MusicHiFi: Fast High-Fidelity Stereo Vocoding

Ge Zhu[1,2*], Juan-Pablo Caceres[2], Zhiyao Duan[1], and Nicholas J. Bryan[2]

[1] University of Rochester   [2] Adobe Research

Abstract—Diffusion-based audio and music generation models commonly perform generation by constructing an image representation of audio (e.g., a mel-spectrogram) and then convert it to audio using a phase reconstruction model or vocoder. Typical vocoders, however, produce monophonic audio at lower resolutions (e.g., 16-24 kHz), which limits their usefulness. We propose MusicHiFi — an efficient high-fidelity stereophonic vocoder. Our method employs a cascade of three generative adversarial networks (GANs) that convert low-resolution mel-spectrograms to audio, upsamples to high-resolution audio via bandwidth extension, and upmixes to stereophonic audio. Compared to past work, we propose 1) a unified GAN-based generator and discriminator architecture and training procedure for each stage of our cascade, 2) a new fast, near downsampling-compatible bandwidth extension module, and 3) a new fast downmix-compatible mono-to-stereo upmixer that ensures the preservation of monophonic content in the output. We evaluate our approach using objective and subjective listening tests and find our approach yields comparable or better audio quality, better spatialization control, and significantly faster inference speed compared to past work. Sound examples are at https://MusicHiFi.github.io/web/.

Index Terms—music generation, mel-spectrogram inversion, bandwidth extension, mono-to-stereo upmixing

I. INTRODUCTION

Recent generation methods offer an exciting new way to create media content. Diffusion models [1], in particular, have shown great promise for fast, high-quality image generation [2], [3] and are rapidly being adopted for audio and music generation [4]–[11]. When used for audio [5]–[8], diffusion models are commonly used to generate an image representation of audio (e.g. mel-spectrogram) and then a phase reconstruction model or vocoder is used to convert to audio. This two-stage cascaded approach has been shown beneficial [6], but typically generates low-resolution audio (e.g., mono 16-24 kHz) via a diffusion [12] or generative adversarial network (GAN) vocoder [13]–[15].

To improve diffusion-based generation quality for audio and music, a high-fidelity vocoder is essential. Current methods to do so largely come in two forms: bandwidth extension (expansion) and mono-to-stereo upmixing. Bandwidth extension (BWE) increases the frequency range or bandwidth of an audio signal and can be achieved using a source-filter model [16], time-domain [17]–[19], or spectral-domain [20]–[24] neural network. Mono-to-stereo (M2S) upmixing converts a single-channel audio signal into spatialized (left and right) channels. M2S can be achieved via decorrelation [25] or a recent parametric stereo method [26] that describes audio as a single-channel plus parameters that approx. characterizes the stereo image [27].

This work is done during an internship at Adobe Research. Correspondence to: Ge Zhu (gzhu6@ur.rochester.edu) and Nicholas J. Bryan (njb@ieee.org).
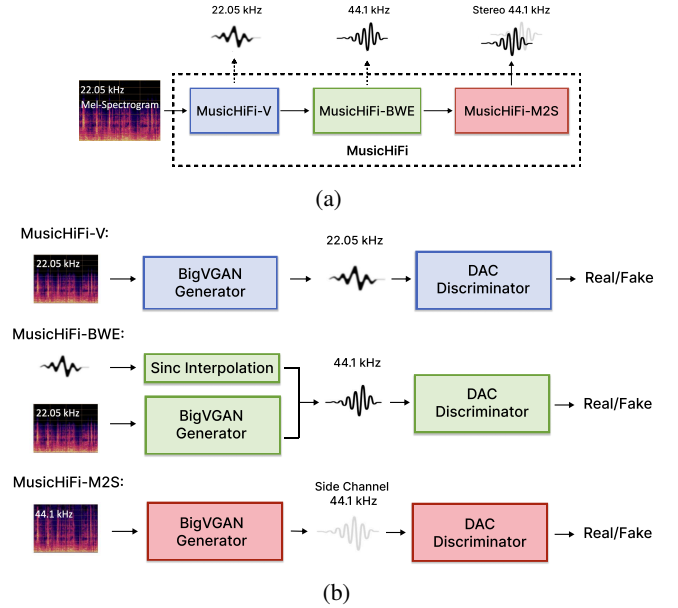

Fig. 1: (a) MusicHiFi inference via a cascade of our vocoder, bandwidth extension (BWE), and mono-to-stereo (M2S) modules that use a shared architecture, but with different weights. (b) MusicHiFi GAN-training for our vocoder, BWE and M2S (top-to-bottom), separately. Our unified inference and training scheme enables novel, high-performing BWE and M2S.

We propose MusicHiFi — a new efficient high-fidelity stereophonic vocoder shown in Fig. 1a. Our method uses a GAN-triplet cascade that converts low-resolution single-channel mel-spectrograms (e.g., 22.05 kHz) to stereo high-resolution waveforms (e.g., stereo 44.1 kHz) via vocoder, BWE, and M2S modules. Our method can be integrated into mel-spectrograms music generators, used to enhance recording fidelity, and/or for spatialization. Compared to diffusion-based BWE methods [24] and AR spatialization, our GAN-based method is much faster and differentiable. We evaluate our approach via objective and subjective listening tests and find our approach yields equal or better vocoding, BWE, and M2S quality with much inference speed. Our contributions are:

- A unified GAN-based generator, discriminator, and training recipe for vocoding, BWE, and M2S upmixing,
- A new fast BWE method that maintains low-frequency content in upsampled audio via skip connections, and
- A new fast mono-to-stereo method that uses a middle-side stereo encoding [28] that fully preserves monophonic content and offers superior spatialization width control.

By creating a unified approach to vocoding, BWE, and M2S,

we achieve a significant improvement in audio quality, spatialization, and inference speed compared to past methods.

II. BACKGROUND

The most relevant recent works to our method include BigV-GAN [15] and the Descript audio codec (DAC) [29]. BigVGAN is a recently proposed vocoder method that has achieved state-of-the-art performance for generating high-fidelity waveforms from mel-spectrograms at speeds significantly faster than real time on a single GPU [15]. The BigVGAN generator employs a neural network architecture composed of a stack of transposed 1D convolutions, each followed by an anti-aliasing multi-periodicity composition (AMP) block which internally uses a Snake activation function [30]. Past work demonstrated AMP blocks are able to generate waveforms with fewer high-frequency artifacts and provide substantial improvements in both objective and subjective assessments. Furthermore, AMP blocks have been found to improve robustness for out-of-distribution vocoding and strong extrapolation ability.

DAC is a neural-based codec that follows the SoundStream generator [31] architecture with Snake activations, an improved GAN-based discriminator, an enhanced reconstruction training objective, and a residual vector quantization scheme that achieves state-of-the-art compression. When we look at the discriminator differences, both the BigVGAN and DAC use time-domain multi-period discriminators (MPD) to capture multiple periodic structures as well as spectral-domain discriminators, but DAC replaces the BigVGAN magnitude-only spectral discriminators with a multi-band multi-resolution complex spectrogram discriminators (MMSD) to enhance high-frequency prediction and mitigate aliasing [29]. Compared to the reconstruction loss proposed by BigVGAN, DAC uses multiple mel-bins over multi-scale spectrograms to improve training stability and convergence speed.

III. METHODOLOGY

A. Overview

We introduce MusicHiFi, a new vocoding method based on unified triplet-GAN cascade that progressively upsamples audio as shown in Fig. 1a. Our approach involves three stages, each being modular and useful for different applications. First, single-channel mel-spectrograms of low sampling rate are transformed into waveforms of the same sampling rate using a vocoder (MusicHiFi-V). Then, waveforms of low sampling rate are converted into waveforms with full-bandwidth via our BWE module (MusicHiFi-BWE). Finally, the single-channel waveforms are upmixed to stereo audio through our M2S module (MusicHiFi-M2S). At each stage, we use an identical generator architecture, discriminator architecture, training objective as shown in Fig. 1b. The main differences between the three modules are the varying input and output dimensions, and the inclusion of an additional skip connection from the input in the BWE module.

In more detail, for all three of our generator stages, we adopt the BigVGAN transpose 1D convolution + AMP block generator architecture [15] that inputs a mel-spectrogram and outputs audio. For our discriminator architecture, we use the

DAC discriminator [29]. For our training objective, we also adopt the DAC reconstruction loss and adversarial loss and remove the codebook learning objective since our focus is high-fidelity audio synthesis. Our final training objective for our generator ($\mathcal{L}_G$) and discriminator ($\mathcal{L}_D$) are:

$$\mathcal{L}_G = \sum_{k=1}^{K} \left[\mathcal{L}_{adv}(G; D_k) + \lambda_{fm}\mathcal{L}_{fm}(G; D_k) \right] + \lambda_{rc}\mathcal{L}_{rc}(G),$$

$$\mathcal{L}_D = \sum_{k=1}^{K} \left[\mathcal{L}_{adv}(D_k; G) \right],$$

(1)

where we apply K discriminators, D_k denotes the k-th sub-discriminator from MPD or MMSD, $\mathcal{L}_{adv}, \mathcal{L}_{rc}, \mathcal{L}_{fm}$ represent least-square adversarial loss [32], reconstruction loss and L_1-based feature matching loss, respectively and λ_{rc} and λ_{fm} represent corresponding loss weighting term. Specifically, $\mathcal{L}_{rc} = \sum_i ||\log S_i - \log \hat{S}_i||_1$, where S_i indicates the i-th mel-spectrogram from a list of mel-spectrograms with different fixed resolutions. L_{fm} aims to minimize the distances between real and generated features from discriminator layers [13].

B. MusicHiFi-V

Our vocoder inputs mel-spectrograms of low sampling-rate and outputs audio waveforms with the same sampling rate, and follows our unified generator, discriminator, and training recipe described above. We note that the original BigVGAN training recipe exhibits instabilities and is susceptible to mode collapse when scaled up to larger models [15]. To enhance performance while stabilizing training, we double the input length of the audio samples originally defined in BigVGAN, reduce the number of convolution layers in the AMP blocks, increase the convolution channel width to 2048. Furthermore, this configuration approximately matches the floating point operations per second (FLOPS) of HiFi-GAN [14].

C. MusicHiFi-BWE

Our BWE module takes audio with a low sampling rate as input and outputs full-band audio and follows our unified generator and discriminator architectural design, and training objectives. For our generator architecture, however, we make two small, but significant changes. First, we compute an intermediate mel-spectrogram representation for the input audio with half the hop size used for the vocoder to double the sequence length and match full-bandwidth waveform output.

Second, we add a residual or skip connection between the input audio signal of low sampling rate and the full-bandwidth audio output, with a sinc interpolation block in between that performs 2x upsampling. The residual connection enables our BWE generator to more easily generate low-bandwidth content and allows our BWE generator to focus model capacity on generating high-frequency content. The discriminator also operates on the higher sample rate, full bandwidth audio [20], [21]. Note, during preliminary testing, we experimented not using the residual connection which did not work and applying a low-pass (LP) filter to the generated waveform, which slowed training and did not enhance performance.

D. MusicHiFi-M2S

For our mono-to-stereo (M2S) upmixer, we follow our unified generator architecture, discriminator architecture, and training recipe for a third time. To create a stereo effect from a mono audio signal, however, we leverage a mid-side encoding [28] to convert stereo left and right signals into a summation channel (mid-channel) and a difference channel (side-channel) [28]. We then train our M2S module to input the mel-spectrogram of the mid-channel M and output the side channel waveform S, where $M = \frac{L+R}{2}$ and $S = \frac{L-R}{2}$, L and R are the left and right stereo channel, respectively. Subsequently, we reconstruct left and right output channels via $L = M + S$ and $R = M - S$.

As a result of using mid-side encodings, our method is downmix-compatible in that we can take a mono channel, upsample it to stereo, downsample back to mono, and recover the original mono channel perfectly. This is not the case with alternative methods which typically degrade results after repeated applications. Furthermore, we can also add a control mechanism to adjust the spatialization width by controlling the energy ratio between the side and mid-channels. We can do this by normalizing the mid-channel and side-channel energies to 0 decibels (dB) and then adjusting the mid/side energy ratio via $\hat{S} \leftarrow \alpha S$, where $\alpha = 10^{\gamma/20}$ and γ is a scalar factor in decibels. When $\gamma > 0$, there will be more side energy and when $\gamma < 0$, there will be less side energy.

IV. EXPERIMENT AND RESULTS

A. Training details

We train all of our models using an internal dataset of 1800 hours of licensed instrumental music (stereo 44.1 kHz). For training, we randomly crop a sequence of 16,384 samples and then apply module-specific pre-processing. For our vocoder, we use channel averaging and downsampling to 22.05 kHz with STFT settings of a 1024-sample window, 256-sample hop size, and 128-band log-mel spectrogram. For our BWE module, we use the same pre-processing as the vocoder but halve the window and hop size. For our M2S module, we use channel averaging with identical STFT settings as the vocoder. We use the scalar weights $\lambda_{fm} = 1$ and $\lambda_{rc} = 360$ for the training objective in Eq. (1) and train all modules with a batch size of 45 for 500k steps, and select the optimal checkpoint via the minimum validation multi-resolution STFT distance (STFT-D). Model sizes per stage are approximately 46M params and 55 GFLOPS for one second of audio. All other parameters follow from the open-source implementations of the BigVGAN [15] generator and DAC [29] discriminator.

B. Baselines

For vocoding, we compare against BigVGAN [15] and HiFi-GAN [14] all trained on the same data and input features. Our retrained HiFi-GAN model has 14M params and take 52 GFLOPS for one second of audio. We also train a large HiFi-GAN-large model with 1024 input channels with 55M params, while taking 211 GFLOPS for one second of audio. For BWE, we compare against Aero [23], a recent state-of-the-art BWE method that uses a encoder-decoder architecture

TABLE I: Vocoder objective evaluation. Our vocoder module yields better results than baselines, but is mildy slower.

Dataset	Method	Mel-D↓	STFT-D↓	ViSQOL↑	RTF↑
DSD100	HiFi-GAN [14]	1.09	0.65	4.47	**3488**
	HiFi-GAN-large [14]	1.06	0.60	4.48	3409
	BigVGAN [15]	0.94	0.41	4.61	1818
	MusicHiFi-V	**0.87**	**0.33**	**4.67**	1786
FMA	HiFi-GAN [14]	1.09	0.64	4.52	**3703**
	HiFi-GAN-large [14]	1.04	0.57	4.56	3614
	BigVGAN [15]	0.94	0.41	4.62	1829
	MusicHiFi-V	**0.87**	**0.35**	**4.67**	1807

TABLE II: BWE objective evaluation for full-band audio. Low/high-band results are in parentheses. * AudioSR outputs have a high-frequency EQ boost that causes evaluation issues.

Dataset	Method	Mel-D↓	STFT-D↓	ViSQOL↑	RTF↑
DSD100	Aero [23]	**0.51** (0.05/1.16)	0.12 (0.02/0.54)	**4.18**	19
	AudioSR* [24]	1.23 (0.64/2.25)	0.51 (0.36/1.68)	3.54	4
	MusicHiFi-BWE	0.55 (0.08/1.18)	**0.11** (0.02/0.56)	4.14	**1639**
FMA	Aero [23]	**0.89** (0.07/1.60)	**0.24** (0.03/0.73)	**4.12**	19
	AudioSR [24]	1.68 (0.68/3.18)	0.68 (0.39/2.33)	3.25	4
	MusicHiFi-BWE	1.01 (0.09/1.76)	0.26 (0.02/0.79)	4.08	**1613**

with LSTM and temporal-based attention layers with 19M params and takes 85 GFLOPS. We also compare a pre-trained checkpoint of AudioSR [24] for BWE (no training code available). For M2S, we compare against a CPU-only open-source decorrelation method [33], [34], denoted as DSP, that divides the signal into transients, noise, and harmonics and decorrelates non-transient content.

C. Objective evaluation

For objective evaluation, we use 673 clips from FMA-small [36] and the test split of the accompaniment track from the DSD100 test dataset [37]. For both test datasets, every segment has a duration of 30 seconds. For objective evaluation metrics, we use a suite of four metrics including ViSQOL [38], mel distance (Mel-D) and STFT-D. ViSQOL is a perceptual quality metric that estimates a mean opinion score based on the spectral similarity to the ground truth. The Mel-D and STFT-D measure the spectral distances between the reconstructed and ground-truth audio under a mel and a linear frequency scale, respectively. We use the real time factor (RTF) metric or time processed over time elapsed on an A100 GPU to measure speed.

D. Objective evaluation result

Vocoder objective evaluation results are shown in Table I. We find our proposed method outperforms BigVGAN and HiFi-GAN on Mel-D, STFT-D and ViSQOL results while maintaining a lower RTF on both datasets. We also find the RTF of our vocoder method is lower compared to HiFi-GAN, but is still very fast and almost 2000x real-time on an A100.

BWE objective evaluation results are shown in Table II. We find both our proposed method and Aero yield comparable STFT-D, Mel-D, and ViSQOL results. When compared to AudioSR, we found that AudioSR is easily influenced by scale

TABLE III: Comparison of objective metrics for M2S system. Values in Mel-D and STFT-D include mid/side channels.

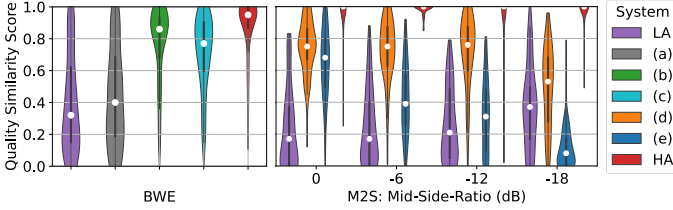| Dataset	Method	Mel-D↓	STFT-D↓	ViSQOL↑	RTF↑
DSD100-test	DSP [35]	1.07/1.87	1.09/1.70	4.69	5 (CPU)
	MusicHiFi-M2S	**0.00/1.70**	**0.00/1.53**	**4.73**	**1539** (GPU)
FMA-small	DSP [35]	0.99/2.29	1.08/2.16	4.70	4 (CPU)
	MusicHiFi-M2S	**0.00/2.03**	**0.00/1.88**	**4.73**	**1554** (GPU)

Fig. 2: Subjective listening test violins plots. BWE test (left) and M2S test (right). Test conditions include (a) AudioSR, (b) cascaded MusicHiFi-V and Aero and (c) MusicHiFi-V+BWE (d) MusicHiFi-V+BWE with DSP (e) full MusicHiFi.

variations and also has a notable presence of high frequency components. In an effort to address this issue, we computed a scale adjustment factor by downsampling the generated waveform to 22.05 kHz normalizing the energy to the ground truth. Despite these adjustments, a significant gap remains in the objective metrics remains, likely from difference in training datasets [24]. We also find the RTF of our BWE module is approx. 80-400x faster than alternatives.

M2S objective evaluation results are shown in Table III. We find our method outperforms the DSP decorrelation method on STFT-D, Mel-D, and ViSQOL. Furthermore, it is important to note that the error of the mid channel is zero for our method since our M2S method is downmix compatible and maintains the original mid channel and only estimates the side channel. We also find the RTF of our BWE module is over 300x faster than the DSP method by way of efficient GPU compute.

E. Subjective evaluation

We performed two subjective listening tests to evaluate our BWE and M2S independently [39]. For our BWE and M2S test, we recruited 20 and 23 participants with diverse audio backgrounds, respectively, and used a multiple stimuli with hidden reference and anchor (MUSHRA) protocol and the Web Audio Evaluation Tool [40]. The goal of our BWE tasks was to have participants rate *quality similarity* relative to ground-truth 44.1 kHz mono music. The goal of our M2S task was to rate *quality similarity* relative to ground-truth stereo as well as test spatial controllability by varying the target mid-side energy ratio (i.e., from 0 to -18dB) given that performance varies heavily on the spatialization level.

For our BWE listening test, we created six test examples from the FMA-small dataset, with each sample being 4 seconds long. Test conditions included (a) AudioSR, (b) cascaded MusicHiFi-V and Aero and (c) MusicHiFi-V+BWE as well as the LA and HA. The goal of this task is to understand the perceptual quality of comparable BWE algorithms. For the M2S evaluation, we prepared twelve listening samples, each 3

seconds long, selected from an internal test dataset instead of the FMA dataset as we found a large number of FMA clips were poorly spatialized. Test conditions included (d) cascaded MusicHiFi-V+BWE with DSP and (e) the full MusicHiFi for M2S comparison as well as the LA and HA. All samples for the two tasks are loudness normalized to -23 dBFS before and after being fed into the cascaded methods. For both tests, 22.05 kHz mono signals were used as the low-anchor (LA) and a hidden reference as high-anchor (HA).

F. Subjective evaluation results

The results of our listening tests are shown in Fig. 2. When we compare the BWE subjective evaluation results, samples from (a) AudioSR ranks the least compared to other baselines. This result matches our earlier qualitative analysis that AudioSR has a strong high-frequency boost. We also find that (b) MusicHiFi-V + Aero ranks slightly above our BWE method, but we believe this is reasonable considering Aero uses an U-Net architecture with internal BiLSTM layers versus our convolutional architecture that is dramatically faster. We further conducted multiple post-hoc paired t-tests with Bonferroni correction [41] for each condition vs. our method. We find there is no statistical significance between our method and Aero, while our method and Aero rank above AudioSR.

For M2S evaluation, we find that our (e) MusicHiFi performs best under different M/S panning coefficients and samples produced from the method (d) MusicHiFi-V+BWE with DSP perform similarly to ours when the energy ratio between mid and side channel is the same (0dB). The difference between our approach and the DSP baseline is statistically significant for side/mid rations 6, 12, and 18 via multiple post-hoc paired t-tests with Bonferroni correction [41].

For further evaluation, please find sound examples using mel-spectrograms extracted from real music and generated via a diffusion model [11] at https://MusicHiFi.github.io/web/.

V. CONCLUSION

We proposed a new efficient, high-fidelity stereophonic vocoding method named MusicHiFi. Our method works via a cascade of three GAN models that convert mel-spectrograms to low-quality audio waveforms, upsamples the low-resolution audio to high-resolution audio via bandwidth extension, and finally renders stereophonic high-resolution audio. Our method can be integrated into mel-spectrogram based music generators, used to enhance the fidelity of a low-resolution audio, and/or used to spatialize monophonic music. Compared to past work, we contribute a unified GAN-based discriminator and generator design, a new downsampling compatible BWE module, and a novel mono-preserving mono-to-stereo module. We evaluated our method using both objective evaluation and two subjective listening tests and found our method yields comparable or better vocoding and BWE results while outperforming comparable M2S methods, has better spatialization width control, and is extraordinarily efficient.

REFERENCES

[1] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Neural information processing systems (NeuraIPS)*, 2020.

[2] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," 2022. [Online]. Available: https://arxiv.org/abs/2204.06125

[3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[4] S. Forsgren and H. Martiros, "Riffusion - Stable diffusion for real-time music generation," 2022. [Online]. Available: https://riffusion.com/about

[5] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley, "AudioLDM: Text-to-audio generation with latent diffusion models," in *International Conference on Machine Learning (ICML)*, 2023.

[6] Q. Huang, D. S. Park, T. Wang, T. I. Denk, A. Ly, N. Chen, Z. Zhang, Z. Zhang, J. Yu, C. Frank *et al.*, "Noise2Music: Text-conditioned music generation with diffusion models," 2023. [Online]. Available: https://arxiv.org/abs/2302.03917

[7] C. Hawthorne, I. Simon, A. Roberts, N. Zeghidour, J. Gardner, E. Manilow, and J. Engel, "Multi-instrument music synthesis with spectrogram diffusion," in *International Society for Music Information Retrieval (ISMIR)*, 2022.

[8] K. Chen, Y. Wu, H. Liu, M. Nezhurina, T. Berg-Kirkpatrick, and S. Dubnov, "MusicLDM: Enhancing novelty in text-to-music generation using beat-synchronous mixup strategies," 2023. [Online]. Available: https://arxiv.org/abs/2308.01546

[9] S.-L. Wu, C. Donahue, S. Watanabe, and N. J. Bryan, "Music ControlNet: Multiple time-varying controls for music generation," 2023. [Online]. Available: https://arxiv.org/abs/2311.07069

[10] Y. Wang, Z. Ju, X. Tan, L. He, Z. Wu, J. Bian *et al.*, "AUDIT: Audio editing by following instructions with latent diffusion models," *Neural Information Processing Systems (NeurIPS)*, 2024.

[11] Z. Novack, J. McAuley, T. Berg-Kirkpatrick, and N. J. Bryan, "DITTO: diffusion inference-time T-optimization for music generation," 2024. [Online]. Available: https://arxiv.org/abs/2401.12179

[12] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "DiffWave: A versatile diffusion model for audio synthesis," in *International Conference on Learning Representations (ICLR)*, 2020.

[13] K. Kumar, R. Kumar, T. De Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. De Brebisson, Y. Bengio, and A. C. Courville, "MelGAN: Generative adversarial networks for conditional waveform synthesis," *Neural information processing systems (NeurIPS)*, 2019.

[14] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," *Neural Information Processing Systems (NeurIPS)*, 2020.

[15] S. gil Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S. Yoon, "BigV-GAN: A universal neural vocoder with large-scale training," in *International Conference on Learning Representations (ICLR)*, 2023.

[16] B. Iser, W. Minker, and G. Schmidt, *Bandwidth extension of speech signals*. Springer, 2008.

[17] J. Su, Y. Wang, A. Finkelstein, and Z. Jin, "Bandwidth extension is all you need," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[18] S. Han and J. Lee, "NU-Wave 2: A general neural audio upsampling model for various sampling rates," in *Interspeech*, 2022.

[19] K. Zhang, Y. Ren, C. Xu, and Z. Zhao, "WSRGlow: A glow-based waveform generative model for audio super-resolution," in *Interspeech*, 2021.

[20] S. E. Eskimez and K. Koishida, "Speech super resolution generative adversarial network," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[21] R. Kumar, K. Kumar, V. Anand, Y. Bengio, and A. Courville, "NU-GAN: High resolution neural upsampling with gan," 2020. [Online]. Available: https://arxiv.org/abs/2010.11362

[22] S. Hu, B. Zhang, B. Liang, E. Zhao, and S. Lui, "Phase-aware music super-resolution using generative adversarial networks," in *Interspeech*, 2020.

[23] M. Mandel, O. Tal, and Y. Adi, "AERO: Audio super resolution in the spectral domain," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[24] H. Liu, K. Chen, Q. Tian, W. Wang, and M. D. Plumbley, "AudioSR: Versatile audio super-resolution at scale," 2023. [Online]. Available: https://arxiv.org/abs/2309.07314

[25] J. Jot and C. Avendano, "Spatial enhancement of audio recordings," in *Audio Engineering Society Conference*. Audio Engineering Society, 2003.

[26] J. Serrà, D. Scaini, S. Pascual, D. Arteaga, J. Pons, J. Breebaart, and G. Cengarle, "Mono-to-stereo through parametric stereo generation," in *International Society of Music Information Retrieval (ISMIR)*, 2023.

[27] E. Schuijers, J. Breebaart, H. Purnhagen, and J. Engdegard, "Low complexity parametric stereo coding," in *Audio Engineering Society Convention*. Audio Engineering Society (AES), 2004.

[28] J. D. Johnston and A. J. Ferreira, "Sum-difference stereo transform coding," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1992.

[29] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, "High-fidelity audio compression with improved RVQGAN," in *Neural Information Processing Systems (NeurIPS)*, 2023.

[30] L. Ziyin, T. Hartwig, and M. Ueda, "Neural networks fail to learn periodic functions and how to fix it," *Neural Information Processing Systems (NeurIPS)*, 2020.

[31] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "Soundstream: An end-to-end neural audio codec," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 2021.

[32] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[33] M. R. Schroeder, "An artificial stereophonic effect obtained from a single audio signal," *Journal of the Audio Engineering Society (JAES)*, 1958.

[34] D. Fitzgerald, "Upmixing from mono-a source separation approach," in *IEEE International Conference on Digital Signal Processing (DSP)*, 2011.

[35] s3a spatialaudio, "s3a decorrelator," https://github.com/s3a-spatialaudio/s3a-decorrelation-toolbox, 2019.

[36] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, "FMA: A dataset for music analysis," in *International Society for Music Information Retrieval (ISMIR)*, 2017.

[37] A. Liutkus, F.-R. Stöter, Z. Rafii, D. Kitamura, B. Rivet, N. Ito, N. Ono, and J. Fontecave, "The 2016 signal separation evaluation campaign," in *International Conference Latent Variable Analysis and Signal Separation LVA/ICA*, P. Tichavský, M. Babaie-Zadeh, O. J. Michel, and N. Thirion-Moreau, Eds. Springer International Publishing, 2017.

[38] A. Hines, J. Skoglund, A. C. Kokaram, and N. Harte, "ViSQOL: an objective speech quality model," *EURASIP Journal on Audio, Speech, and Music Processing*, 2015.

[39] B. Series, "Recommendation ITU BS.1534-3," 2014.

[40] N. Jillings, D. Moffat, B. De Man, and J. D. Reiss, "Web Audio Evaluation Tool: A browser-based listening test environment," in *Sound and Music Computing Conference*, 2015.

[41] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian journal of statistics*, 1979.