MUSIC ENHANCEMENT VIA IMAGE TRANSLATION AND VOCODING

Nikhil Kandpal[*]

University of North Carolina at Chapel Hill
Computer Science Department
Chapel Hill, NC, USA

Oriol Nieto, Zeyu Jin

Adobe Research
San Francisco, CA, USA

ABSTRACT

Consumer-grade music recordings such as those captured by mobile devices typically contain distortions in the form of background noise, reverb, and microphone-induced EQ. This paper presents a deep learning approach to enhance low-quality music recordings by combining (i) an image-to-image translation model for manipulating audio in its mel-spectrogram representation and (ii) a music vocoding model for mapping synthetically generated mel-spectrograms to perceptually realistic waveforms. We find that this approach to music enhancement outperforms baselines which use classical methods for mel-spectrogram inversion and an end-to-end approach directly mapping noisy waveforms to clean waveforms. Additionally, in evaluating the proposed method with a listening test, we analyze the reliability of common audio enhancement evaluation metrics when used in the music domain.

***Index Terms*—** Music Enhancement, Image-to-Image Translation, Diffusion Probabilistic Models, Vocoding

1. INTRODUCTION

With the rise of Internet influencers and music hobbyists, a large portion of music content is created with cheap and accessible recording devices in non-treated environments. While being audible, these recordings often have degraded quality stemming from background noise, unpleasant reverb, and resonance caused by the microphone and the environment. This prompts us to investigate quality enhancement for music signals, transforming low-quality amateur recordings into professional ones.

The main difficulty of such an endeavor is that so many aspects of the low-quality recording setup are unknown. Parameters of the recording device, such as frequency response characteristics, vary drastically across different hardware. Additionally, acoustic properties such as the size, shape, and reflectivity of the recording environment vary between different recording setups. Finally, background noise is hard to capture and generalize, especially non-stationary noise. A solution that faithfully transforms a low-quality recording into what it would sound like recorded professionally must implicitly or explicitly infer all of these aspects from the signal alone. In speech enhancement, end-to-end methods such as HiFi-GAN [1] and Demucs [2] achieve this by extracting the speech source from a mixture of sources. However, music signals are often polyphonic, i.e., there can be an arbitrary number of sources to be extracted at once. Moreover, the perception of music quality typically differs from that of speech. For example, human listeners may find reverb pleasant in music, while it is usually undesired in speech. Therefore, we aim to develop a solution that works for polyphonic signal enhancement and reflects the unique qualities of music perception.

Our approach performs enhancement on the recording's mel-spectrogram representation. This is achieved by treating the mel-spectrogram as an image and training an image-to-image translation model similar to Pix2Pix [3] to transform a low-quality mel-spectrogram into that of a high-quality signal. We hypothesize that it is easier to enhance polyphonic signals in the mel-spectrogram domain as polyphonic sources are additive and have a very small temporal span compared to waveforms. Finally, to map generated high-quality mel-spectrograms to perceptually realistic waveforms, we train a vocoding model based on DiffWave [4]. Training this model on only the high quality samples of music performance makes it robust to the artifacts that reside in the synthetic mel-spectrogram.

We evaluate our approach by performing a listening test with 211 participants, and we show that this approach achieves a much better perceptual enhancement than several state-of-the-art techniques. We also compare the subjective listening test scores with widely used audio quality metrics and suggest that, similar to speech enhancement, these metrics correlate poorly with human perception [1, 5]. With this work, we hope to motivate both future research in music enhancement as well as music quality perceptual metrics akin to those in the speech literature [6, 7]. To promote further research, audio samples generated in our experiments and source code are provided at our project website[1].

In this paper, we refer to Pix2Pix models operating on mel-spectrograms as *Mel2Mel* models and vocoding applied to the music domain as *musecoding*. We summarize our contributions as follows:

- A music enhancement model leveraging recent work on conditional image synthesis and vocoding.
- A generative process for simulating realistic low-quality music recordings from professional-quality recordings.
- An analysis of the reliability of common audio enhancement evaluation metrics in the music domain.

2. RELATED WORK

To our knowledge there is little prior work studying music quality enhancement. The work most similar to our contributions focuses on speech enhancement, conditional speech synthesis, or music source separation.

Early approaches to speech enhancement have used classical signal processing techniques such as Wiener filtering [8] and non-negative matrix factorization [9]. More recently, deep learning-based methods have achieved state-of-the-art on speech enhancement. These methods either manipulate the audio in its magnitude

[*] Work done during an internship with Adobe Research

[1] https://nkandpa2.github.io/music-enhancement

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

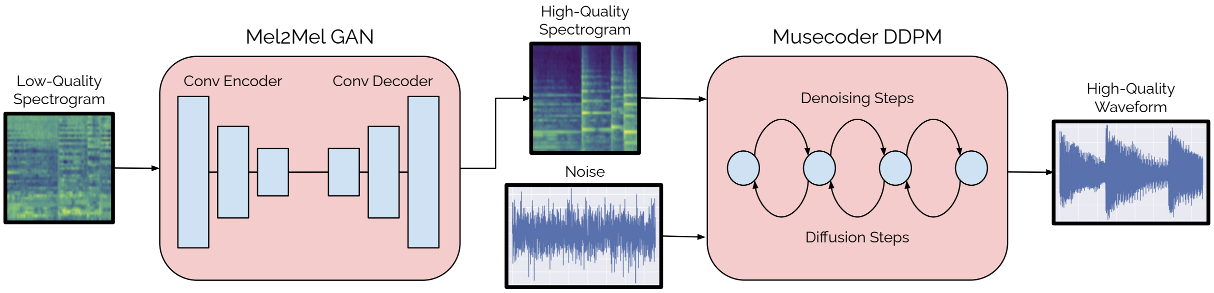

Fig. 1. Model architecture of our Mel2Mel + Diffwave model. First, a low-quality mel-spectrogram is enhanced by a conditional GAN. The resulting synthetic mel-spectrogram is then "musecoded" into a waveform by a Denoising Diffusion Probabilistic Model (DDPM).

spectrogram representation (followed by a spectrogram inversion method to recreate the corresponding waveform) [10, 11, 12] or map directly from the low-quality waveform to a cleaned waveform [13, 2, 1]. Methods that operate on the time-frequency domain generally produce audible artifacts due to the use of phase reconstruction algorithms like the Griffin-Lim algorithm [14]. A recent work addresses this with neural-network based vocoders [15], yet its quality is not on par with an end-to-end approach [16]. Alternatively, methods that work on the time domain typically require more training steps [1].

Conditional speech synthesis techniques produce speech waveforms from conditioning information such as magnitude spectrograms, a problem commonly known as vocoding. Some state-of-the-art vocoding methods involve using generative adversarial networks [17, 18] or denoising diffusion probabilistic models [4, 19] for generating audio.

Music source separation focuses on taking a mix of multiple music "stems" (vocals, drums, etc.) and separating the mix into its individual sources. Some approaches to music source separation operate by masking spectrograms [20] or directly mapping the mix waveform to individual source waveforms [21, 22]. The music enhancement problem is different than music source separation, since our goal is not only to extract all musical sources from a noisy mixture but also to reduce reverb and adjust EQ such that the listening experience is improved.

3. METHODS

3.1. Modeling Approach

In this paper, we investigate the approach of enhancing music in its mel-spectrogram domain, as it is easier to represent complex harmonic structures and polyphonic sound sources. We then transform the resulting mel-spectrograms to waveforms through a Diffwave-based vocoder (a process that in this context could be more aptly named "musecoding"). Decoupling waveform generation from mel-spectrogram enhancement allows us to train a musecoder that is not only robust to noise and other artifacts, but can also be used for any generation and enhancement task without the need of retraining. Figure 1 depicts a block diagram of our proposed architecture. This approach is motivated by recent advances in vocoding that generate natural-sounding speech from mel-spectrograms [4].

3.2. Data Simulation

The modeling techniques we consider in this paper require aligned pairs of high- and low-quality music recordings. To construct such a dataset, we assume access to high-quality recordings and define a generative process for simulating low-quality ones. First, we simulate the reverb and varied microphone placements of a non-professional recording environment by convolving the high-quality music signal with a room impulse response. Next, we apply additive background noise scaled to achieve a randomly sampled SNR between 5 and 30 dB. Finally, we simulate a low-quality microphone frequency response by applying 4-band equalization with randomly sampled gains between -15 and 15 dB and frequency bands from 0-200, 200-1000, 1000-4000, and 4000-8000 Hz.

3.3. Mel-Spectrogram Enhancement with Mel2Mel

Our first step in music enhancement is modeling the distribution of high-quality mel-spectrograms conditioned on their low-quality counterparts. To estimate this distribution, we use existing work on image-to-image translation with conditional adversarial networks [3] in an approach similar to [12].

In this framework a generator and a discriminator are trained using an aligned dataset of low and high-quality recording pairs. The generator maps from low to high-quality mel-spectrograms with the objective of maximizing the discriminator's loss and minimizing the ℓ_1 distance between the generated mel-spectrogram and the ground truth high-quality mel-spectrogram. The discriminator is trained to classify whether a given mel-spectrogram is generated or comes from the true data distribution. It performs this classification on a patch-wise basis, predicting a class for each patch in the input mel-spectrogram. For this reason, the discriminator acts as a learned loss function for the generator which enforces realistic local features and the ℓ_1 loss enforces global consistency with the ground truth mel-spectrogram.

3.4. Musecoding

Recent work has shown that deep learning models can generate perceptually realistic waveforms from speech mel-spectrograms. In our experiments, we evaluate the Diffwave [4] vocoder applied to music, a process that we call "musecoding".

Diffwave is a denoising diffusion probabilistic model (DDPM). This class of models defines a forward diffusion process which iteratively adds gaussian noise to audio waveforms from the training dataset. A model is then trained to estimate the reverse transition distributions of each noising step conditioned on the mel-spectrogram of the clean audio. Sampling from this model requires sampling noise from a standard gaussian and iteratively denoising using the reverse transition probability distributions from the model. For further discussion of DDPMs see [4] and [23].

Model	MOS ↑
Clean	4.39 ± 0.05
Mel2Mel + Diffwave	$\mathbf{4.06 \pm 0.06}$
Mel2Mel + Griffin-Lim	3.01 ± 0.09
No Enhancement	2.85 ± 0.09

Table 1. Mean Opinion Scores in a human listening test.

As a musecoding baseline, we also consider mel-spectrogram inversion with inverse mel-scaling and the Griffin-Lim algorithm [14].

4. EXPERIMENT SETUP

4.1. Dataset

We train and evaluate models on the Medley-solos-DB dataset [24], containing 21,572 three-second, single-instrument samples recorded in professional studios. We exclude the distorted electric guitar samples to avoid fitting our models to production effects. We use 5841 samples for training, 3494 for validation and the rest for testing. We start by downsampling our data to 16 kHz following the setup of prior vocoding work [4, 17]. This sample rate has shown to be favored by most speech enhancement work [1, 2] and can be potentially super-resolved to 48 kHz with bandwidth extension techniques [5]. Using the procedure described in Section 3.2, we generate a dataset of high- and low-quality recording pairs. For simulation of low-quality recordings, we source room impulse responses from the DNS Challenge dataset [25] and realistic background noise from the ACE Challenge dataset [26]. As a final step, we apply a low-cut filter to remove nearly inaudible low frequencies below 35 Hz and normalize the waveforms to have a maximum absolute value of 0.95. We find that this treatment helps improve our models' training stability. When evaluating, we apply the same treatment (low-cut filter at 35 Hz and normalization) before applying our enhancement models.

4.2. Model Architectures and Hyperparameters

In all experiments, we compute mel-spectrograms with 128 mel bins, an FFT size of 1024, and a 256 sample hop length. When training models that generate or are conditioned on mel-spectrograms, we use log-scale amplitudes to reduce the range of values and to avoid positive restrictions on our models' domain or range.

The Mel2Mel generator described in Section 3.3 consists of 2 downsampling blocks, each containing a 2D convolutional kernel of size 3 and stride 2, instance normalization [27] and ReLU activation functions. This is followed by 3 ResNet blocks [28] with kernel size 3 and instance normalization. Finally, the representation is upsampled back to the original dimensionality of the input with two upsampling blocks, each containing a transposed convolutional kernel of size 3 and stride 2, instance normalization, and ReLU activation functions. The Mel2Mel discriminator is a fully convolutional model made up of three blocks, each containing a convolutional kernel of size 4 and stride 2, instance normalization, and LeakyReLU activation function. The last layer does not have any normalization or activation function. Both the generator and discriminator are trained with batch size of 64 and learning rate of 0.0002 for 200 epochs.

The Diffwave model described in Section 3.4 uses the architecture and training objective described in [4]. We train this model for 3000 epochs using a batch size of 8 and a learning rate of 0.0002.

4.3. Baselines

We evaluate our approach against two separate baselines. First, we pair Mel2Mel for mel-spectrogram enhancement with inverse mel-scaling and the Griffin-Lim algorithm for musecoding. Both inverse mel-scaling and Griffin-Lim require solving optimization problems [29], so we run both solvers for 100 iterations, which yields a per-sample runtime comparable to that of the Diffwave musecoder.

Our second baseline is an end-to-end approach for music enhancement. Namely, we use the Demucs model architecture [21] and train it using the ℓ_1 reconstruction loss on our dataset of low- and high-quality recording pairs. This matches the original training objective used for this architecture on the task of music source separation. We train this model for 360 epochs with batch size 64 and learning rate 0.0003. We find that after this number of epochs the validation loss plateaus.

4.4. Evaluation Metrics

To evaluate the results of different enhancement models we conducted a Mean Opinion Score (MOS) test with human listeners on Amazon Mechanical Turk (AMT). Additionally, we evaluate enhancement methods by computing the frequency-weighted segmental SNR (fwSSNR) [30], multi-resolution spectrogram loss (MRS) [31], ℓ_1 spectrogram distance, and Fréchet Audio Distance (FAD) [32] between enhanced and clean reference signals. In Section 5.3 we analyze the effectiveness of these objective metrics at approximating human listener ratings in the music domain.

5. RESULTS

5.1. Mean Opinion Score Test

To evaluate our proposed Mel2Mel + Diffwave music enhancement model, we conducted an MOS test with human listeners on AMT. We used 200 audio samples from our test set, added 8 different types of simulated degradation, and passed these low-quality waveforms through our method, Mel2Mel + Griffin-Lim, and Demucs. The low-quality, enhanced, and ground truth high-quality samples were then presented to human listeners who were asked to give a quality score from 1 to 5. We used the ground truth high-quality recording as high anchor and the same recording with 0 dB white noise as low anchor.

Each Human Intelligence Task (HIT) started with a screening test in which human listeners were required to identify which one of 5 audio samples sound the same as a reference sample. 4 out of the 5 samples are passed a small amount of effects including low pass filters, high pass filters, comb filters, and added noise. Passing the screening test was required to continue. The rest of the HIT consisted of 34 tests in which 4 were validation tests to check if listeners were paying attention. If they failed the validation test, the entire HIT was invalidated.

In the end we collected 9,095 answers from 211 listeners. The results shown in Table 1 suggests that Mel2Mel with a Diffwave musecoder achieves the highest MOS with a score near that of clean audio from the dataset.

5.2. Perturbation Ablation Study

To gain insight into which perturbations are handled most effectively by each enhancement model, we perform an ablation study isolating each perturbation introduced in the low-quality signal generative process. Table 2 contains mean opinion scores for each enhancement

Model	Random EQ	SNR 5	SNR 10	SNR 15	DRR 0	DRR 3	DRR 6
Clean	4.35 ± 0.06	4.24 ± 0.07	4.27 ± 0.06	4.46 ± 0.06	4.28 ± 0.04	4.19 ± 0.06	4.42 ± 0.06
Mel2Mel + Diffwave	$\mathbf{4.15 \pm 0.07}$	$\mathbf{4.01 \pm 0.08}$	$\mathbf{4.24 \pm 0.06}$	$\mathbf{3.96 \pm 0.09}$	3.77 ± 0.06	3.84 ± 0.06	3.96 ± 0.08
Mel2Mel + Griffin-Lim	2.98 ± 0.1	3.10 ± 0.08	3.53 ± 0.09	3.18 ± 0.11	2.82 ± 0.07	2.77 ± 0.09	2.99 ± 0.10
Demucs	3.39 ± 0.10	2.55 ± 0.10	3.07 ± 0.1	2.85 ± 0.11	3.13 ± 0.07	3.21 ± 0.07	3.30 ± 0.10
No Enhancement	3.99 ± 0.08	2.48 ± 0.11	2.71 ± 0.1	3.04 ± 0.12	$\mathbf{4.01 \pm 0.06}$	$\mathbf{3.91 \pm 0.07}$	$\mathbf{4.21 \pm 0.07}$

Table 2. Mean Opinion Scores in a human listening test. Each column contains the ratings for a single perturbation type: EQ, additive background noise at different signal-to-noise ratios (SNR), and reverb at different direct-to-reverberant ratios (DRR).

Enhancement Metric	Rank Correlation with MOS
fwSSNR	0.5
$-$MRS	0.56
$-$L1	0.4
$-$FAD	0.53

Table 3. Spearman rank correlation between MOS test ratings and audio enhancement metrics.

Model	fwSSNR ↑	MRS ↓	L1 ↓	FAD ↓
Independent Training	**9.04**	**1.40**	**1.50**	4.73
Joint Fine-tuning	7.61	1.57	1.57	4.54
Joint Training	6.58	1.65	1.69	**3.98**
Sequential Training	8.23	1.80	1.83	5.54
No Enhancement	6.96	1.89	2.16	5.90

Table 4. Performance of Mel2Mel + Diffwave enhancement models using different training schemes

model applied to signals with randomly sampled EQ, additive noise with signal-to-noise ratios (SNR) of 5, 10, and 15 dB, and reverb with direct-to-reverberant ratios (DRR) of 0, 3, and 6 dB.

This ablation shows that the Mel2Mel + Diffwave model excels at removing noise even at SNR values as low as 5 dB and at undoing 4-band equalization simulating a non-flat microphone frequency response. Interestingly, none of the models tested perform dereverberation very well, and in fact degrade signals that contain no noise and only simulated reverb. This may be due to train-test mismatch, since all samples enhanced during training time contained some level of additive noise.

This ablation also lends insight into the types of perturbations that affect human listeners' perception of music. From the difference between the scores given to clean samples and non-enhanced samples, it is clear that additive noise impacts the listener's perception significantly while reverb is mostly ignored.

5.3. Perceptual Alignment of Objective Metrics

The results of the MOS test also provide a mechanism to evaluate how well objective metrics for audio quality align with human perception in the music domain. We measure fwSSNR, MRS, FAD, and ℓ_1 spectrogram distance on the same samples submitted for MOS evaluation. We then take the mean score across all samples with a given perturbation type (i.e. SNR 5, DRR 0, etc.) and perform a Spearman rank correlation with the mean scores measured in the human MOS test. In Table 3 we show the rank correlation for each objective metric. We find that none of the four metrics evaluated correlate very strongly with human opinion scores, the highest achieving a rank correlation of 0.56.

We also identify particular failure modes of these metrics. All four metrics fail to identify robotic artifacts induced by the Griffin-Lim algorithm and actually rate the Mel2Mel + Griffin-Lim model as the best of all models we tested. Additionally, fwSSNR, MRS, and ℓ_1 spectrogram distance all fail to identify additive noise effectively, and rate non-enhanced samples at SNR values of 10 and 15 dB as being better than any enhancement model output. FAD does not have this failure mode.

5.4. Alternate Training Schemes

In Section 3.1 we motivated approaching music enhancement by training two decoupled models that separately handle mel-spectrogram enhancement and musecoding. Here, we investigate training schemes for these models other than independently training them on their respective tasks. In addition to independent training, we (1) finetune the Mel2Mel generator and Diffwave musecoder jointly using the Diffwave objective, (2) train the models sequentially by first training the musecoder and then training the Mel2Mel generator with musecoder parameters frozen, and (3) train the Mel2Mel generator and musecoder jointly as a single model using the Diffwave objective.

Table 4 shows the performance of the resulting models. In Section 5.3 we discussed the reliability of using these metrics for evaluating algorithms, and find that FAD is the most perceptually aligned metric when it comes to denoising. Given this observation, our results suggest that joint training may yield better denoising performance than independent training. Joint training has the added benefit that only a single model is trained using a non-adversarial objective. However, this comes with the downside that the trained model cannot be split into enhancement and musecoding sub-models. Future work could focus on further exploring such training schemes.

6. CONCLUSION

We propose a music enhancement model that decomposes the task into mel-spectrogram enhancement and waveform synthesis from mel-spectrograms. This model was trained using high-quality samples from a public dataset paired with low-quality samples generated by simulating artifacts that typically appear in amateur recordings. A human MOS test shows that this model outperforms state-of-the-art baselines. Additionally, we found that current objective metrics for audio enhancement do not accurately reflect human perception of music. We hope this work encourages researchers to further advance the rather unexplored and yet timely topic of automatic music enhancement, either by designing more performant models or by proposing metrics that better align with human music perception.

7. REFERENCES

[1] Jiaqi Su, Zeyu Jin, and Adam Finkelstein, “Hifi-gan: High-fidelity denoising and dereverberation based on speech deep features in adversarial networks,” 2020.

[2] Alexandre Defossez, Gabriel Synnaeve, and Yossi Adi, “Real time speech enhancement in the waveform domain,” *Interspeech2020*, 2020.

[3] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros, “Image-to-image translation with conditional adversarial networks,” 2018.

[4] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro, “Diffwave: A versatile diffusion model for audio synthesis,” 2021.

[5] Jiaqi Su, Yunyun Wang, Adam Finkelstein, and Zeyu Jin, “Bandwidth extension is all you need,” in *ICASSP 2021-2021*. IEEE, 2021, pp. 696–700.

[6] Pranay Manocha, Zeyu Jin, Richard Zhang, and Adam Finkelstein, “Cdpam: Contrastive learning for perceptual audio similarity,” in *ICASSP 2021-2021*. IEEE, 2021, pp. 196–200.

[7] 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, 2021, pp. 6493–6497.

[8] P. Scalart and J.V. Filho, “Speech enhancement based on a priori signal to noise estimation,” in *1996 IEEE ICASSP Proceedings*, 1996, vol. 2, pp. 629–632 vol. 2.

[9] Hideaki Kagami, Hirokazu Kameoka, and Masahiro Yukawa, “Joint separation and dereverberation of reverberant mixtures with determined multichannel non-negative matrix factorization,” in *2018 IEEE ICASSP*, 2018, pp. 31–35.

[10] Kun Han, Yuxuan Wang, DeLiang Wang, William S. Woods, Ivo Merks, and Tao Zhang, “Learning spectral mapping for speech dereverberation and denoising,” *IEEE/ACM TASLP*, vol. 23, no. 6, pp. 982–992, 2015.

[11] Donald S. Williamson and DeLiang Wang, “Speech dereverberation and denoising using complex ratio masks,” in *2017 IEEE ICASSP*, 2017, pp. 5590–5594.

[12] Daniel Michelsanti and Zheng-Hua Tan, “Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification,” *Interspeech 2017*, Aug 2017.

[13] Santiago Pascual, Joan Serrà, and Antonio Bonafonte, “Towards generalized speech enhancement with generative adversarial networks,” 2019.

[14] D. Griffin and Jae Lim, “Signal estimation from modified short-time fourier transform,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[15] Adam Polyak, Lior Wolf, Yossi Adi, Ori Kabeli, and Yaniv Taigman, “High fidelity speech regeneration with application to speech enhancement,” in *ICASSP 2021-2021*. IEEE, 2021, pp. 7143–7147.

[16] Jiaqi Su, Zeyu Jin, and Adam Finkelstein, “Hifi-gan-2: studio-quality speech enhancement via generative adversarial networks conditioned on acoustic features,” in *2015 IEEE WASPAA*, 2021.

[17] Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brebisson, Yoshua Bengio, and Aaron Courville, “Melgan: Generative adversarial networks for conditional waveform synthesis,” 2019.

[18] Jaeseong You, Dalhyun Kim, Gyuhyeon Nam, Geumbyeol Hwang, and Gyeongsu Chae, “Gan vocoder: Multi-resolution discriminator is all you need,” 2021.

[19] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J. Weiss, Mohammad Norouzi, and William Chan, “Wavegrad: Estimating gradients for waveform generation,” 2020.

[20] Romain Hennequin, Anis Khlif, Felix Voituret, and Manuel Moussallam, “Spleeter: a fast and efficient music source separation tool with pre-trained models,” *Journal of Open Source Software*, vol. 5, pp. 2154, 06 2020.

[21] Alexandre Défossez, Nicolas Usunier, Léon Bottou, and Francis Bach, “Music source separation in the waveform domain,” 2021.

[22] Yi Luo and Nima Mesgarani, “Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation,” *IEEE/ACM TASLP*, vol. 27, no. 8, pp. 1256–1266, Aug 2019.

[23] Jonathan Ho, Ajay Jain, and Pieter Abbeel, “Denoising diffusion probabilistic models,” 2020.

[24] Vincent Lostanlen and Carmine-Emanuele Cella, “Deep convolutional networks on the pitch spiral for musical instrument recognition,” 2017.

[25] Chandan K A Reddy, Harishchandra Dubey, Kazuhito Koishida, Arun Nair, Vishak Gopal, Ross Cutler, Sebastian Braun, Hannes Gamper, Robert Aichner, and Sriram Srinivasan, “Interspeech 2021 deep noise suppression challenge,” 2021.

[26] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, “The ace challenge — corpus description and performance evaluation,” in *2015 IEEE WASPAA*, 2015, pp. 1–5.

[27] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky, “Instance normalization: The missing ingredient for fast stylization,” 2017.

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” in *2016 IEEE CVPR*, 2016, pp. 770–778.

[29] Yao-Yuan Yang, Moto Hira, Zhaoheng Ni, Anjali Chourdia, Artyom Astafurov, Caroline Chen, Ching-Feng Yeh, Christian Puhrsch, David Pollack, Dmitriy Genzel, Donny Greenberg, Edward Z. Yang, Jason Lian, Jay Mahadeokar, Jeff Hwang, Ji Chen, Peter Goldsborough, Prabhat Roy, Sean Narenthiran, Shinji Watanabe, Soumith Chintala, Vincent Quenneville-Bélair, and Yangyang Shi, “Torchaudio: Building blocks for audio and speech processing,” *arXiv preprint arXiv:2110.15018*, 2021.

[30] Y. Hu and Philipos C. Loizou, “Evaluation of objective quality measures for speech enhancement,” *IEEE TASLP*, vol. 16, pp. 229–238, 2008.

[31] Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim, “Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram,” 2020.

[32] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi, “Fréchet audio distance: A metric for evaluating music enhancement algorithms,” 2019.