. We further propose Time-Domain Phase Repair (TD-PR), which significantly improves the perceptual quality of TD-CNN baselines. Since the proposed TD-PR only repairs the phase components of waveform, the improved perceptual quality in turn indicates that phase distortion has been the cause of the annoying artifacts of TD-CNNs.

Then, based on the discussion of our first proposed method TD-PR, we proposed a large-scale wave-to-wave model referred to as BigWavGAN for music SR. The results of a set of subjective evaluations demonstrate that BigWavGAN can achieve significantly better perceptual quality compared to the baseline model Demucs. Notably, BigWavGAN surpasses the state-of-the-art music SR model in both simulated and real-world scenarios. Moreover, BigWavGAN achieves superior generalization ability to address out-of-distribution data including real historical recordings. Therefore, BigWavGAN successfully unleashes the potential of the large-scale Demucs in music SR.

References

[1] Shichao Hu, et al. Phase-aware music super-resolution using generative adversarial networks. arXiv preprint arXiv:2010.04506, 2020.

[2] Haohe Liu, et al. Neural vocoder is all you need for speech super-resolution. arXiv preprint arXiv:2203.14941, 2022.

[3] Volodymyr Kuleshov, et al. Audio super resolution using neural networks. arXiv preprint arXiv:1708.00853, 2017.

[4] Marco Tagliasacchi, et al. Seanet: A multi-modal speech enhancement network. arXiv preprint arXiv:2009.02095, 2020.

[5] Alexandre, Défossez, et al. Demucs: Deep extractor for music sources with extra unlabeled data remixed. arXiv preprint arXiv:1909.01174, 2019.

[6] Jiaqi Su, et al. Bandwidth extension is all you need. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 696–700. IEEE, 2021.

[7] Eloi Moliner, et al. Behm-gan: Bandwidth extension of historical music using generative adversarial networks. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31:943–956, 2022.

[8] Sang-gil Lee, et al. Bigvgan: A universal neural vocoder with large-scale training. arXiv preprint arXiv:2206.04658, 2022.

[9] Curtis Hawthorne, et al. Enabling factorized piano music modeling and generation with the maestro dataset. arXiv preprint arXiv:1810.12247, 2018.

Music Super-Resolution Using Deep Neural Networks

A Thesis Submitted to the Department of Computer Science and Communications Engineering, the Graduate School of Fundamental Science and Engineering of Waseda University in Partial Fulfillment of the Requirements for the Degree of Master of Engineering

Submission Date: July 22nd, 2024

Yenan Zhang

(5122FG28-7)

Advisor: Prof. Hiroshi Watanabe

Research guidance: Research on Audiovisual Information Processing

Acknowledgements

Firstly, I would like to express my deepest gratitude to my advisor, Professor Hiroshi Watanabe, for providing his invaluable guidance, patience, encouragement and a nurturing and motivating environment throughout my master's studies, making my journey both enriching and enjoyable. Besides academic guidance, he has always offered me steadfast support, genuine care and concern in daily life, boosting my confidence to pursue my goals.

Then, I am profoundly grateful to all the members of Advanced Multimedia Systems Laboratory for their helpful discussions, encouragements, and friendships. Through frequent discussions about each other's research progress and the latest technological developments, my knowledge extended beyond a single field. These exchanges have significantly broadened my understanding and sparked my curiosity in various areas, greatly enriching my academic journey. I would also like to extend my heartfelt thanks to all members in the lab for their enthusiastic participation in the subjective listening experiments we conducted. This thesis would not have been possible without their active involvement and support.

Finally, I would like to thank my family and friends for their unconditional love and support, which have given me the courage to pursue the life I desired. When I was stressful or in a low mood, their accompany and encouragement helped me through the toughest moments. They has made my master's journey even more vibrant and fulfilling.

Abstract

Audio Super-Resolution (SR), involving the transformation of low-resolution (*i.e.*, narrow-band) input into high-resolution (*i.e.*, wide-band) audio, which gives the low-resolution audio more detail and brighter tone, is a vital research topic as low-resolution recordings are ubiquitous in daily life. In this thesis, we explore the music SR task through solo piano music, which is challenging due to the wide frequency response and dynamic range of music. Many SR models exploit Time-Domain Convolutional Neural Networks (TD-CNNs), which benefit from the joint processing of magnitude and phase information of audio signals. However, prior works indicate that TD-CNN approaches tend to produce annoying artifacts, and the cause of the artifacts is yet to be identified. To this end, we demonstrate that the artifacts in TD-CNNs are caused by the phase distortion via a subjective experiment for the first time. We further propose Time-Domain Phase Repair (TD-PR), which uses a neural vocoder pretrained on the wide-band data to repair the phase components in the waveform outputs of TD-CNNs. The proposed TD-PR obtained better mean opinion score than TD-CNN baselines, which demonstrates TD-PR significantly improves the perceptual quality of TD-CNNs. Since the proposed TD-PR only repairs the phase components of the waveforms, the improved perceptual quality in turn indicates that phase distortion has been the cause of the annoying artifacts of TD-CNNs. Moreover, the proposed TD-PR can be easy applied to arbitrary TD-CNNs without additional adaptation. Audio samples of TD-PR are available on the demo page[1].

Based on the analysis of TD-PR, we further explore music SR, aiming to further improve the performance of TD-PR to achieve the optimal perceptual quality. Gen-

[1] https://mannmaruko.github.io/demopage/tdpr.html

erally, Deep Neural Networks (DNNs) are expected to have high performance when their model size is large. However, large models failed to produce high-quality results commensurate with their scale in the music SR task in the experiments of the proposed method TD-PR. We attribute this to that DNNs cannot learn information commensurate with their size from standard mean square error losses. To unleash the potential of large DNN models in music SR, we propose BigWavGAN, which incorporates Demucs, a large-scale wave-to-wave model, with the state-of-the-art discriminators and adversarial training strategies. Our discriminator consists of Multi-Scale Discriminator (MSD) and Multi-Resolution Discriminator (MRD). During inference, since only the generator is utilized, there are no additional parameters or computational resources required compared to the baseline model Demucs. Objective evaluations affirm the effectiveness of BigWavGAN in music SR. Subjective evaluations indicate that BigWavGAN can generate music with significantly high perceptual quality over the baseline model. Notably, BigWavGAN surpasses the state-of-the-art music SR model in both simulated and real-world scenarios. In addition, BigWavGAN represents its superior generalization ability to address out-of-distribution data. The conducted ablation study reveals the importance of our discriminators and training strategies. Samples of BigWavGAN are available on the demo page [2].

Keywords: Music super-resolution, Audio super-resolution, Convolutional neural network, Large-scale wave-to-wave model, Generative adversarial network, Music information retrieval

[2]https://mannmaruko.github.io/demopage/BigWavGAN/d.html

Contents

Acknowledgements **ii**

Abstract **iv**

List of Figures **vii**

List of Tables **viii**

1 Introduction **1**

1.1 Research Background 1

1.2 Research Objectives 2

1.3 Thesis Outline 4

2 Phase Repair for Time-Domain Convolutional Neural Networks in Music Super-Resolution **6**

2.1 Time-Domain and Frequency-Domain Approaches for Super-Resolution 6

2.1.1 Frequency-Domain Approaches 6

2.1.2 Time-Domain Convolutional Neural Network Approaches . 7

2.2 Proposed Method: TD-PR 8

2.2.1 TD-PR: Time-Domain Phase Repair 8

2.2.2 Simulation Pipeline 10

2.2.3 Loss Function 10

2.3 Experiments for TD-PR 12
2.3.1 Dataset and Implementation 12
2.3.2 Investigation into Effectiveness of Ground Truth Phase Components 13
2.3.3 Comparison Between TD-PR and TD-CNN Baselines . . . 14
2.4 Evaluations Results and Discussion 15
2.4.1 Impact of Ground Truth Phase Components 15
2.4.2 Subjective and Objective Evaluations Results on TD-PR . . 15
2.4.3 Qualitative Evaluation of TD-PR 17

3 BigWavGAN: A Wave-To-Wave Generative Adversarial Network for Music Super-Resolution 19

3.1 Existing Challenges in Music Super-Resolution 19
3.1.1 Large-scale Time-Domain Neural Network Fails to Achieve Optimal Performance Commensurate with Its Model Size . 19
3.1.2 Generative Adversarial Network Approaches in Audio Super-Resolution 20
3.2 Proposed Method: BigWavGAN 21
3.2.1 Architecture of BigWavGAN 21
3.2.2 Training Objectives 23
3.3 Experiments for BigWavGAN 24
3.4 Evaluations on BigWavGAN 25
3.4.1 Objective Evaluations 25
3.4.2 Subjective Evaluations 26

4 Conclusion 29

5 Future Work 31

5.1 Explore Music Super-Resolution for Various Types of Music . . . 31

5.2 Explore Automatic Sound Quality Assessment 31

6 List of Publications **33**

Bibliography **33**

List of Figures

2.1 Overview of the proposed TD-PR: The TD-CNN is trained to perform super-resolution for various narrow-band inputs. The neural vocoder takes only the magnitude of the TD-CNN's output as input, and re-synthesizes another waveform that contains repaired phase components. Then, the distorted phase components in TD-CNN's output are replaced by that from the vocoder. 9

2.2 Results of the preliminary AB listening test: 95.38% of the TD-CNN w/ GT-phase is voted to have fewer artifacts. 13

2.3 Results of MOS listening test: The box plot of the ratings across input, TD-CNN, TD-PR and GT. TD-PR is applied to three different TD-CNN baselines. 15

2.4 Visualization of a set of phase spectrograms: (a) low-resolution input; (b) ground truth; (c-1) SEANet; (c-2) SEANet w/ TD-PR (proposed); (d-1) AudioUNet; (d-2) AudioUNet w/ TD-PR (proposed); (e-2) Demcus; (e-2) Demucs w/ TD-PR (proposed). 17

2.5 Visualization of a set of magnitude spectrograms: (a) low-resolution input; (b) ground truth; (c-1) SEANet; (c-2) SEANet w/ TD-PR (proposed); (d-1) AudioUNet: (d-2) AudioUNet w/ TD-PR (proposed); (e-2) Demcus; (e-2) Demucs w/ TD-PR (proposed). 18

3.1 Overview of the architecture of BigWavGAN. 22

3.2 Results of A/B listening tests: (a) is tested on MAESTRO; (b) is tested on MusicNet; (c) is tested on denoised real historical recordings. 27

List of Tables

2.1 LSD results with different input bandwidth and parameter amount. . 16

3.1 LSD scores on the MAESTRO dataset. The bold represents the top two LSD scores. 25

Chapter 1

Introduction

1.1 Research Background

Audio Super-Resolution (SR), also known as bandwidth extension and bandwidth expansion, aims to predict the high-resolution components from the low-resolution input audio to give the low-resolution input more detail and brighter tone. Audio SR is an important research topic as low-resolution audio is common in daily life, *e.g.*, historical recordings or unprofessional-made modern recordings. In recent years, deep learning based methods have become the mainstream in audio SR [1, 2, 3, 4, 5], but only few works focus on the field of music [2, 5].

Music SR plays a crucial role in the field of audio restoration, for providing high-fidelity listening experience of music production, enhancing streaming services, restoring historical recordings. It aims to deliver superior sound quality that brings out the full richness and detail of the music. Among existing prior works, several challenges in music SR remain and require thorough discussion. In this thesis, we explore the music SR task through solo piano music, which is challenging due to the wide frequency response and dynamic range of music. Among the diverse applications of music SR, restoring historical music recordings stands out as one of the most significant tasks. Historical music recordings are invaluable as

cultural heritage for representing the artistic legacy of the golden age. A vast number of historical music recordings are preserved in archives, allowing contemporary audiences to experience music as it was originally performed. They also provide a window into the evolution of musical genres, styles, and performance practices over time. However, historical music recordings suffer from severe and multiple degradation due to the technological limitations of the era, such as multiple kinds of surface noises, distortion, and a narrow frequency bandwidth [5, 6]. Therefore, one of our goals in this thesis is the bandwidth extension of band-limited signals of historical music recordings. Meanwhile, performing music SR in real world is also one of our concerns, which is challenging due to a variety of bandwidths of real-world low-resolution recordings.

1.2 Research Objectives

Although deep learning methods for audio SR has been received increasing attention in recent years, only few works focus on music SR, which remains several challenges in the field of music SR requiring thorough investigation.

First, we aim to develop music SR models that are capable of handling real-world applications, which is challenging due to a variety of bandwidths of real-world low-resolution recordings. In terms of up-sampling ratio, various models are developed to perform audio SR on a fixed ratio (*e.g.*, $2\times$) [1, 2], which would be a limitation when apply these models to real world scenarios. To this end, we develop our music SR models that handle input of various narrow-bandwidths within a certain range. Note that our models can handle arbitrary narrow bandwidth within the range.

Next, among previous works in audio SR, many models are developed in time domain to jointly process magnitude and phase of audio signals. However, prior works indicate that approaches using Time-Domain Convolutional Neural Network (TD-CNN) tend to produce annoying artifacts in their waveform outputs. And the

cause of the artifacts is yet to be identified. To this end, we investigate the artifacts of TD-CNNs in the following ways:

- First, we train three TD-CNN models to handle low-resolution music with various bandwidth, which is applicable to real world problems. The SR capability of three TD-CNN models as well as the artifacts are successfully reproduced.
- Second, we conduct an AB listening test to demonstrate the artifacts in TD-CNNs are caused by the phase distortion via a subjective experiment. To the best of our knowledge, this is the first to demonstrate this problem via a subjective experiment.
- Last but not least, we propose a method referred to as Time-Domain Phase Repair (TD-PR), which utilizes a vocoder pretrained on wide-band music signals to repair the distorted phase components in the waveform output of the TD-CNN. Since the vocoder and TD-CNNs are trained independently, a single pretrained vocoder can be directly applied to arbitrary TD-CNNs without additional adaptation. Therefore, we apply TD-PR to the aforementioned three TD-CNNs. The proposed TD-PR consistently and significantly improved the perceptual quality of all three TD-CNN baselines. Since TD-PR only repair the phase components of waveform, the improved perceptual quality in turn indicates that phase distortion has been the cause of the annoying artifacts of TD-CNNs.

Moreover, based on the investigation of TD-PR, we further explore music SR, aiming to further improve the performance of TD-PR to achieve the optimal perceptual quality. Deep Neural Networks (DNNs) are anticipated to achieve high performance when their model size is large. However, large models failed to produce high-quality results commensurate with their scale in music SR according to our investigation of TD-PR. We attribute this to that models cannot learn information (e.g., correct phase information) commensurate with their size through standard

Mean Square Error (MSE) losses. To unleash the potential of large DNN models in music SR, we propose BigWavGAN, which incorporates Demucs, a large-scale wave-to-wave model containing 134M parameters, with a state-of-the-art discriminators and adversarial training strategies. Specifically, the combination of Multi-Scale Discriminator (MSD) and Multi-Resolution Discriminator (MRD) constitutes the discriminator of BigWavGAN. During inference, only the generator is utilized, resulting in no additional parameters or computational requirements compared to the baseline model Demucs. We evaluate BigWavGAN from both objective and subjective perspectives:

- The objective evaluations affirm the effectiveness of BigWavGAN in music SR.
- The subjective evaluations indicate that the proposed BigWavGAN is capable of producing high-resolution music with better perceptual quality than its baseline.
- Moreover, BigWavGAN represents its strong ability to handle out-of-distribution data.
- Notably, BigWavGAN surpasses the state-of-the-art music SR model in both simulated and real-world scenarios (i.e., historical music recordings). The results indicate that BigWavGAN successfully unleashes the potential of the baseline model without additional computation or parameters.
- At last, the ablation study unveils the importance of our discriminators and training strategies.

1.3 Thesis Outline

The outline of this thesis is as follows:

Chapter 1: we first briefly introduce the task of music super-resolution, including the applications and challenges in this task. Then, the research objectives and the outline of this thesis are presented.

Chapter 2: In this chapter, we investigate the common annoying artifacts in Time-Domain Neural Networks (TD-CNNs) and demonstrate the artifacts in TD-CNNs are caused by the phase distortion via a subjective experiment. Then, we further propose Time-Domain Phase Repair (TD-PR), which uses a neural vocoder pre-trained on the wide-band data to repair the phase components in the waveform outputs of TD-CNNs. Evaluation results indicate that TD-PR significantly improves the perceptual quality of TD-CNN baselines.

Chapter 3: In this chapter, we propose BigWavGAN, which incorporates Demucs, a large-scale wave-to-wave model containing 134M parameters, with the state-of-the-art discriminators and adversarial training strategies. BigWavGAN are proposed to further improve the performance of TD-PR and to unleash the potential of large DNN models in music SR. The evaluations indicate that the proposed BigWavGAN is capable of producing high-resolution music with better perceptual quality than its baseline. Moreover, BigWavGAN represents its strong ability to handle out-of-distribution data. Notably, BigWavGAN surpasses the state-of-the-art music SR model in both simulated and real-world scenarios (i.e., historical music recordings).

Chapter 4: In the concluding chapter, we summarize the contents in this thesis.

Chapter 5: In this chapter, we describe the future work that we are interested, including generalizing the optimal performance in music super-resolution for other types of music and exploring a task of automatic sound quality assessment.

Chapter 6: In this chapter, the list of publications are presented.

Chapter 2

Phase Repair for Time-Domain Convolutional Neural Networks in Music Super-Resolution

2.1 Time-Domain and Frequency-Domain Approaches for Super-Resolution

2.1.1 Frequency-Domain Approaches

Various works have delved into the deep learning based approaches for audio SR. Some of them work in frequency domain. Frequency-domain approaches aim to directly recover the high-resolution components in the magnitude spectrogram, and generally require additional signal processing to estimate the corresponding phase information, such as Griffin-Lim algorithms [2] or neural vocoders [4]. Li *et al.* proposed an FD approach for speech SR, which consists of 2 steps [7]. The first step is mapping the magnitude components from narrow-bandwidth to wide-bandwidth by DNN. The second step is to estimate the corresponding phase by signal processing. Following this work, Hu *et al.* introduced Generative Adversarial Network (GAN)

into both steps and got the better performance [2]. However, training two GAN-based models is difficult due to the instability of GAN training. Furthermore, this SR system works on a fixed up-sampling ratio, which limits its application to real world problems. Liu *et al.* used a GAN-based neural vocoder for the second step without using GAN in the first step, which successfully performed speech SR with the ability of handling various up-sampling ratios [4]. It is worth pointing out that the FD approaches mentioned above requires strict matching of mel-spectrogram settings between the FD-CNN model and the neural vocoder. Therefore, some FD-CNN models trained with an unmatched mel-spectrogram settings cannot directly work with the pretrained vocoder.

2.1.2 Time-Domain Convolutional Neural Network Approaches

Compared with frequency-domain approaches, Time-Domain Convolutional Neural Networks (TD-CNNs) that directly learn a wave-to-wave mapping, are considered being able to avoid the phase problem in audio SR due to the direct waveform processing [2]. AudioUNet is one of the pioneers of tackling audio SR by a TD-CNN [1]. Tagliasacchi *et al.* proposed SEANet [8], a GAN-based model for speech SR. The generator of SEANet is a light-weight but effective TD-CNN. In this chapter, we utilize the generator of SEANet to music SR as one of our baselines. Defossez *et al.* proposed a TD-CNN model referred to as Demucs, which is a large model with over 130M parameters and is initially designed to address music source separation [9]. Considering the fact that Demucs has shown strong performance in tasks besides source separation [10], we utilize the Demucs model in the SR task in this chapter. To the best of our knowledge, this is the first time to apply Demucs to the music SR task.

However, TD-CNNs tend to produce annoying artifacts in their waveform output. To alleviate the artifacts, Lim *et al.* proposed a time-frequency hybrid

model [11] based on AudioUNet. Wang *et al.* made efforts on objective function that employing the frequency domain losses [12] during the TD-CNN's training. The data augmentation strategy was proposed in [13] to improve the robustness of TD-CNNs.

Although the above efforts for TD-CNNs improved audio SR quality measured by objective metrics, none of the above TD-CNN approaches succeeds in removing the artifacts according to their open-available audio samples. We hypothesize that the inconsistency between objective and subjective evaluation results could have been caused by some signal components that cannot be measured by the objective metrics. We observe that phase components are not explicitly measured by typical objective metrics such as log-spectral distance. This observation encourages us to explore the importance of phase in audio SR.

2.2 Proposed Method: TD-PR

2.2.1 TD-PR: Time-Domain Phase Repair

In order to alleviate the artifacts caused by distorted phase components, we propose Time-Domain Phase Repair (TD-PR). The TD-PR framework consists of two separately pretrained DNN modules and a phase replacement operation.

The overview of the proposed method is shown in Fig. 3.1. Specifically, the TD-PR pipeline involves the following steps. First, a TD-CNN is trained to perform music SR. To handle low-resolution music with various bandwidths which is common in real world, we apply a simulation pipeline to high-resolution music data to get the corresponding low-resolution version. With the simulated pseudo paired data, the training of TD-CNN for music SR is made possible. Details of the simulation pipeline and training objectives are explained in the succeeding section.

Second, we pretrain a neural vocoder on the unprocessed high-resolution music

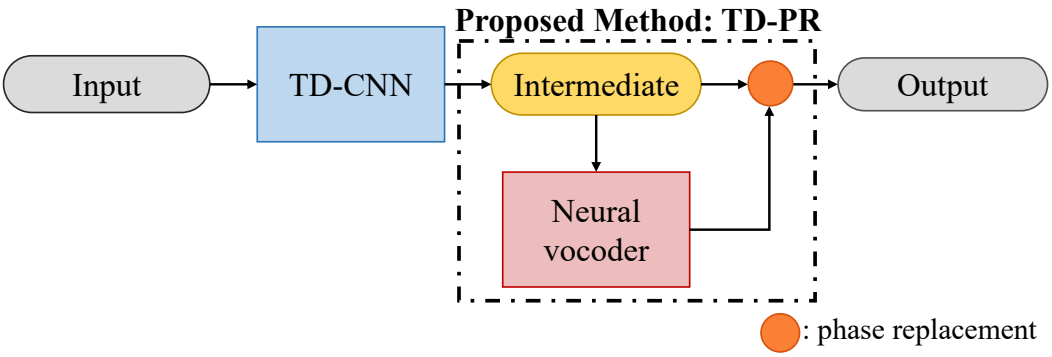

Figure 2.1: Overview of the proposed TD-PR: The TD-CNN is trained to perform super-resolution for various narrow-band inputs. The neural vocoder takes only the magnitude of the TD-CNN's output as input, and re-synthesizes another waveform that contains repaired phase components. Then, the distorted phase components in TD-CNN's output are replaced by that from the vocoder.

data. Since a neural vocoder can generate realistic waveform signals with only the magnitude input, it can be inferred that a vocoder can generate realistic phase components that are coherent with the input magnitude components. This inspires us to utilize a neural vocoder to repair distorted phase.

Last, we introduce TD-PR to repair the phase components of the output from the TD-CNN. The intermediate waveform produced by the TD-CNN is decomposed into magnitude and phase components by Short-Time Fourier Transform (STFT). We empirically use an STFT of 1024-point hann window and 256 hop length for a sampling rate of 16 kHz. The neural vocoder takes only the magnitude of the TD-CNN's output as input, and re-synthesizes another waveform that contains repaired phase components. Then, the distorted phase components in TD-CNN's output is replaced by that from the vocoder, and a phase-repaired waveform output is produced by inverse STFT. Although the vocoder also outputs waveform, we decide not to use it as the final results, because empirically we found that the vocoder could introduce distortions in the lower frequency part.

According to the above description, the vocoder and TD-CNNs are trained independently, which indicates a single pretrained vocoder can be directly applied to

arbitrary TD-CNNs without additional adaptation, making the method flexible. It is worth noting that since TD-PR only repair the phase components of the waveforms, the improved perceptual quality in turn indicates that phase distortion has been the cause of the annoying artifacts of TD-CNNs.

2.2.2 Simulation Pipeline

The design of simulation pipeline has been shown to be critical to the performance and robustness of audio SR models [12, 13]. The simulation pipeline we utilize mainly follows the principles in [12, 13]. Specifically, we simulate each low-resolution input by randomly choosing a low-pass filter from 7 low-pass filters, including Butterworth, Chebyshev type 1, Chebyshev type 2, Elliptic, Bessel, subsampling (*i.e.*, resample_poly in scipy), STFT filter (*i.e.*, replacing the high frequency components with zero elements) with the filter order randomly selected from 6 to 10. We use the implementation of low-pass filters provided by Liu *et al.*[1] [4].

Since 3 kHz has been analyzed to be the typical bandwidth of real historical recordings [5], we sample an low-resolution bandwidth between 2.5 kHz and 4 kHz via a uniform distribution. The low-pass filtering is conducted on-the-fly during training.

2.2.3 Loss Function

Inspired by [12], we perform cross-domain loss to guide TD-CNNs to capture features in both time and frequency domains. The loss function (denoted as L) is comprised of two parts, multi-resolution STFT loss (L_{MRSTFT}) [14] and multi-resolution wave loss (L_{MRwave}) which is similar to L_{MRSTFT}. The loss function is defined as below:

[1] https://github.com/haoheliu/ssr_eval

$$L = L_{\text{MRSTFT}} + \lambda L_{\text{MRwave}}, \tag{2.1}$$

where λ denotes the hyperparameter balancing the two loss terms. In our case, we empirically set $\lambda = 1000$ to balance the weights between two losses.

The definition of L_{MRSTFT} and L_{MRwave} are shown as follows:

$$L_{\text{MRSTFT}} = \frac{1}{M}\sum_{m=1}^{M} L_{\text{STFT}}^{(m)}(y,\hat{y}), \tag{2.2}$$

$$L_{\text{MRwave}} = \frac{1}{N}\sum_{n=1}^{N} L_{\text{wave}}^{(n)}(y,\hat{y}), \tag{2.3}$$

where y and $\hat{y}$ denote the ground truth and generated sample respectively. M denotes the number of STFT losses with different analysis parameters (*i.e.*, FFT size = [512, 1024, 2048]; hop size = [256, 512, 1024]; window size = [512, 1024, 2048]). We use the implementation of L_{MRSTFT} from [15]. N denotes the number of wave losses with different sampling rate (*i.e.*, original sampling rate, $2\times$ down sampling rate, $4\times$ down sampling rate).

L_{wave} is defined as follows:

$$L_{\text{wave}}(y,\hat{y}) = \frac{1}{P}\|y-\hat{y}\|_1, \tag{2.4}$$

where P denotes the number of wave samples and $\|\cdot\|_1$ denotes the L1 norms.

2.3 Experiments for TD-PR

2.3.1 Dataset and Implementation

We trained and evaluated our model on the MAESTRO dataset [16]. It is composed of about 200 hours of high-quality classical piano recordings in waveform. Although these recordings have the sampling rate of 44.1 kHz or 48 kHz, we empirically found that 16 kHz is high enough for the piano solo. Hence, we performed music SR with the target bandwidth of 8 kHz, *i.e.*, a target sampling rate 16 kHz. We used the official split of the MAESTRO dataset for training, validation and test. We cut all of the waveform into 30-second short clips for efficient training.

To implement the proposed TD-PR framework, we trained a TFGAN [17], a light-weight vocoder, from scratch on MAESTRO training set by using an unofficial implementation[2]. We followed the original settings, except resetting the sampling rate to 16 kHz, and trained it for 1M iterations.

Since TD-PR is feasible for arbitrary TD-CNNs with a single pretrained neural vocoder as mentioned in Sec. 2.2.1, we evaluated TD-PR with three representative TD-CNN models as baselines: AudioUNet [1], Demucs [9] and SEANet generator [8]. We trained them from scratch with the loss function mentioned in Sec. 2.2.3 by applying the simulation pipeline in Sec. 2.2.2 to the dataset. We used the Pytorch implementation of AudioUNet[3] and Demucs[4]. We implemented the SEANet generator by ourselves. We used an Adam optimizer and the initial learning rate 0.0001 to optimize each TD-CNN model for 200 epochs with the batch size of 12 and the input duration of 5s.

[2]https://github.com/rishikksh20/TFGAN
[3]https://github.com/serkansulun/deep-music-enhancer
[4]https://github.com/facebookresearch/demucs/tree/v2

2.3.2 Investigation into Effectiveness of Ground Truth Phase Components

Before delving into the evaluation of TD-PR, we present a preliminary study to show the impact of phase on the artifacts issue of TD-CNN models. In this study, we used SEANet a representative, and replaced the phase of the TD-CNN output with the phase of the corresponding Ground Truth (GT) music, which denoted as TD-CNN w/ GT-phase. Note that GT phase is not available in real world applications.

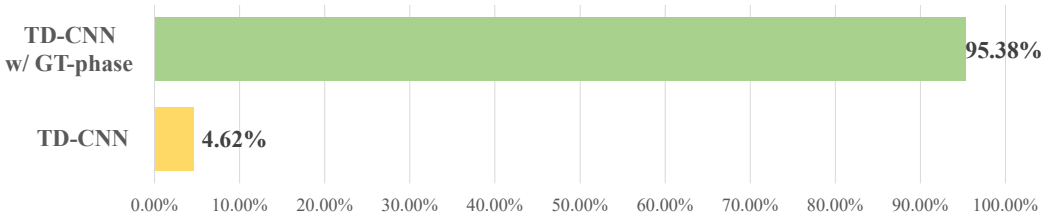

Figure 2.2: Results of the preliminary AB listening test: 95.38% of the TD-CNN w/ GT-phase is voted to have fewer artifacts.

We then conducted an AB listening test, in which we asked participants to choose the one containing fewer artifacts between the TD-CNN baseline and TD-CNN w/ GT-phase. We selected eleven music pieces for the listening test which cover different periods and styles of different musicians from the MAESTRO test set. Eleven audio pairs are presented in the AB test, in which one pair is for practice and the left ten pairs are for evaluation. Each clip is cut into the duration of 5s. We also regularized the volume of all the samples by Audacity[5]. The input bandwidth for this listening test is set to 3 kHz, as it has been analyzed to be the typical bandwidth of historical recordings [5].

[5]https://www.audacityteam.org/

2.3.3 Comparison Between TD-PR and TD-CNN Baselines

TD-PR is proposed to improve the perceptual quality of TD-CNN baselines via phase repair. We evaluated the proposed TD-PR from both objective and subjective aspects. In terms of the objective evaluation, we used the Log-Spectral Distance (LSD) as the metric, which has been widely used in audio SR tasks [1, 2, 4]. LSD is designed as:

$$LSD = \frac{1}{L}\sum_{l=1}^{L}\sqrt{\frac{1}{F}\sum_{f=1}^{F}\left(\log|Y_{l,f}|^2 - \log|\hat{Y}_{l,f}|^2\right)^2}, \tag{2.5}$$

where $Y_{l,f}$ and $\hat{Y}_{l,f}$ are the ground truth and the estimated magnitude via STFT at l-th time step $(l = 1,...,L)$ and f-th frequency bin $(k = 1,...,F)$, respectively.

The subjective evaluation aims at collecting Mean Opinion Score (MOS) from participants to compare the perceptual quality across the input low-resolution music, TD-CNN baseline, TD-CNN w/ TD-PR and ground truth high-resolution music. MOS is commonly used in audio SR tasks to represent the perceptual quality [4, 10]. Participants are asked to rate audio samples according to the similarity with the reference audio, *i.e.*, the ground truth high-resolution music. The range of MOS in our work is set from 1 to 5, where 5 denotes excellent quality (*i.e.*, is the closest to the reference) and 1 denotes bad quality. To avoid auditory fatigue caused by giving too many samples to participants, we evaluated the three TD-CNN models separately in three independent listening tests, which means the MOS values across different tests cannot be directly compared. For each TD-CNN, eleven people with no background in audio engineering participated the listening test. The same eleven music pieces and pre-processing as in the preliminary AB test are used.

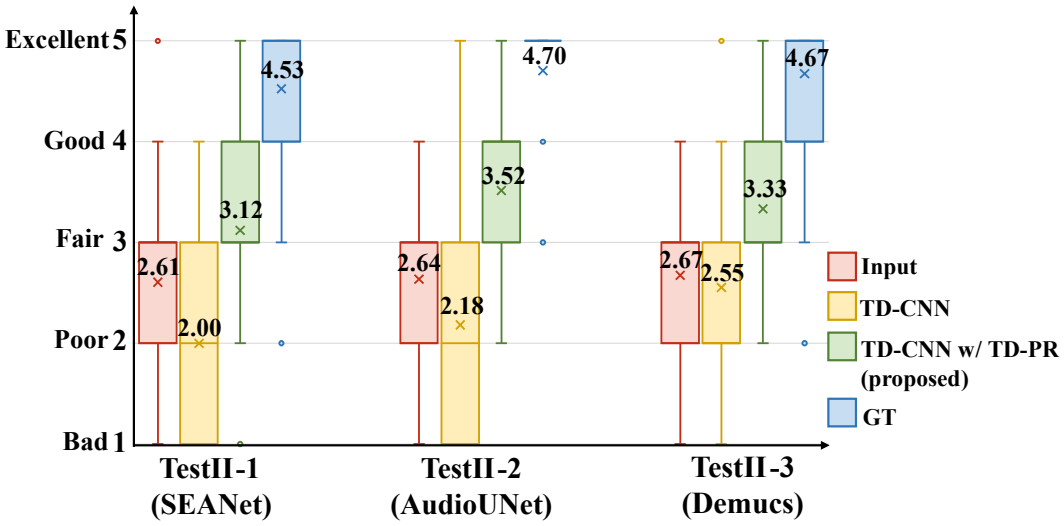

Figure 2.3: Results of MOS listening test: The box plot of the ratings across input, TD-CNN, TD-PR and GT. TD-PR is applied to three different TD-CNN baselines.

2.4 Evaluations Results and Discussion

2.4.1 Impact of Ground Truth Phase Components

The preference of the AB listening test between TD-CNN baseline and TD-CNN w/ GT-phase described in Sec. 2.3.2 is shown in Fig. 2.2. TD-CNN w/ GT-phase is voted to have fewer artifacts with a large margin (95.38% vs 4.62%). Therefore, we concluded that the artifacts in TD-CNN approaches for audio SR tasks is caused by the phase distortion, and the distortion can be repaired by replacing the distorted phase with a more realistic one.

2.4.2 Subjective and Objective Evaluations Results on TD-PR

We conducted the MOS listening test described in Sec. 2.3.3. The box plot of the MOS test results and the corresponding average for each method are shown in Fig. 2.3. First, the proposed TD-PR obtained better MOS scores than all three TD-CNN baselines by a large margin, *e.g.*, the proposed TD-PR has higher boxes, and

higher average MOS scores of 1.12 (SEANet), 1.34 (AudioUNet), 0.78 (Demucs), revealing that the TD-PR improved the perceptual quality of TD-CNN baselines significantly. Successfully improving three different baselines with a single pretrained vocoder indicates the flexibility of the proposed TD-PR method.

From the perspective of the average MOS scores between input low-resolution music and TD-CNN baselines, it is obversed that TD-CNN baselines obtained lower MOS than the low-resolution input by the deterioration of -0.61 (SEANet), -0.46 (AudioUNet), -0.12 (Demucs). This indicates that the artifacts in TD-CNNs severely harmed the perceptual quality. However, we will show later that TD-CNN baselines obtained better LSD scores (objective metric) than the low-resolution input, indicating that LSD is not a reliable metric to evaluate audio SR and perceptual quality.

Table 2.1: LSD results with different input bandwidth and parameter amount.

	2.5kHz	3kHz	3.5kHz	4kHz	AVG	Parameter
Input	2.43	2.19	1.97	1.78	2.09	-
SEANet	0.89	0.78	0.72	0.68	0.77	11M
SEANet w/ TD-PR(proposed)	0.94	0.86	0.82	0.80	0.86	11+6M
AudioUNet	0.83	**0.74**	0.69	0.66	0.73	56M
AudioUNet w/ TD-PR(proposed)	0.89	0.82	0.79	0.77	0.82	56+6M
Demucs	**0.82**	**0.74**	**0.68**	**0.64**	**0.72**	134M
Demucs w/ TD-PR(proposed)	0.89	0.83	0.79	0.77	0.82	134+6M
Ground truth	0	0	0	0	0	-

In terms of the gap of the average MOS between input and TD-CNN baselines, Demucs shows the smallest gap to the input, which implies that Demucs is the strongest among the three baselines. This observation is also in consistency with its largest parameter amount.

The LSD scores on 4 representative low-resolution bandwidth (2.5 kHz, 3 kHz, 3.5 kHz, 4 kHz) is shown in Table 2.1. Note that the proposed method can deal

with any bandwidth between 2.5 kHz and 4 kHz. The results show that both TD-PR and their TD-CNN baselines got much lower LSD than low-resolution input, indicating that music SR is successfully achieved. Although the proposed method got sightly worse LSD scores than the baselines, we argue this is trivial, because the aforementioned MOS listening test revealed a significant gap in perceptual quality between TD-PR and baselines. Although LSD can well reflect how well the high frequency magnitude is recovered in each model, it can't reflect the degree of the phase distortion and has been observed not highly correlated with perceptual audio quality in previous literature [4].

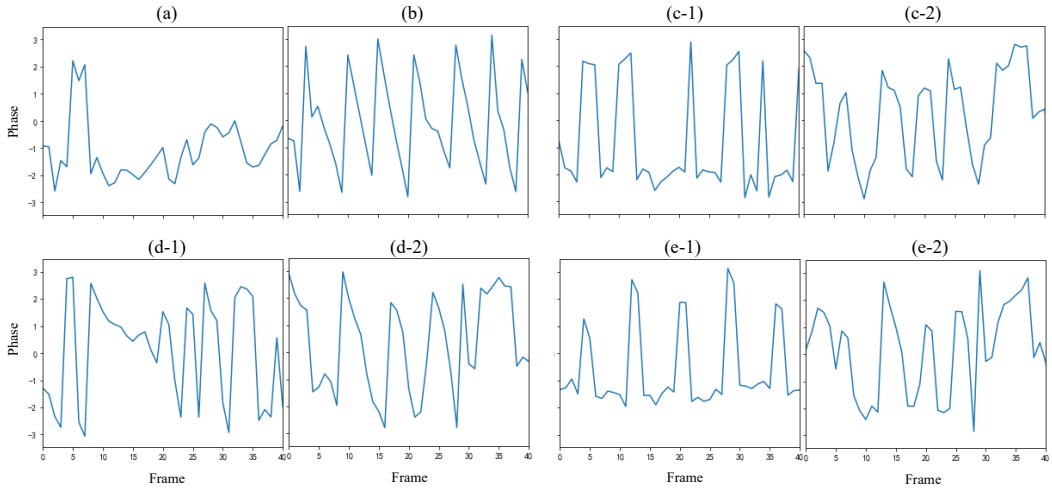

Figure 2.4: Visualization of a set of phase spectrograms: (a) low-resolution input; (b) ground truth; (c-1) SEANet; (c-2) SEANet w/ TD-PR (proposed); (d-1) AudioUNet; (d-2) AudioUNet w/ TD-PR (proposed); (e-2) Demcus; (e-2) Demucs w/ TD-PR (proposed).

2.4.3 Qualitative Evaluation of TD-PR

We visualize a part of phase spectrograms in Fig. 2.4 and their corresponding magnitude spectrograms in Fig. 2.5 to qualitatively evaluate the proposed TD-PR method. The visualizations include the spectrograms of low-resolution input,

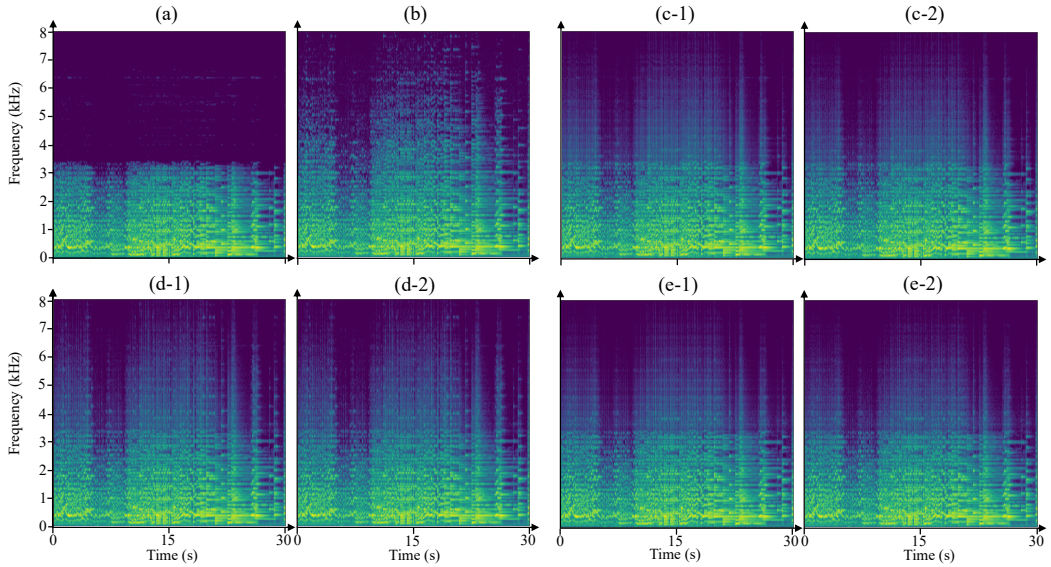

Figure 2.5: Visualization of a set of magnitude spectrograms: (a) low-resolution input; (b) ground truth; (c-1) SEANet; (c-2) SEANet w/ TD-PR (proposed); (d-1) AudioUNet: (d-2) AudioUNet w/ TD-PR (proposed); (e-2) Demcus; (e-2) Demucs w/ TD-PR (proposed).

ground truth, three TD-CNN baselines and their corresponding TD-PR outputs. For a clear view in Fig. 2.4, we plot only the phase of a single frequency bin for the first 40 time frames of an audio sample, as the phase spectrogram across multiple frequency bins is difficult to understand. The visualizations reveal that the proposed TD-PR successfully produced a phase distribution that is closer to ground truth's compared to TD-CNN baselines. Meanwhile, as TD-PR only repairs the phase components, we cannot observe significant differences in magnitude spectrograms shown in Fig. 2.5. Nevertheless, perceptual quality is improved significantly by TD-PR. The visualizations again validate that phase distortion has been the cause of the annoying artifacts in TD-CNNs.

Chapter 3

BigWavGAN: A Wave-To-Wave Generative Adversarial Network for Music Super-Resolution

3.1 Existing Challenges in Music Super-Resolution

3.1.1 Large-scale Time-Domain Neural Network Fails to Achieve Optimal Performance Commensurate with Its Model Size

DNNs are generally associated with high performance when the model size is large. However, the discussion in Chapter 2 indicates that the large-scale model referred to as Demusc, cannot generate music with the quality that is commensurate with its model size in music SR, mainly due to phase distortion [18]. Demucs is a large-scale model initially designed for music source separation [9] but also generated fairly good results in other tasks, such as music SR [18] and music enhancement [19]. Due to the large size of Demucs, it was anticipated to produce high-quality results in music SR. However, Demucs still yielded results with annoying artifacts

in music SR. We attribute this to that models cannot learn information (*e.g.*, correct phase information) commensurate with their size through standard Mean Square Error (MSE) losses. In addition, we also investigated that besides Demucs, AudioUNet [1] and SEANet [8] cannot generate high-quality audio due to phase problem when trained by standard MSE losses. To address the artifacts, we employed a neural vocoder to rectify the distorted phase generated by Demucs. Nevertheless, the improvements brought by phase repair remain limited, which indicates that introducing adversarial training into the model can lead the model to learn more information.

3.1.2 Generative Adversarial Network Approaches in Audio Super-Resolution

Recent publications have delved into GAN-based models in audio SR. Compared to models trained with standard MSE losses, GAN-based models exhibit a superior capability to generate results with better perceptual quality [5]. BEHMGAN is the state-of-the-art of GAN-based music SR model. It comprises a complex U-net as the generator and the Multi-Scale Discriminator (MSD) from MelGAN [20]. MelGAN is the first work that successfully synthesizes realistic speeches by training GANs without additional distillation or perceptual loss functions. In recent years, several works utilized neural vocoders to address audio SR tasks, mapping mel-spectrogram to raw waveform [2, 4, 18]. TFGAN is a lightweight vocoder for speech, which employs MSD and a single-resolution frequency discriminator as its discriminator [17]. TFGAN has been used in audio SR [4, 18]. Jiang *et al.* proposed an advanced neural vocoder named UnivNet, in which Multi-Resolution Discriminator (MRD) was proposed and was proved to effectively improve the performance of MelGAN [21]. Notably, a state-of-the-art neural vocoder referred to as BigVGAN, which characterized by a large-size generator with an unprecedented scale of up to 112M parameters is proposed by Lee *et al.* [22]. BigVGAN can synthesize high-fidelity audio and shows its superior zero-shot performance across

various out-of-distribution scenarios. However, in the task of audio SR, there is no wave-to-wave GAN-based model in such a large model size. This inspired us to explore the large-scale wave-to-wave GAN model in music SR with high performance and superior generalization ability.

3.2 Proposed Method: BigWavGAN

Although Demucs is a large-scale model with 134M parameters, it did not generate high-quality waveforms commensurate with its large size in music SR [18]. To unleash the potential of Demucs, we propose BigWavGAN for wave-to-wave music SR, which incorporates Demucs with state-of-the-art discriminators and adversarial training strategies.

3.2.1 Architecture of BigWavGAN

The overview of BigWavGAN's architecture is shown in Fig. 3.1. The generator of BigWavGAN has the identical architecture with Demcus from [9]. It is a wave domain U-net model leveraging a Long Short-Term Memory (LSTM) recurrent neural network layer as the bottleneck.

BigWavGAN benefits from the two types of discriminators: MSD and MRD. MSD works in the time domain, where each sub-discriminator receives downsampled 1-D waveform signals at downsampling ratios of 1, 2, and 4. MRD works in the frequency domain, which also comprises several sub-discriminators operating on multiple 2-D spectrograms with different Short-Time Fourier Transform (STFT) resolutions. On top of standard MSE losses, applying different types of discriminators to cross domains (*i.e.*, time and frequency domains) guides the generator to restore high-resolution music that is realistic in multiple domains and resolutions, minimizing annoying artifacts that are common for wave-to-wave models.

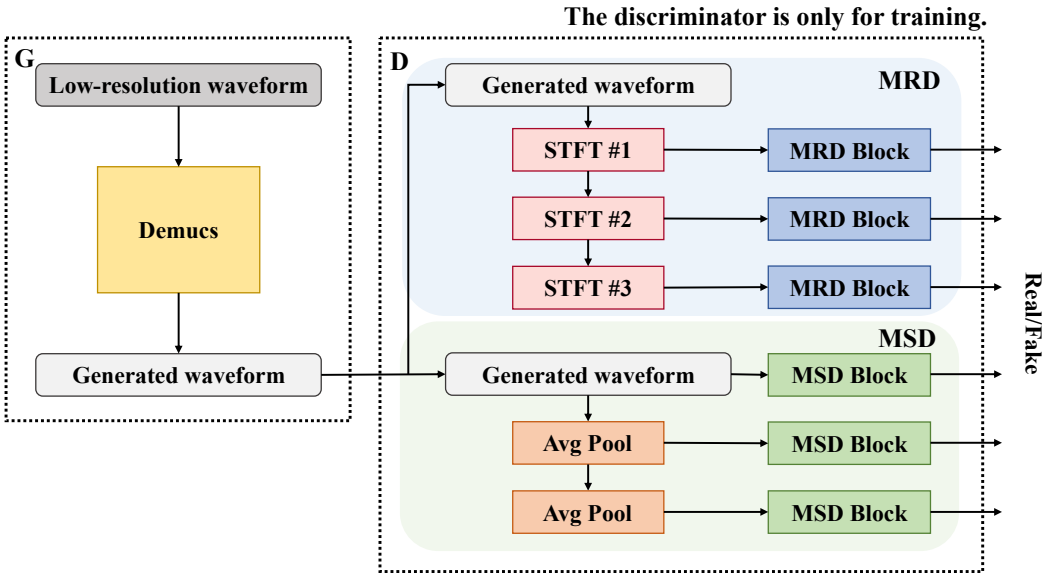

Figure 3.1: Overview of the architecture of BigWavGAN.

Our choice of MRD with MSD is not common in related vocoder publications, in which the Multi-Period Discriminator (MPD) is widely used [23, 21, 22]. However, since MPD reshapes the 1-D waveform into 2-D matrices at multiple periods, it requires much more computational resources than MSD, making the training difficult for low-resource environments.

To improve training efficiency, we decided to replace MPD by MSD. Although the design of MPD and MSD is different, they all work in the time domain, which implies that MSD could be an alternative to MPD in order to similarly capture details in the waveform. The evaluation results in section 3.4 reveal BigWavGAN's superior performance, validating the success of combining MSD and MRD as the discriminator.

Adversarial training of large-scale models tend to be unstable. To stabilize the training, we utilized the training strategies of BigVGAN. Lee *et al.* [22] made lots of efforts on maintaining the stability of large-scale GAN training and the high-speed practical usability. We believe that these training strategies are suitable for training non-vocoder models with a similar scale, and introduced these strategies into BigWavGAN's training to ensure training stability.

3.2.2 Training Objectives

In terms of training objectives, we applied L_G for generator and L_D for discriminator, respectively:

$$L_G = \sum_{k=1}^{K} \Big[L_{adv}(G;D_k) + \lambda_{fm} L_{fm}(G;D_k) \Big] + \lambda_{mel} L_{mel}(G), \tag{3.1}$$

$$L_D = \sum_{k=1}^{K} \Big[L_{adv}(D_k;G) \Big], \tag{3.2}$$

where $K = 3$, D_k denotes the k-th MSD or MRD sub-modules. L_{adv} stands for adversarial losses, L_{fm} stands for feature matching losses, L_{mel} stands for mel losses. We use the scalar weights $\lambda_{fm} = 2$ and $\lambda_{mel} = 45$ identically as [22].

L_{adv} uses the least-square GAN as follows:

$$L_{adv}(G;D_k) = E_s \Big[(D_k(G(s)) - 1)^2 \Big], \tag{3.3}$$

$$L_{adv}(D_k;G) = E_{(x,s)} \Big[(D_k(x) - 1)^2 + (D_k(G(s)))^2 \Big], \tag{3.4}$$

where s is the input low-resolution waveform, x is the ground-truth waveform.

The feature matching loss L_{fm} minimizes the l_1 distance for every intermediate features from the discriminator layers:

$$L_{fm}(G;D_k) = E_{(x,s)} \Big[\sum_{i=1}^{T} \frac{1}{N} ||D_k^i(x) - D_k^i(G(s))||_1 \Big], \tag{3.5}$$

where T is the number of layers of the sub-discriminator D_k.

The generator loss L_G also has the spectral l_1 regression loss between the mel spectrogram of the synthesized waveform and the corresponding ground-truth:

$$L_{mel}(G) = E_{(x,s)}\left[||\phi(x) - \phi(G(s))||_1\right], \tag{3.6}$$

where ϕ is the STFT with mel filter bank that converts the waveform into mel-spectrogram.

3.3 Experiments for BigWavGAN

We used the MAESTRO dataset [16] for training. We simulated the low-resolution music by means of following [4, 18]. To handle real-world low-resolution music recordings which have various bandwidths, we simulated the input bandwidth ranging from 2.0 kHz to 4.0 kHz on the fly during training. The models involved in our evaluation all work at the sampling rate of 16 kHz with a target bandwidth of 8 kHz, except BEHMGAN. The configurations of the low-pass filters used to simulate low-resolution audio are identical to that in [18]. Hereby, the proposed BigWavGAN can deal with any bandwidths between 2.0 kHz and 4.0 kHz.

The implementation of BigWavGAN's generator (*i.e.*, Demucs) is from [9]. We implemented MSD and MRD by utilizing the open-source code from [22] and [23] respectively. During training, the batch size is 10, each music segment is 2.56 seconds long. We trained BigWavGAN for 1M iterations with the same training strategies as BigVGAN [22]. As BigVGAN is similar to our BigWavGAN in model size, keeping the same training strategy contributed to the stable training of BigWavGAN.

For the baseline Demucs, we used the checkpoint from [18]. We used the official checkpoints of BEHMGAN [5] for comparison. Music generated by BEHMGAN were resampled from 22.05 kHz to 16 kHz for a fair evaluation. Furthermore, in or-

der to explore the effectiveness of the discriminator and training strategies, we also trained a model denoted as BigWavGAN w/o MRD which is trained by discriminators and strategies from TFGAN [17]. TFGAN combines MSD with a single-resolution frequency discriminator instead of MRD. We implemented this training by using an unofficial implementation[1] and trained this model for 1M iterations.

3.4 Evaluations on BigWavGAN

We evaluated the proposed BigWavGAN from both objective and subjective perspectives.

3.4.1 Objective Evaluations

We used Log-Spectral Distance (LSD) as the objective metric, which is widely used in audio SR tasks [4, 5]. We calculated the LSD scores at four representative bandwidths (*i.e.*, 2.5 kHz, 3.0 kHz, 3.5 kHz, 4.0 kHz). The results of LSD are illustrated in Tab. 3.1. Note that the proposed BigWavGAN can handle any bandwidth from 2.0 kHz to 4.0 kHz.

Table 3.1: LSD scores on the MAESTRO dataset. The bold represents the top two LSD scores.

	MSD	MRD	2.5 kHz	3.0 kHz	3.5 kHz	4.0 kHz	AVG LSD
Input	-	-	2.43	2.19	1.97	1.78	2.09
BEHMGAN [5]		-	1.89	1.01	1.79	1.77	1.61
BigWavGAN (proposed)			**0.83**	**0.79**	**0.76**	**0.73**	**0.78**
- w/o MRD		-	0.93	0.88	0.82	**0.73**	0.84
- w/o MSD (Demucs)	-	-	**0.82**	**0.74**	**0.68**	**0.64**	**0.72**

[1] https://github.com/rishikksh20/TFGAN

In terms of LSD scores, the four models all successfully achieved music SR since all the generated results received much better LSD scores than low-resolution inputs. BEHMGAN was trained on inputs with bandwidths around 3.0 kHz, as 3.0 kHz was believed to be the typical bandwidth of real historical recordings [5]. Consequently, BEHMGAN performed well at 3.0 kHz. Nevertheless, the proposed BigWavGAN still outperformed BEHMGAN at this bandwidth.

In order to explore the importance of the discriminator and training strategies, we compared BigWavGAN with a variant that has only a single-resolution frequency discriminator combined with the MSD, *i.e.*, BigWavGAN w/o MRD. We found that the proposed BigWavGAN, which utilizes MRD and MSD with the training strategies from [23, 22], outperformed the "w/o MRD" variant overall. Since LSD is a metric working on the frequency domain, compared with the single-resolution frequency discriminator, the multi-resolution frequency discriminator (MRD) seems to have improved the LSD score by forcing the model to concentrate more on the fidelity of music in the frequency domain.

The proposed BigWavGAN acquired a slightly worse LSD score than the baseline model Demucs. We consider this difference in LSD score as the result of the common phenomenon that objective metrics tend to give generative methods lower scores than their non-generative counterparts [4, 18]. This can be explained as that generative models tend to generate results similar rather than exactly identical to ground truth. To show that BigWavGAN can restore music with better perceptual quality, subjective evaluations were conducted.

3.4.2 Subjective Evaluations

Although LSD can well reflect how well the high frequency in the magnitude is recovered, it cannot reflect the degree of the artifacts and has been observed not to correlate with perceptual audio quality [4, 18]. To this end, we conducted a set of subjective evaluations to identify the advantage of the proposed BigWavGAN. The

subjective evaluation is in the style of A/B test, rather than mean opinion score test, because A/B test can better measure tiny differences between two models. Since A/B test cannot handle multiple models at once, we conduct multiple A/B tests (*e.g.*, BigWavGAN vs Demucs, BigWavGAN vs BEHMGAN) for a more comprehensive analysis.

We conducted A/B tests on 3 different datasets: (a) four tracks from MAESTRO [16], (b) four piano tracks from MusicNet [24], and (c) for real historical piano recordings provided in [5]. We only trained our BigWavGAN on MAESTRO dataset, then applied it to out-of-distribution data (*i.e.*, MusicNet, and real-world historical recordings) in the zero-shot condition to evaluate its generalization ability. To avoid too many testing samples and the consequent auditory fatigue in participants, the input bandwidth is set to 3.0 kHz, except for the real-world recordings. We selected the above 12 tracks to cover different musicians, periods, and styles. The duration of each music clip has been standardized to 10 seconds, and all audio clips have a normalized loudness for accurate evaluation. Twelve and eleven people with no background in audio engineering participated our subjective tests for BigWavGAN vs Demucs and BigWavGAN vs BEHMGAN respectively

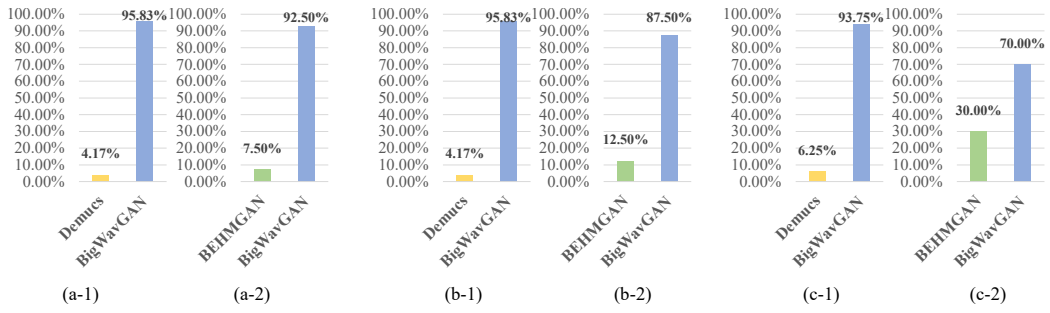

Figure 3.2: Results of A/B listening tests: (a) is tested on MAESTRO; (b) is tested on MusicNet; (c) is tested on denoised real historical recordings.

The results of subjective evaluations are illustrated in Fig 3.2. In all three datasets, BigWavGAN significantly improved Demucs in terms of perceptual qual-

ity by a large margin. This also reveals that BigWavGAN achieved superior generalization to out-of-distribution data. Similar advantages of BigWavGAN are observed when it is compared with BEHMGAN, the state-of-the-art music SR model. We further analyze the trends in the preference of BigWavGAN and BEHMGAN. First, in Fig. 3.2 we can see that BEHMGAN's preference increased from MAESTRO (a-2) to MusicNet (b-2), *i.e.*, the preferences changed from 7.50% vs 92.50% to 12.50% vs 87.50%. This is because BEHMGAN was trained on MusicNet. Although MusicNet is out-of-distribution data to our BigWavGAN, we outperformed BEHMGAN by a large margin, validating the strong generalization of BigWavGAN again.

Although in the real-world historical recordings, the preference of BEHMGAN further increased to 30.00% vs 70.00%, BigWavGAN still outperformed BEHMGAN with a large margin. We think the less advantage of BigWavGAN in real-world condition is due to our limited simulation in training data. In the future, we would like to explore more realistic simulation techniques and further improve our model's performance in real-world historical recordings.

Chapter 4

Conclusion

In this thesis, we explore music Super-Resolution (SR) through solo piano music.

First, we delved into Time-Domain Convolutional Neural Networks (TD-CNNs), trying to identify the cause of the annoying artifacts and improve TD-CNNs' perceptual quality by alleviating the artifacts. To the best of our knowledge, this work is the first to demonstrate the artifacts in TD-CNNs are caused by the phase distortion via a subjective experiment. We further propose Time-Domain Phase Repair (TD-PR), which uses a neural vocoder pretrained on the wide-band data to repair the phase components in the waveform output of TD-CNNs. The proposed TD-PR achieved better mean opinion score, significantly improving the perceptual quality of TD-CNN baselines. Moreover, a single pretrained vocoder can be directly applied to arbitrary TD-CNNs without additional adaptation. Since the proposed TD-PR only repairs the phase components of waveform, the improved perceptual quality in turn indicates that phase distortion has been the cause of the annoying artifacts of TD-CNNs. The findings and comprehensive evaluations presented in this work offer a new perspective for the future improvement of audio super-resolution algorithms. This work inspires us to combine the advantages of TD-CNNs and neural vocoders, to develop a model that can better address the challenges in music super-resolution.

Then, based on the discussion of our first proposed method TD-PR, we proposed a large-scale wave-to-wave model referred to as BigWavGAN for music SR. The model integrates a large-size generator (*i.e.*, Demucs with up to 134M parameters), with the state-of-the-art discriminators and adversarial training strategies. The discriminator of the proposed BigWavGAN consists of Multi-Scale Discriminator (MSD) and Multi-Resolution Discriminator (MRD). During inference phase since only the generator will be used, there are no additional parameters or computational resources required during inference compared to the baseline model Demucs. We evaluated BigWavGAN from both objective and subjective perspectives. The objective evaluation indicates the effectiveness of BigWavGAN in music SR. The results of a set of subjective evaluations demonstrate that BigWavGAN can produce high-resolution music in significantly better perceptual quality compared to the baseline model Demucs. Notably, the subjective evaluations also indicate that BigWavGAN surpasses the state-of-the-art music SR model in both simulated and real-world scenarios (i.e., historical music recordings). Moreover, it also implies that BigWavGAN achieves superior generalization ability to address out-of-distribution data including real historical recordings. Therefore, BigWavGAN successfully unleashes the potential of the large-scale Demucs in music SR.

Chapter 5

Future Work

5.1 Explore Music Super-Resolution for Various Types of Music

Given the superior performance of the proposed method BigWavGAN, in the future, we hope to extend the applications of BigWavGAN to other types of music instead of solo piano music, e.g., violin music, flute music, symphony music, pop music, etc. Moreover, we hope to explore more realistic simulation techniques to further improve BigWavGAN in real-world scenarios as well as to further extend BigWavGAN to more tasks.

5.2 Explore Automatic Sound Quality Assessment

In order to comprehensively evaluate the performance of music super-resolution models, we conducted listening tests several times as subjective evaluations. Subjective listening tests have been considered as the golden standard for sound quality assessment. However, a subjective test is costly and not scalable to very huge test data [25]. To automate the sound quality assessment process, various objective met-

rics have been developed and have been widely utilized. However, prior works have indicated that objective metrics poorly correlated with human perception. Therefore, we are interested in developing a deep learning driven method for automatic sound quality assessment by predicting the mean opinion score.

Chapter 6

List of Publications

- Yenan Zhang, Guilly Kolkman, and Hiroshi Watanabe. Phase repair for time-domain convolutional neural networks in music super-resolution. 2024 Sound and Music Computing (SMC), July, 2024.
- Yenan Zhang, and Hiroshi Watanabe. BigWavGAN: A Wave-To-Wave Generative Adversarial Network for Music Super-Resolution. 2023 IEEE 12th Global Conference on Consumer Electronics (GCCE). pp. 593–597, October, IEEE, 2023.

Bibliography

[1] Volodymyr Kuleshov, S Zayd Enam, and Stefano Ermon. Audio super resolution using neural networks. *arXiv preprint arXiv:1708.00853*, 2017.

[2] Shichao Hu, Bin Zhang, Beici Liang, Ethan Zhao, and Simon Lui. Phase-aware music super-resolution using generative adversarial networks. *arXiv preprint arXiv:2010.04506*, 2020.

[3] Yunpeng Li, Marco Tagliasacchi, Oleg Rybakov, Victor Ungureanu, and Dominik Roblek. Real-time speech frequency bandwidth extension. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 691–695. IEEE, 2021.

[4] Haohe Liu, Woosung Choi, Xubo Liu, Qiuqiang Kong, Qiao Tian, and DeLiang Wang. Neural vocoder is all you need for speech super-resolution. *arXiv preprint arXiv:2203.14941*, 2022.

[5] Eloi Moliner and Vesa Välimäki. Behm-gan: Bandwidth extension of historical music using generative adversarial networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:943–956, 2022.

[6] Paulo AA Esquef. Audio restoration. In *Handbook of Signal Processing in Acoustics*, pages 773–784. Springer, 2008.

[7] Kehuang Li and Chin-Hui Lee. A deep neural network approach to speech bandwidth expansion. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4395–4399. IEEE, 2015.

[8] Marco Tagliasacchi, Yunpeng Li, Karolis Misiunas, and Dominik Roblek. Seanet: A multi-modal speech enhancement network. *arXiv preprint arXiv:2009.02095*, 2020.

[9] Alexandre Défossez, Nicolas Usunier, Léon Bottou, and Francis Bach. Demucs: Deep extractor for music sources with extra unlabeled data remixed. *arXiv preprint arXiv:1909.01174*, 2019.

[10] Jiaqi Su, Yunyun Wang, Adam Finkelstein, and Zeyu Jin. Bandwidth extension is all you need. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 696–700. IEEE, 2021.

[11] Teck Yian Lim, Raymond A Yeh, Yijia Xu, Minh N Do, and Mark Hasegawa-Johnson. Time-frequency networks for audio super-resolution. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 646–650. IEEE, 2018.

[12] Heming Wang and DeLiang Wang. Towards robust speech super-resolution. *IEEE/ACM transactions on audio, speech, and language processing*, 29:2058–2066, 2021.

[13] Serkan Sulun and Matthew EP Davies. On filter generalization for music bandwidth extension using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 15(1):132–142, 2020.

[14] Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *ICASSP 2020*, pages 6199–6203. IEEE, 2020.

[15] Christian J Steinmetz and Joshua D Reiss. auraloss: Audio focused loss functions in pytorch. In *Digital music research network one-day workshop (DMRN+ 15)*, 2020.

[16] Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, Sander Dieleman, Erich Elsen, Jesse Engel, and Douglas Eck. Enabling factorized piano music modeling and generation with the maestro dataset. *arXiv preprint arXiv:1810.12247*, 2018.

[17] Qiao Tian, Yi Chen, Zewang Zhang, Heng Lu, Linghui Chen, Lei Xie, and Shan Liu. Tfgan: Time and frequency domain based generative adversarial network for high-fidelity speech synthesis. *arXiv preprint arXiv:2011.12206*, 2020.

[18] Yenan Zhang, Guilly Kolkman, and Hiroshi Watanabe. Phase repair for time-domain convolutional neural networks in music super-resolution. *arXiv preprint arXiv:2306.11282*, 2023.

[19] Nikhil Kandpal, Oriol Nieto, and Zeyu Jin. Music enhancement via image translation and vocoding. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3124–3128. IEEE, 2022.

[20] Kundan Kumar, Rithesh Kumar, Thibault De Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre De Brebisson, Yoshua Bengio, and Aaron C Courville. Melgan: Generative adversarial networks for conditional waveform synthesis. *Advances in neural information processing systems*, 32, 2019.

[21] Won Jang, Dan Lim, Jaesam Yoon, Bongwan Kim, and Juntae Kim. Univnet: A neural vocoder with multi-resolution spectrogram discriminators for high-fidelity waveform generation. *arXiv preprint arXiv:2106.07889*, 2021.

[22] Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. Bigvgan: A universal neural vocoder with large-scale training. *arXiv preprint arXiv:2206.04658*, 2022.

[23] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33:17022–17033, 2020.

[24] John Thickstun, Zaid Harchaoui, and Sham Kakade. Learning features of music from scratch. *arXiv preprint arXiv:1611.09827*, 2016.

[25] Chandan KA Reddy, Vishak Gopal, and Ross Cutler. Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6493–6497. IEEE, 2021.