Enhancing GAN-based Vocoders with Contrastive Learning
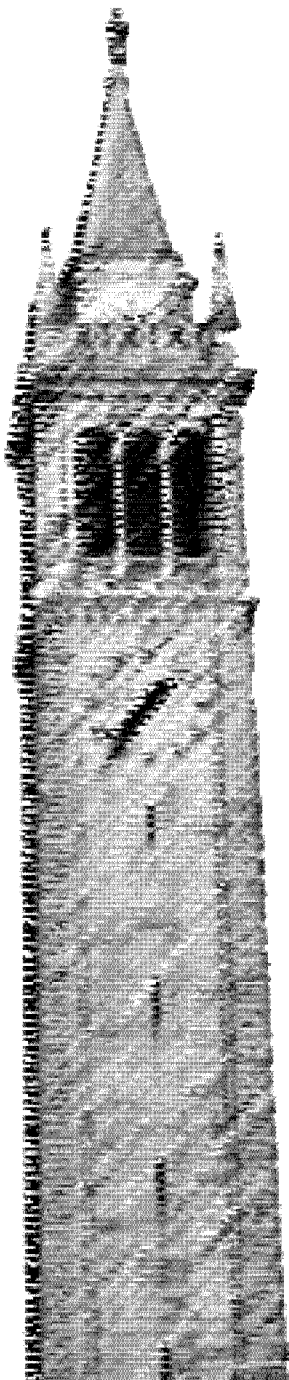
Haoming Guo
Gerald Friedland
Gopala Krishna Anumanchipalli

Electrical Engineering and Computer Sciences
University of California, Berkeley

May 19, 2023

Acknowledgement

Enhancing GAN-based Vocoders with Contrastive Learning

by

Haoming Guo

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Master of Science

in

Electrical Engineering and Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Gerald Friedland, Chair
Professor Gopala Anumanchipalli

Spring 2023

Enhancing GAN-based Vocoders with Contrastive Learning

by Haoming Guo

Research Project

Submitted to the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, in partial satisfaction of the requirements for the degree of **Master of Science, Plan II.**

Approval for the Report and Comprehensive Examination:

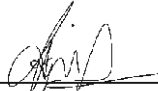**Committee:**

_____

Professor Gerald Friedland
Research Advisor

_____
May 19th, 2023

(Date)

* * * * * * *

_____

Professor Gopala Anumanchipalli
Second Reader

_____
May 19th, 2023

(Date)

Enhancing GAN-based Vocoders with Contrastive Learning

Abstract

Enhancing GAN-based Vocoders with Contrastive Learning

by

Haoming Guo

Master of Science in Electrical Engineering and Computer Science

University of California, Berkeley

Professor Gerald Friedland, Chair

Vocoder models have recently achieved substantial progress in generating authentic audio comparable to human quality while significantly reducing memory requirement and inference time. However, these data-hungry generative models require large-scale audio data for learning good representations. In this paper, we apply contrastive learning methods in training the vocoder to improve the perceptual quality of the vocoder without modifying its architecture or adding more data. We design an auxiliary task with mel-spectrogram contrastive learning to enhance the utterance-level quality of the vocoder model in data-limited conditions. We also extend the task to include waveforms to improve the multi-modality comprehension of the model and address the discriminator overfitting problem. We optimize the additional task simultaneously with GAN training objectives. Our result shows that the tasks improve model performance substantially in data-limited settings. Our analysis based on the result indicates that the proposed design successfully alleviate discriminator overfitting and produce audio of higher fidelity.

Contents

List of Figures

List of Tables

Acknowledgments

I would like to express my deepest gratitude to my supervisor, Professor Gerald Friedland, who started advising me when I was an undergraduate. I deeply appreciate his support and commitment as my research advisor. I am also grateful to Professor Gopala Anumanchipalli for taking precious time out of his busy schedule to be the second reader for my thesis. I am truly fortunate to have their mentorship that instrumentally shaped this thesis and me as a researcher.

I would like to thank my collaborator, Zhihao Zhao, who made significant contributions to this project in both ideas and experiments. This research would not have been possible without him. I would also like to express my appreciation to Jiachen Lian, who provided valuable advice to this work to reach its current quality.

Finally, I would like to thank my parents for their unwavering support and encouragement. I would also like to thank all my friends, family, and everyone who has supported me throughout my undergraduate and graduate education. Their support played a crucial role in my personal and academic growth during my five years UC Berkeley.

Chapter 1

Introduction

1.1 Background

Recently, neural speech synthesis has received increasing attention in the machine learning research community. It has achieved superior performance in the last ten years over traditional methods such as Unit Selection and Statistical Parametric Synthesis. Neural speech synthesis has two main approaches, the end-to-end approach and multi-phase approach. The end-to-end approach captures complex linguistic and acoustic patterns directly from text to generate speech waveforms. The multi-phase approach first extracts linguistic features from text, then converts the linguistic features to acoustic features like mel-spectrograms, and finally generate speech waveforms from the acoustic features. The tools used for the last step are called vocoders, and they have been widely used in the field of speech processing, music production, and voice communication since the 20th century. Compared to the end-to-end approach which has been rapid growing recently, the multi-phase approach with vocoders allows for more interpretability, control over synthesis parameters, and less data to reach high fidelity.

Recent research in vocoders has focused on improving the quality of synthesized speech and music using deep learning techniques. Since mel-spectrograms is similar to image in terms of data shape, many successful methods used in computer vision have been applied in vocoders to process the input mel-spectrograms. One of these methods is Generative Adversarial Networks (GANs) [9], which has been widely used in different domains for generation tasks. A GAN consists of a generator which handles the generation task, and a discriminator which is trained to combat the generator by distinguishing the generated samples from the real samples. As the discriminator improves, it challenges the generator to produce more authentic outputs, thereby enhancing the capability of the generator.

Vocoders based on GAN have been widely used and have achieved the state-of-the-art in the domain [31, 28, 33]. We survey these works in detail in section 2.2.

1.2   Motivations

Training GAN vocoders still meet two challenges, data insufficiency and discriminator overfitting.

In the realm of single-speaker speech synthesis, the limited size of available datasets poses a significant challenge. To enhance the performance of vocoders operating under such constraints, we propose the use of unsupervised learning techniques to extract additional self-supervised signals for training. Drawing on the exceptional transfer learning capabilities of self-supervised learning, we seek to harness this power in the realm of Vocoder modeling, focusing specifically on the application of contrastive learning. Although contrastive learning has been explored in the context of speech recognition [45], we are unaware of any previous efforts to apply this approach to Vocoder modeling. In this work, our aim is to leverage contrastive learning as an auxiliary task to enhance the vocoding performance of GAN generators under data-limited conditions.

The second challenge, discriminator overfitting, is also shown to be crucial especially on small dataset [66, 57, 1, 25], and the convergence of GAN also depends on the quality of discriminators [51]. Contrastive learning on the discriminator has been proved to alleviate this problem in image generation [21], and the method in general is also shown to increase model robustness on vision tasks [61, 13, 8, 67]. However, in speech synthesis, a naive approach of mel-spectrogram contrastive learning will only involve the generator, which encodes mel-spectrograms, but not the discriminator, which encodes the waveform. Therefore, we propose to extend the training to the discriminator by using a multi-modal contrastive task between mel-spectrograms and waveforms. We show later that the method yields a more robust discriminator and consequently a better generator.

1.3   Research Objectives

Our research objectives can be summarized as the following.

1. We aim to design a contrastive learning task for better training on mel-spectrograms to improve the performance of GAN-based vocoders on limited data. The goal of the task is to extract additional signals from existing data to boost performance in data-limited conditions.

2. We want to design an alternative novel contrastive learning task of matching mel-spectrogram to waveforms to enhance the understanding between the modalities, regularize the discriminator and improve perceptual quality of the generator. This task should also alleviate the discriminator overfitting problem, in addition to extracting more supervising signals like the first task.

3. We aim to implement a framework of integrating the contrastive learning into the GAN training pipeline. This step ensures the proposed tasks to work well with the original training objectives of the GAN-based vocoders.

4. We provide experimental results and in-depth analysis of the methods' effectiveness compared to the baseline. This should be comprehensive and cover experiments on multiple model model architecture, multiple data constraints, and multiple evaluation metrics to validate the effectiveness of the proposed contrastive learning tasks.

Chapter 2

Related Work

2.1   A Brief History of Vocoders

The term "vocoder" is a combination of the words "voice" and "encoder". Homer Dudley at Bell Labs developed the first vocoders in the 1930s [40], which was designed for military use to transmit speech over long distances using fewer transmission channels. In the history of TTS, vocoders first became significant when statistic parametric speech synthesis are proposed [65, 55, 56]. Statistic parametric speech synthesis uses a multi-phase approach consisted of a text analysis model that generates linguistic features, an acoustic model that generates acoustic features, and a heuristics-based vocoder that recovers speech waveforms [17, 18, 26].

Neural vocoders have recently replaced traditional heuristic methods and become the state-of-the-art vocoding method since Tacotron 2 [52] used WaveNet [46] as a vocoder to to convert mel-spectrograms to audio waveforms. WaveNet [46] is an autoregressive model that uses layers of dilated causal convolution with gated activation units. It is the fundamental work of subsequent autoregressive models. WaveRNN [23] and LPCNet [58] both use recurrent neural networks [22] to generate natural speech in an autoregressive way more efficiently, and they still remain widely-used today for efficient architecture. Different from autoregressive models, flow-based models, such as Parallel Wavenet [43] and WaveGlow [48], use normalizing flow techniques [7] to directly approximate the data distribution with feed-forward layers. Approaches based on Denoising Diffusion Probabilistic Models (DDPMs) [14], including WaveGrad [4], DiffWave[29], PriorGrad [34] and WaveFit[27], have also been shown to achieve comparable performance.

2.2   GAN-based Vocoders

Generative adversarial networks (GANs), as one of the most dominant deep generative models, have been applied to speech synthesis by many. Neekhara et al. are one of the first to use GANs to learn mappings from mel-spectrogram to magnitude spectrogram as a key

step for raw audio generation [41]. In their work, GANs effectively transforms the non-invertible mel-spectrogram to an invertible magnitude spectrogram, thereby enabling the conversion of the derived spectrogram back into raw audio. This development circumvents the traditional issue of information loss during the conversion from magnitude spectrogram to audio waveforms. Their approach entails training a generator network to convert the low-resolution mel-spectrograms into high-resolution magnitude spectrograms, while a discriminator network is concurrently trained to differentiate between the generated and real spectrograms. This dual-structure training process enabled the generator to progressively improve the quality of the generated spectrograms, thereby enhancing the fidelity of the resultant audio.

Yamamoto et al. tries to use GANs to directly generate speech and concludes that adversarial loss requires a KL-divergence term to achieve high quality audio generation [64]. The authors subsequently propose multi-resolution STFT loss to improve GAN training and achieve promising performance[63]. The multi-resolution STFT loss includes multiple resolution of spectral convergence and log STFT magnitude loss, enabling the generator to better learn the time-frequency characteristics of speech and avoid overfitting to a fixed STFT pattern. Their work not only opens the door of directly generating waveforms with GAN, but also highlight the required element of doing so, an auxiliary loss that aligns the input and output converted to the same data modality.
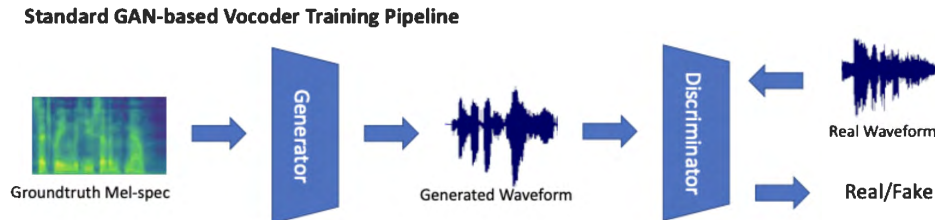
Figure 2.1: **Illustration of GAN framework.**

Subsequent researchers built upon Yamamoto et al.'s work, further enhancing the quality and fidelity of synthesized speech. MelGAN [31] and HiFi-GAN [28] with multi-scale and multi-period discriminators are recent milestones of GAN-based vocoders performing at state-of-the-art in vocoding. They use similar GAN framework with previous works by Yamamoto et al., as shown in figure 2.1. It uses a fully convolutional network as the generator and 1-D and 2-D convolutional neural networks respectively for multi-period and multi-scale discriminators. Each discriminator is composed of several sub-discriminator with similar architecture applied on different audio period and scale to capture periodic signals, consecutive patterns and long-term dependencies. HiFi-GAN uses multiple losses to stably train the models. The first loss is the GAN loss, as illustrated in equation 2.1 and 2.2, where D is the discriminator, G is the generator, X is input mel-spectrogram and Y is the ground truth waveform.

$$\mathcal{L}_{adv}(D;G) = \mathbf{E}_{(X,Y)}[(D(Y) - 1)^2 + D(G(X))] \tag{2.1}$$

$$\mathcal{L}_{adv}(G;D) = \mathbf{E}_{(X)}[(D(G(X)) - 1)^2] \tag{2.2}$$

The second loss is mel-spectrogram reconstruction loss, inspired by the STFT loss [64, 63] and the L1 image loss [19]. It measures the L1 distance between the ground truth and generated waveform converted to mel-spectrogram, as shown in equation 2.3, where ϕ is the function that transforms a waveform into the corresponding mel-spectrogram. This loss helps synthesize a realistic waveform corresponding to an input condition, and stabilizes the adversarial training process[28].

$$\mathcal{L}_{mel}(G) = \mathbf{E}_{(X,Y)}[||\phi(Y) - \phi(G(X))||_1] \tag{2.3}$$

The third loss is the feature matching loss, designed to provide additional supervision signals to align intermediate representation within the network. It computes the difference in features of the discriminator between a ground truth sample and a generated sample, as illustrated in 2.4, where M denotes total number of layers, D^i and N_i denote the features and number of features after the i-th layer of the discriminator. HiFi-GAN is trained to optimize the weighted sum of the three loss.

$$\mathcal{L}_{fm}(G;D) = \mathbf{E}_{(X,Y)}[\sum_{i=1}^{M} \frac{1}{N_i}||D^i(Y) - D^i(G(X))||_1] \tag{2.4}$$

A range of research has proposed improvements to the successful MelGAN and HiFi-GAN. Lee el al. proposes differentiable augmentation techniques to remove periodicity artifacts in the generated waveform [32]. Bak et al. designs additional discriminators to remove discovered artifacts and enhance the generated speech quality [3]. Many works also make progress on the runtime efficiency of GAN-based vocoders with modified architecture design [11, 39, 24, 47]. However, despite these advancements, the training of GAN-based vocoders is not without challenges. The training stability, data efficiency and overfitting of GAN still remains obstacles in in further improving GAN-based vocoder models..

2.3   Contrastive Learning

Contrastive learning is first used in the computer vision domain as a data-efficient way of learning rich representation of images in a self-supervised way [59, 54, 12, 2, 44]. SimCLR

[6] proposes a widely-used contrastive learning framework without requiring specialized architectures or a memory bank. It uses the backbone of a neural network to obtain a latent embedding of an input image, map it to another hidden space with an MLP layer, and perform a contrastive matching task with the mapped embedding. The task assigns augmented versions of the same image as positive pairs, and those of different images as negative pairs, and train the network to push the embedding of the positive pairs closer to each other with the Info NCE loss [44]. Generalizing to other modalities, CLIP (Contrastive Language-Image Pretraining) [49] is a contrastive learning framework that combines images and text to learn joint representations in a shared embedding space. It enables the model to understand visual and textual information together, allowing it to perform tasks like image classification and generating natural language descriptions of images.

Self-supervised learning (SSL) methods have demonstrated efficacy in a diverse array of speech domains, including representation learning [45, 15, 5, 50, 16], synthesis [42, 35, 37, 38], and multi-modality [53, 36]. However, we have not seen contrastive learning, one kind of SSL, being applied to train neural vocoders.

Chapter 3

Methodology

3.1   Mel-spectrogram Contrastive Learning

In our GAN model, the generator takes a mel-spectrogram as input and outputs a raw waveform through a stack of convolutional layers. We use a learnable feed-forward layer to project the features of the convolutional layers onto a latent space R^D , where elements of similar semantics are close to each other through contrastive learning. For each anchor in a batch of N samples, we apply masking on randomly selected intervals in time and frequency to create a positive sample, while all other $(N-1)$ input samples and $(N-1)$ masked samples are used as negative samples. Together, the method results in 1 positive pair and $2(N-1)$ negative pairs in the batch. We then adapt the InfoNCE loss [44] used in CLIP [49] for our loss function as follows:

$$\mathcal{L}_{cl} = -\frac{1}{N} \sum_{i=1}^{N} \left(\log \frac{\exp(\tau \mathbf{v}_i \cdot \mathbf{v}_k)}{\sum_{j=1;i \neq j}^{2N} \exp(\tau \mathbf{v}_i \cdot \mathbf{v}_j))} \right) \tag{3.1}$$

where $\mathbf{v}_k \in R^D$ is the masked sample from $\mathbf{v}_i \in R^D$ and τ is a temperature parameter. This method is shown in figure 3.1.

3.2   Mel-spectrogram Waveform Contrastive Learning

In addition to training solely the generator, we propose a novel task that involves contrastive spectrogram-waveform matching. This task serves to train both the generator and the discriminators, promoting rich semantic representation and preventing overfitting of the discriminators to the real or fake classification. The method is illustrated in figure 3.2. For a batch of pairs of mel-spectrograms and waveforms, we assign the labels of the true pairs to be positive and those of the other pairs to be negative, resulting in N positive pairs and
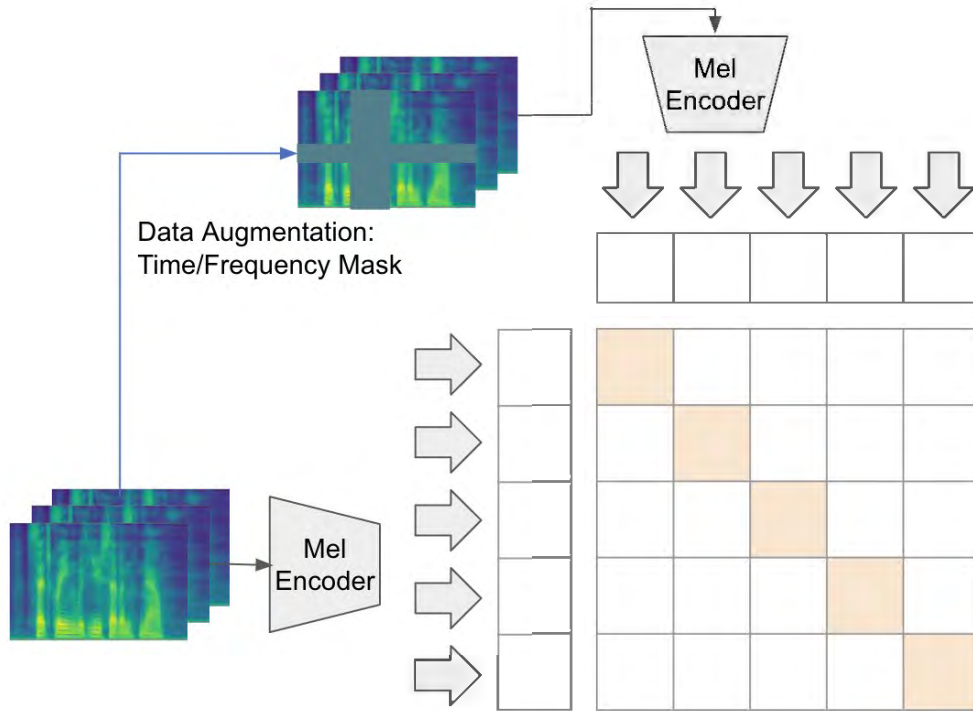
Figure 3.1: **Illustration of Mel-spectrogram Contrastive Learning.** The Mel Encoder is the backbone of the generator. This method only trains the generator in a GAN framework.

$N(N-1)$ negative pairs in a batch of N samples. We use the backbone of the generator to encode the mel-spectrogram and the backbone of the discriminator to encode the waveform. Similar to the method in section 3.1, we use two separate feed-forward layer to project each encoded features to the same latent dimension R^D. Then, we perform the modified loss function

$$\mathcal{L}_{cl} = -\frac{1}{N} \sum_{i=1}^{N} \left(\log \frac{\exp(\tau \mathbf{v}_i \cdot \mathbf{w}_i)}{\sum_{j=1; i \neq j}^{N} \exp(\tau \mathbf{v}_i \cdot \mathbf{w}_j))} \right) \tag{3.2}$$

where $\mathbf{w}_i \in R^D$ is the latent embedding of the waveform corresponding to the ith mel-spectrogram, $\mathbf{v}_i \in R^D$ is the latent embedding of the ith mel-spectrogram, and τ is a temperature parameter. HiFi-GAN contains multiple discriminators, so we calculate a contrastive loss between the mel-spectrogram embedding and each of the waveform embeddings and sum them up. For simplicity, we refer them as one discriminator in this paper unless otherwise mentioned.
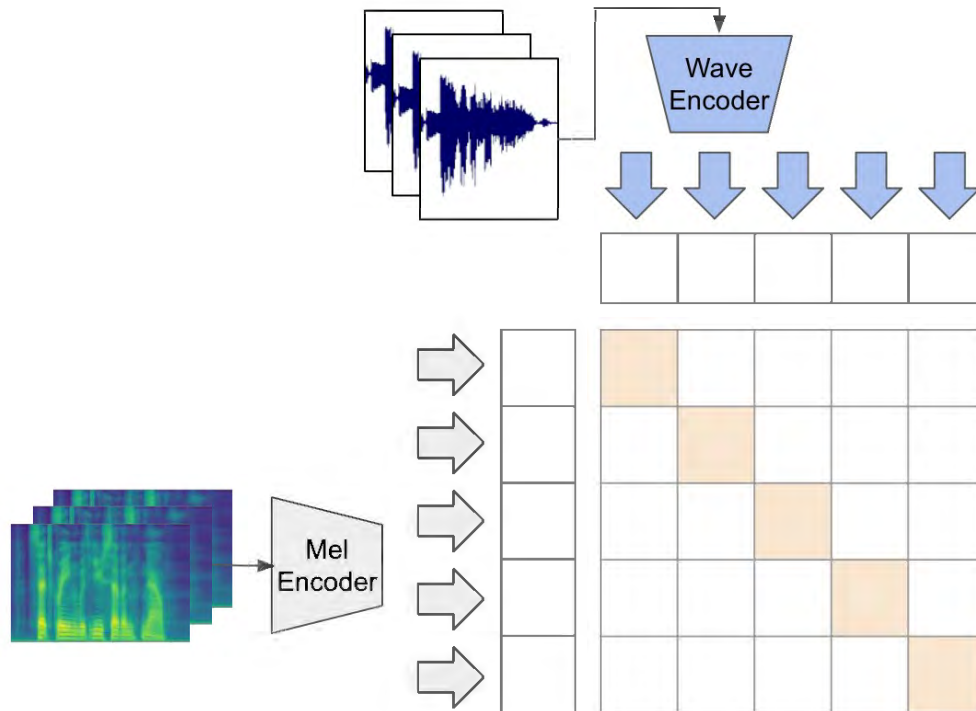
Figure 3.2: **Illustration of Mel-spectrogram & Waveform Contrastive Learning.** The Mel Encoder is the backbone of the generator, and the Wave Encoder is the backbone of the discriminator. Therefore, this method trains both the generator and discriminator.

3.3 Pretraining Framework

Inspired by recent advances of pretraining pipeline in vision and language domain [60, 49, 6], we purpose a pretraining framework for GAN-based vocoder models, as shown in Figure 3.3. The pretraining pipeline consists of generator pretraining, discriminator pretraining, and finetuning. Instead of random initialization of the weights for the generator, we want to make the generator learn a good prior that tells the model to generate realistic waveforms from mel-spectrogram. In this way, we hope that the generator can benefit from auxiliary tasks that lead to a good initialization of parameters for later finetuning. Even if we initialize the generator well, without pretraining discriminator might not give enough guidance for the generator to converge [51]. Thus we also assign auxiliary task on the discriminator as well, which we mainly focus on learning the relation between waveform features and mel-spectrogram features in order to better learn the data distribution. After learning this strong prior knowledge, it would be a harder task for the generator to generate realistic waveforms that the discriminator could detect. As pretraining completes, we load the pretrained parameters into the finetuning stage of the vocoder module. While the auxiliary tasks give

both model good prior knowledge on representation learning, we hope to get better speech synthesis performance compared to standard training procedure.

3.4 Multi-tasking Framework

Alternatively, to integrate contrastive learning with GAN tasks, we adopt a multi-tasking framework that makes auxiliary task a joint optimization objective with original learning goals [68]. As illustrated in Figure 3.3, we create additional heads for training generator and discriminator with auxiliary tasks. The total loss for training the vocoder model thus becomes:

$$\mathcal{L}_G = \mathcal{L}_{adv} + \lambda_{fm}\mathcal{L}_{fm} + \lambda_{mel}\mathcal{L}_{mel} + \lambda_{cl}\mathcal{L}_{cl} \tag{3.3}$$

$$\mathcal{L}_D = \mathcal{L}_{adv} + \mathcal{I}_{disc}\lambda_{cl}\mathcal{L}_{cl} \tag{3.4}$$

where $\mathcal{L}_G$ is the total loss for the generator and $\mathcal{L}_D$ is the total loss for the discriminator. $\mathcal{L}_{adv}$ is the adversarial loss, $\mathcal{L}_{fm}$ is the feature matching loss, and $\mathcal{L}_{mel}$ is the mel-spectrogram reconstruction loss in the original HiFi-GAN training pipeline. $\mathcal{L}_{mel}$ can be either of the contrastive loss described in section 3.1 or 3.2, and $\mathcal{I}_{disc}$ is an indicator of whether the latter is used. Each loss is weighted with a λ coefficient which can be set as hyperparameters. We use a λ_{fm} of 2, λ_{mel} of 45 from the HiFi-GAN setting [28] and a λ_{cl} of 1.

3.5 Evaluation metrics

To objectively evaluate our models compared to the baseline, we measure the mean average error (MAE) and mel-cepstral distortion (MCD) [30] on mel-spectrograms. MAE directly computes the euclidean distance between the synthesized and reference speech, as shown in equation 3.5, where X represents the reference mel-spectrogram, $\hat{X}$ represents the synthesized audio converted to mel-spectrogram, T and C are the time and frequency channels of the mel-spectrogram. In this study, T is set to 24 and C is set to 80. Lower MAE indicate closer alignment with the ground truth.

$$MAE = \frac{1}{TC}\sum_{i=1}^{C}\sum_{j=1}^{T}(X_{i,j} - \hat{X}_{i,j}) \tag{3.5}$$

MCD is a measure of difference between two sets of mel-cepstral coefficients and provides a quantitative metric of the perceptual similarity between the two sets. A smaller mel-cepstral distortion indicates a closer similarity between the synthesized and reference speech. We present MCD generalized to multiple time steps in equation 3.6, where C' is a set number of frequency channels different than C. We use $C' = 24$ that match most studies in speech
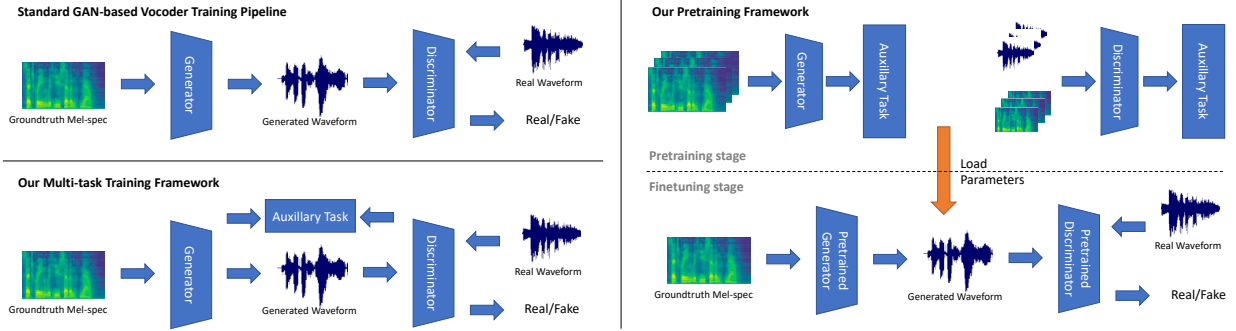
Figure 3.3: **Illustration of our multi-tasking frameworks.** GAN-based Vocoder model, such as MelGAN [31] and HiFi-GAN [28], follows an adversarial network (**top-left**) consisting of a generator that generates raw waveforms from mel-spectrograms and a discriminator that aims to distinguish real from generated waveform samples. Under the multi-tasking framework (**bottom-left**), we set the contrastive task as additional learning objectives along with the original GAN optimization objectives. This framework applies to both contrastive learning methods described in section 3.1 and 3.2. Under the pretraining framework (**right**), we pretrain the generator and the discriminator with designated auxiliary tasks, and load the pretrained parameters of both models during the finetuning stage.

synthesis. It is noteworthy that the original MCD does not average over C', so its scale can largely differ if a different C' is used.

$$MCD = \frac{1}{T} \sum_{j=1}^{T} \frac{10\sqrt{2}}{\ln 10} \sqrt{\sum_{i=1}^{C'} (X_{i,j} - \hat{X}_{i,j})} \qquad (3.6)$$

We also include a 5-scale mean opinion score (MOS) on audio quality as subjective evaluation performed on 50 samples excluded from the training set. A score of 5 represents the best quality and a score of 1 represents the worst quality. Our MOS test are performed by three college students who rate each sample exactly once.

Chapter 4

Experiments

4.1 Experimental Setting

In this section, we describe the details of our experimental settings including the dataset, model choice, hyperparameters and evaluation metrics.

Dataset

In order to have a fair comparison with other vocoder models, we train the model on LJSpeech dataset [20] which is also used in other vocoder works like HiFi-GAN [28]. LJSpeech is a public single speaker dataset with 13100 short English audio clips whose durations span from 1 second to 10 seconds. We use the default data split with 12950 training samples and 150 validation samples. We use the same preprocessing configurations with HiFi-GAN, including 80 bands mel-spectrograms as input and FFT size of 1024, window size of 1024, and hop size 256 for conversion from waveform to mel-spectrograms.[28]

Implementation details

For experimental comparison on audio quality, we choose the most powerful HiFi-GAN V1 and the most lightweight HiFi-GAN V3 as the baseline methods, and we use the same model architecture as the backbone to apply the contrastive tasks described in section 3.1 and 3.2. Under the multi-tasking framework, we train HiFi-GAN along with the contrastive learning methods with a batch size of 16, an AdamW optimizer, and a learning rate of 0.0002. For the following experiments on the full dataset, all models are trained for 400k steps (about 96 hours) on one Nvidia TITAN RTX GPU. The experiments on 20% of the dataset train for 300k steps (about 72 hours) on the same device, and those on 4% of the dataset train for 200k steps. The model inference time on GPU is about 70ms for V1 models and 32ms for V3 models.

4.2 Results

We present the results of models trained on full data with the multi-tasking framework in table 4.4. Below, we refer Mel CL as the mel-spectrogram contrastive learning in section 3.1, and Mel-Wave CL as the mel-spectrogram waveform contrastive learning in section 3.2. For V1 models, the baseline performs slightly better than the proposed methods by margins of 0.02 on MAE, 0.025 on MCD, and 0.01 on MOS. For V3 models, on the objective tests, we observe that the model trained with mel-spectrogram contrastive loss has comparable performance with the baseline, while the one trained with mel-spectrogram waveform contrastive loss achieves the highest scores on both metrics. The results show that our proposed methods have at least comparable performance to the baseline HiFi-GAN when training on the full dataset. On the subjective tests, the V3 model with Mel CL achieves the highest MOS score, 0.03 above the V3 baseline. The model with Mel-Wave CL has similar MOS score with the baseline on the full dataset. Overall, when trained on the full dataset, the proposed methods have limited gains on top of the baseline.

Table 4.1: Objective and subjective evaluation results for models with mel-spectrogram contrastive loss (Mel CL) and mel-spectrogram contrastive loss (Mel-Wave CL). Models are trained on the full training set. CI is 95% confidence interval of the MOS score.

Model	MAE	MCD	MOS (CI)
Ground Truth	-	-	4.32 ($\pm$0.05)
HiFi-GAN V1	**0.111**	**4.203**	**4.21** ($\pm$0.05)
+ Mel CL	0.114	4.289	4.18 ($\pm$0.06)
+ Mel-Wave CL	0.113	4.228	4.20 ($\pm$0.05)
HiFi-GAN V3	**0.203**	7.786	4.10 ($\pm$0.05)
+ Mel CL	0.204	7.766	**4.13** ($\pm$0.07)
+ Mel-Wave CL	**0.203**	**7.723**	4.09 ($\pm$0.06)

To investigate how each model performs under data limitation, we train the three models on 20% of the dataset and evaluate them with the same validation set. We present the results in table 4.2. With less data, the baseline HiFi-GAN V3 suffers a significant performance degradation across all metrics, including 0.371 on MCD and 0.22 on MOS. Meanwhile, the V3 model trained with Mel CL experiences an increase of 0.194 on MCD and a drop of 0.18 on MOS. The V3 model trained with Mel-Wave CL has an increase of 0.251 on MCD and a drop of only 0.05 on MOS. It suggests Mel-Wave CL is most resistant to data insufficiency. The two proposed methods have comparable scores on the objective evaluation, but the model with Mel-Wave CL obtains a significantly higher score on the subjective test, 0.16 higher than the V3 baseline. The findings align with our hypothesized alleviation of discriminator overfitting by Mel-Wave CL, which is a more severe problem on small training dataset. Both

Table 4.2: Objective and subjective evaluation results for models trained with 20% of the training set. The number in parenthesis indicates difference from the results when trained on the full dataset.

Model	MAE ↓	MCD ↓	MOS ↑ (CI)
Ground Truth	-	-	4.32 (±0.05)
HiFi-GAN V1 (20% data)	**0.113** (↑ 0.002)	4.352 (↑ 0.149)	4.13 (↓ 0.08) (±0.06)
+ Mel CL (20% data)	0.116 (↑ 0.002)	4.430 (↑ 0.139)	4.11 (↓ 0.07) (±0.07)
+ Mel-Wave CL (20% data)	**0.113** (↑ 0.000)	**4.295** (↑ 0.067)	**4.16** (↓ 0.04) (±0.06)
Hifi-GAN V3 (20% data)	0.212 (↑ 0.009)	8.157 (↑ 0.371)	3.88 (↓ 0.22) (±0.06)
+ Mel CL (20% data)	**0.207** (↑ 0.003)	**7.960** (↑ 0.206)	3.95 (↓ 0.18) (±0.06)
+ Mel-Wave CL (20% data)	**0.207** (↑ 0.004)	7.974 (↑ 0.251)	**4.04** (↓ 0.05) (±0.07)

of the proposed methods perform substantially better than the baseline by 0.07 and 0.16 respectively.

Similar trend exist in the HiFi-GAN V1 experiments, where Mel-Wave CL achieves the best scores and least performance drop on all metrics. One slightly surprising finding is that the larger model V1 often experiences smaller performance drop compared to the smaller model V3 when trained on 20% data. Typically, larger model is expected to be more prone to overfitting when trained on less data, which should lead to larger performance drop. In this specific case, however, HiFi-GAN V1 has a larger generator but the same discriminator as HiFi-GAN V3 [28], which is our suspected reason for the finding. Overall, the results show the benefits of additional supervision signals from contrastive learning in data-limited situations and the superior performance of Mel-Wave CL on small dataset.

To further validate the usefulness of Mel-Wave CL, we run a more extreme case of training on only 4% of the training set. The results are shown in table 4.3. Mel-Wave CL still outperforms the baseline V1 by significant margins on all metrics, which shows its consistency in improving the model in data-limited situations.

Table 4.3: Objective and subjective evaluation results for models trained with 4% of the training set. The number in parenthesis indicates difference from the results when trained on the full dataset.

Model	MAE	MCD	MOS (CI)
Ground Truth	-	-	4.32 (±0.05)
HiFi-GAN V1 (4% data)	0.137 (↑ 0.026)	5.372 (↑ 1.169)	3.80 (↓ 0.41) (±0.05)
+ Mel-Wave CL (4% data)	**0.135** (↑ 0.022)	**5.201** (↑ 0.973)	**3.86** (↓ 0.34) (±0.06)

We also test the results above trained with the multi-tasking framework in section 3.4 against the pretraining framework in section 3.3. We conduct this set of experiments on

HiFi-GAN V3 on 20% data, because this is the setting where our performance gain is most significant. We finetune the pretrained models with the same hyperparameters as the baseline models. We observe that the pretraining framework has very little boost to the baseline. The pretraining framework with Mel CL is better than the baseline by margins of 0.002 on MAE, 0.064 on MCD, and 0.02 on MOS. These improvements are much less significant than the multi-tasking framework. The pretraining framework with Mel-Wave CL has virtually no performance boost on MAE and MCD, and a 0.06 drop on MOS. These results suggest that our proposed contrastive task is more effective when applied in the multi-tasking framework.

Table 4.4: Objective and subjective evaluation results for models with mel-spectrogram contrastive loss (Mel CL) and mel-spectrogram contrastive loss (Mel-Wave CL). Models are trained on the full training set. CI is 95% confidence interval of the MOS score.

Model	MAE	MCD	MOS (CI)
Ground Truth	-	-	4.32 ($\pm$0.05)
HiFi-GAN V3 (20% data)	0.212	8.157	3.88 ($\pm$0.06)
+ Mel CL multi-task (20% data)	**0.207**	**7.960**	3.95 ($\pm$0.06)
+ Mel CL pretrain (20% data)	0.210	8.093	3.90 ($\pm$0.05)
+ Mel-Wave CL multi-task (20% data)	**0.207**	7.974	**4.04** ($\pm$0.07)
+ Mel-Wave CL pretrain (20% data)	0.213	8.139	3.84 ($\pm$0.07)

4.3   Ablation Studies

In this section, we introduce additional ablation studies on our proposed methods to analyze their improvement over the baseline. We examine the discriminators' overfitting problem by comparing their performance on the training and validation set when trained on 20% of data. In addition, we provides a qualitative analysis on the mel-spectrogram pixel-wise difference between the generated and ground-truth audio.

Discriminator Overfitting

Quantitatively evaluation of the discriminator can be hard due to its dependence on the generator. To examine whether it overfits the training data, we measure the discriminator accuracy on the training and validation set. If the discriminator memorizes the ground-truth in the training set, it will perform worse at classifying the positive samples in the validation set. The same measurement on negative samples indicates whether the discriminator overfits the training-time generator outputs. It is noteworthy that the negative samples are generated by different corresponding generators, which are not consistent across the three models. Therefore, we report True Positive Rate (TPR) and True Negative Rate (TNR) separately

Table 4.5: True Positive Rate (TPR) and True Negative Rate (TNR) of the discriminators on training and validation set.

	TPR(%)			TNR(%)		
Model	Train	Val	Diff	Train	Val	Diff
HiFi-GAN V3 (20%)	93.2	89.3	-3.9	95.1	94.0	-1.1
+ Mel CL (20%)	92.2	92.6	**+0.4**	94.9	94.8	**-0.1**
+ Mel-wave CL (20%)	92.1	92.2	+0.1	95.3	94.9	-0.4

in table 4.5. HiFi-GAN contains several discriminators, so we average the accuracy across them.

We present the results on HiFi-GAN V3 in table 4.5. On the positive samples, the discriminator of the baseline HiFi-GAN has the highest TPR on the training set but the lowest TPR on the validation set, which indicates that it memorizes the training samples to a larger extent. The 3.9% gap between the training and validation TPR is not high, but can accumulate through the long training. On the other hand, the discriminator trained with mel-wave contrastive loss has only a 0.1% difference. This suggests the effect of training the discriminators in the contrastive learning framework.

On the negative samples, the baseline discriminator has the largest margin of 1.1 between the training and validation set, which can result from it overfitting the generator output. On the other hand, the models with contrastive loss may suffer less from such overfitting since it has the smallest margin and highest validation accuracy. Nevertheless, the gap is really small between the models, and the accuracy on the negative samples is also affected by the quality of the generator, which varies across the models. Thus, the information TNR provides is limited.

Qualitative Evaluation

In this section, we visualize the mel-spectrogram pixel-wise difference between the generated outputs and the ground truth on one of the validation sample in figure 4.1. We observe that Mel-Wave CL results in errors of lower magnitude that are more evenly spaced out on the time axis compared to the baseline output, as shown in the red boxes. Comparing the figure of the baseline (top) and Mel-Wave CL (bottom), the baseline has much brighter color which signifies larger distance to the ground truth. These correspond to more noticeable artifacts in the generated audio. On the other hand, the Mel-Wave CL has much dimmer color which corresponds to smaller distance to the ground truth. Therefore, they are less observable by humans and make more natural audio. This aligns with the perceptual quality of its generated samples that have fewer noticeable, unnatural distortions. One explanation of this difference is that the discriminator learns to overfit the training data at specific time frames and forces the generator to adapt to these patterns which do not generalize to the
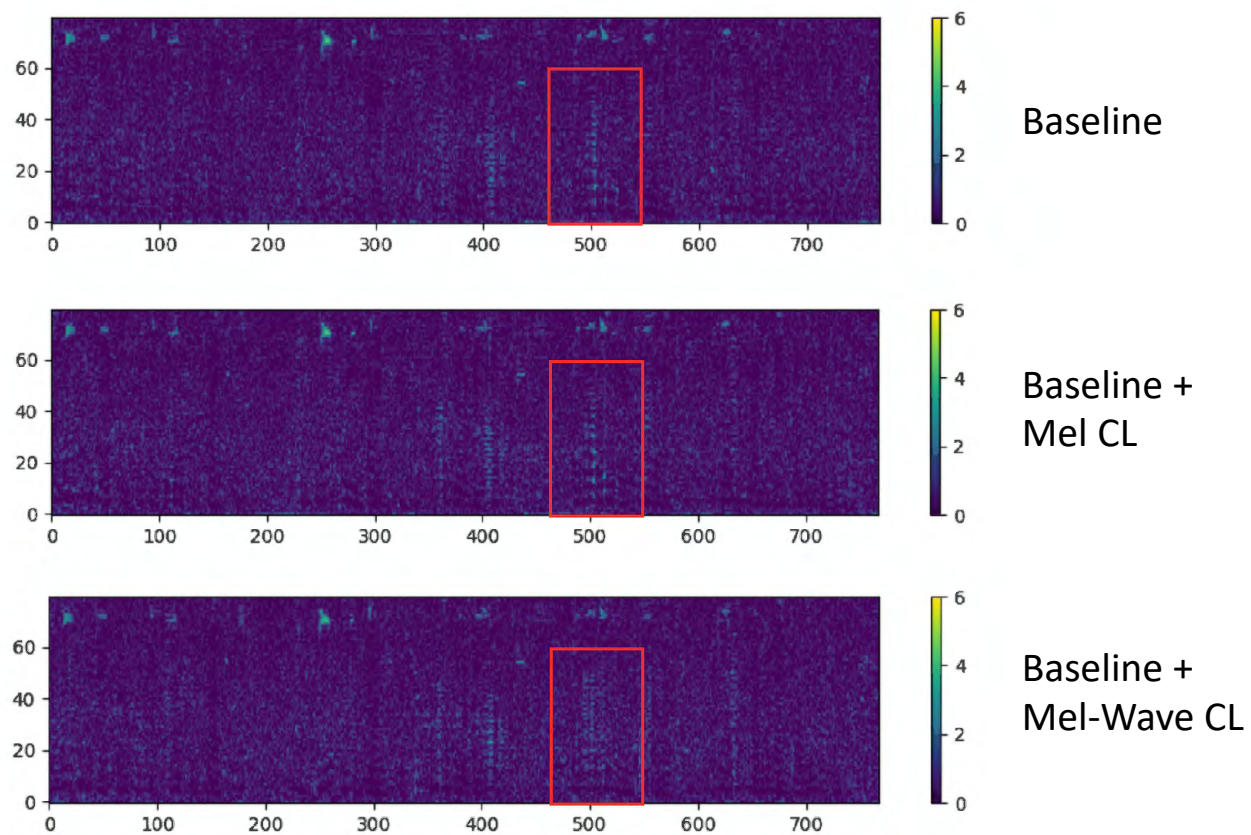
Figure 4.1: **Pixel-wise absolute difference** in the mel-spectrogram domain between the generated waveforms and the ground truth. They are generated by the V3 models.

validation data. Mel-Wave CL is able to reduce overfitting and thus does not induce large errors at these moments.

Chapter 5

Conclusion

5.1 Conclusion

This paper describes our proposed contrastive learning framework to improve GAN vocoders. Our results show the legacy of using contrastive learning as an auxiliary task that facilitates vocoder training without adding more data or modifying model architecture. We demonstrate that the proposed framework is significant especially when training on limited data by extracting additional supervision signals and reducing discriminator overfitting.

5.2 Challenges and Limitations

Despite the significant improvement, there are several challenges and limitations of this study. We discuss them below in details.

1. This project only studies its application in single-speaker speech synthesis on LJSpeech dataset. The results are not validated on multi-speaker dataset such as VCTK[62].

2. Researchers typically use a strong decoder between the learnt representation by contrastive learning and the final output which is the waveform in this study. The decoder part of HiFi-GAN has limited capacity, and it is valuable to test more design of the decoders.

3. Due to limited computation resources, we only train the models for limited time that is about a fourth of the original training time of HiFi-GAN. We still observe a small gap in our baseline and the open-sourced checkpoint. Whether the experiments still hold true when training for more extensive length is not explicitly validated. However, we observe that our proposed methods converges slightly slower than the baseline, and their improvement to the baseline is increasing during the training, so we are confident that the improvement should maintain or grow.

5.3   Future Work

For future work, we plan to repeat the experiments on different datasets and model architecture to test our methods' generalizability. In particular, we want to test its extension to multi-speaker dataset like VCTK, a domain where data insufficiency is critical. This will demonstrate the proposed method's capacity to learn from audio of multiple speaker and generalize to unseen speaker. There is also more space for contrastive learning design on multi-speaker dataset. For example, one can assign positive pairs to audio from the same speaker to encourage the learning of speaker acoustic features. On model architecture, we plan to experiment with decoders of more capacity to better align the contrastive task and the other GAN tasks. We will also explore other metrics to evaluate the discriminator overfitting problem more holistically. The TPR and TNR analysis is only one view of the problem, and more evaluation such as swapping the discriminators of two GAN models can be done to further validate the usefulness of our method in addressing discriminator overfitting. Finally, we will fully train the model with the proposed framework and open-source the checkpoint for public use and contribution.

On a broader scale, there are many promising future research directions in this field. First, self-supervised learning methods beyond contrastive learning is under-explored in vocoders. For example, masked reconstruction loss [10] has proved successful in image pretraining and can be smoothly applied mel-spectrograms. Interestingly, the mel reconstruction loss of HiFi-GAN can be considered as a modified version of the masked reconstruction loss with an additional STFT conversion. Applying the masked reconstruction task may be able to preserve benefits of HiFi-GAN and enhance the understanding of mel-spectrograms. Second, the training efficiency of GAN-based vocoders has much room for improvement. HiFi-GAN V1 requires around 2 million steps to fully converge, which takes weeks on a single GPU. On the other hand, the more efficient HiFi-GAN V3 has a large drop in output fidelity compared to HiFi-GAN V1. More efficient models of affordable performance drop are needed for broader applications of GAN-based vocoders in real life.

Despite the limitations and unsolved challenges, we still expect GAN-based vocoders to play an important role in various applications. Some GAN-based vocoders have achieved superior CPU inference time compared to autoregressive models, making them suitable for on-device applications. They also achieve the best speech quality and fidelity on various datasets, highlighting their capability in generating the finest audio in cases where both clarity and naturalness are desired. With more research in this field, GAN-based vocoders will continue to push the frontier of speech synthesis.

Bibliography

[1]   Martin Arjovsky and Leon Bottou. "Towards Principled Methods for Training Generative Adversarial Networks". In: *International Conference on Learning Representations*. 2017. URL: https://openreview.net/forum?id=Hk4_qw5xe.

[2]   Philip Bachman, R Devon Hjelm, and William Buchwalter. *Learning Representations by Maximizing Mutual Information Across Views*. 2019. arXiv: 1906.00910 [cs.LG].

[3]   Taejun Bak et al. *Avocodo: Generative Adversarial Network for Artifact-free Vocoder*. 2022. DOI: 10.48550/ARXIV.2206.13404. URL: https://arxiv.org/abs/2206.13404.

[4]   Nanxin Chen et al. *WaveGrad: Estimating Gradients for Waveform Generation*. 2020. arXiv: 2009.00713 [eess.AS].

[5]   Sanyuan Chen et al. "Wavlm: Large-scale self-supervised pre-training for full stack speech processing". In: *IEEE Journal of Selected Topics in Signal Processing* 16.6 (2022), pp. 1505–1518.

[6]   Ting Chen et al. *A Simple Framework for Contrastive Learning of Visual Representations*. 2020. DOI: 10.48550/ARXIV.2002.05709. URL: https://arxiv.org/abs/2002.05709.

[7]   Laurent Dinh, David Krueger, and Yoshua Bengio. *NICE: Non-linear Independent Components Estimation*. 2015. arXiv: 1410.8516 [cs.LG].

[8]   Aritra Ghosh and Andrew Lan. "Contrastive Learning Improves Model Robustness Under Label Noise". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. June 2021, pp. 2703–2708.

[9]   Ian J. Goodfellow et al. *Generative Adversarial Networks*. 2014. DOI: 10.48550/ARXIV.1406.2661. URL: https://arxiv.org/abs/1406.2661.

[10] Kaiming He et al. *Masked Autoencoders Are Scalable Vision Learners*. 2021. arXiv: 2111.06377 [cs.CV].

[11] Mengnan He et al. "Improving GAN-based vocoder for fast and high-quality speech synthesis". In: *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*. Ed. by Hanseok Ko and John H. L. Hansen. ISCA, 2022, pp. 1601–1605. DOI: 10.21437/Interspeech.2022-730. URL: https://doi.org/10.21437/Interspeech.2022-730.

[12] Olivier J. Hénaff et al. *Data-Efficient Image Recognition with Contrastive Predictive Coding*. 2020. arXiv: 1905.09272 [cs.CV].

[13] Dan Hendrycks et al. "Using Self-Supervised Learning Can Improve Model Robustness and Uncertainty". In: *Neural Information Processing Systems*. 2019.

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. *Denoising Diffusion Probabilistic Models*. 2020. DOI: 10.48550/ARXIV.2006.11239. URL: https://arxiv.org/abs/2006.11239.

[15] Wei-Ning Hsu et al. *HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units*. 2021. DOI: 10.48550/ARXIV.2106.07447. URL: https://arxiv.org/abs/2106.07447.

[16] Po-Yao Huang et al. *Masked Autoencoders that Listen*. 2022. DOI: 10.48550/ARXIV.2207.06405. URL: https://arxiv.org/abs/2207.06405.

[17] Satoshi Imai. "Cepstral analysis synthesis on the mel frequency scale". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing*. 1983.

[18] Satoshi Imai, Kazuo Sumita, and Chieko Furuichi. "Mel Log Spectrum Approximation (MLSA) filter for speech synthesis". In: *Electronics and Communications in Japan Part I-communications* 66 (1983), pp. 10–18.

[19] Phillip Isola et al. "Image-to-Image Translation with Conditional Adversarial Networks". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 5967–5976. DOI: 10.1109/CVPR.2017.632.

[20] Keith Ito and Linda Johnson. *The LJ Speech Dataset*. https://keithito.com/LJ-Speech-Dataset/. 2017.

[21] Jongheon Jeong and Jinwoo Shin. "Training {GAN}s with Stronger Augmentations via Contrastive Discriminator". In: *International Conference on Learning Representations*. 2021. URL: https://openreview.net/forum?id=eo6U4CAwVmg.

[22] Michael I. Jordan. "Serial Order: A Parallel Distributed Processing Approach". In: *Advances in psychology* 121 (1997), pp. 471–495.

[23] Nal Kalchbrenner et al. *Efficient Neural Audio Synthesis*. 2018. arXiv: 1802.08435 [cs.SD].

[24] Takuhiro Kaneko et al. "MISRNet: Lightweight Neural Vocoder Using Multi-Input Single Shared Residual Blocks". In: *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*. Ed. by Hanseok Ko and John H. L. Hansen. ISCA, 2022, pp. 1631–1635. DOI: 10.21437/Interspeech.2022-11152. URL: https://doi.org/10.21437/Interspeech.2022-11152.

[25] Tero Karras et al. "Training Generative Adversarial Networks with Limited Data". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 12104–12114. URL: https://proceedings.neurips.cc/paper/2020/file/8d30aa96e72440759f74bd2306c1fa3d-Paper.pdf.

[26] Hideki Kawahara. "STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds". In: *Acoustical Science and Technology* 27 (2006), pp. 349–353.

[27] Yuma Koizumi et al. *WaveFit: An Iterative and Non-autoregressive Neural Vocoder based on Fixed-Point Iteration*. 2022. arXiv: 2210.01029 [eess.AS].

[28] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis". In: *ArXiv* abs/2010.05646 (2020).

[29] Zhifeng Kong et al. "DiffWave: A Versatile Diffusion Model for Audio Synthesis". In: *International Conference on Learning Representations*. 2021. URL: https://openreview.net/forum?id=a-xFK8Ymz5J.

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

[31] Kundan Kumar et al. "MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper/2019/file/6804c9bca0a615bdb9374d00a9fcba59-Paper.pdf.

[32] Junhyeok Lee et al. "PHASEAUG: A Differentiable Augmentation for Speech Synthesis to Simulate One-to-Many Mapping". In: *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, June 2023. DOI: 10.1109/icassp49357.2023.10096374. URL: https://doi.org/10.1109%5C%2Ficassp49357.2023.10096374.

[33] Sang-gil Lee et al. "BigVGAN: A Universal Neural Vocoder with Large-Scale Training". In: *The Eleventh International Conference on Learning Representations*. 2023. URL: https://openreview.net/forum?id=iTtGCMDEzS_.

[34] Sang-gil Lee et al. *PriorGrad: Improving Conditional Denoising Diffusion Models with Data-Dependent Adaptive Prior*. 2022. arXiv: 2106.06406 [stat.ML].

[35] Jiachen Lian, Chunlei Zhang, and Dong Yu. "Robust disentangled variational speech representation learning for zero-shot voice conversion". In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2022, pp. 6572–6576.

[36] Jiachen Lian et al. "AV-data2vec: Self-supervised Learning of Audio-Visual Speech Representations with Contextualized Target Representations". In: *arXiv preprint arXiv:2302.06419* (2023).

[37] Jiachen Lian et al. "Towards improved zero-shot voice conversion with conditional dsvae". In: *arXiv preprint arXiv:2205.05227* (2022).

[38] Jiachen Lian et al. "Utts: Unsupervised tts with conditional disentangled sequential variational auto-encoder". In: *arXiv preprint arXiv:2206.02512* (2022).

[39] Manh Luong and Viet-Anh Tran. "FlowVocoder: A small Footprint Neural Vocoder based Normalizing Flow for Speech Synthesis". In: *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*. Ed. by Hanseok Ko and John H. L. Hansen. ISCA, 2022, pp. 1576–1580. DOI: `10.21437/Interspeech.2022-272`. URL: `https://doi.org/10.21437/Interspeech.2022-272`.

[40] Mara Mills. "Media and Prosthesis: The Vocoder, the Artificial Larynx, and the History of Signal Processing". In: *Qui Parle* 21.1 (June 2012), pp. 107–149. ISSN: 1041-8385. DOI: `10.5250/quiparle.21.1.0107`. eprint: `https://read.dukeupress.edu/qui-parle/article-pdf/21/1/107/379716/107Mills.pdf`. URL: `https://doi.org/10.5250/quiparle.21.1.0107`.

[41] Paarth Neekhara et al. *Expediting TTS Synthesis with Adversarial Vocoding*. 2019. DOI: `10.48550/ARXIV.1904.07944`. URL: `https://arxiv.org/abs/1904.07944`.

[42] Junrui Ni et al. "Unsupervised text-to-speech synthesis by unsupervised automatic speech recognition". In: *arXiv preprint arXiv:2203.15796* (2022).

[43] Aaron van den Oord et al. *Parallel WaveNet: Fast High-Fidelity Speech Synthesis*. 2017. arXiv: `1711.10433 [cs.LG]`.

[44] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. *Representation Learning with Contrastive Predictive Coding*. 2018. DOI: `10.48550/ARXIV.1807.03748`. URL: `https://arxiv.org/abs/1807.03748`.

[45] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. "Representation learning with contrastive predictive coding". In: *arXiv preprint arXiv:1807.03748* (2018).

[46] Aaron van den Oord et al. *WaveNet: A Generative Model for Raw Audio*. 2016. DOI: `10.48550/ARXIV.1609.03499`. URL: `https://arxiv.org/abs/1609.03499`.

[47] Sangjun Park et al. "Bunched LPCNet2: Efficient Neural Vocoders Covering Devices from Cloud to Edge". In: *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*. Ed. by Hanseok Ko and John H. L. Hansen. ISCA, 2022, pp. 808–812. DOI: `10.21437/Interspeech.2022-310`. URL: `https://doi.org/10.21437/Interspeech.2022-310`.

[48] Ryan Prenger, Rafael Valle, and Bryan Catanzaro. *WaveGlow: A Flow-based Generative Network for Speech Synthesis*. 2018. arXiv: `1811.00002 [cs.SD]`.

[49] Alec Radford et al. *Learning Transferable Visual Models From Natural Language Supervision*. 2021. DOI: `10.48550/ARXIV.2103.00020`. URL: `https://arxiv.org/abs/2103.00020`.

[50] Aaqib Saeed, David Grangier, and Neil Zeghidour. *Contrastive Learning of General-Purpose Audio Representations*. 2020. DOI: `10.48550/ARXIV.2010.10915`. URL: `https://arxiv.org/abs/2010.10915`.

[51] Axel Sauer et al. "Projected GANs Converge Faster". In: *CoRR* abs/2111.01007 (2021). arXiv: `2111.01007`. URL: `https://arxiv.org/abs/2111.01007`.

[52] Jonathan Shen et al. *Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions*. 2018. arXiv: `1712.05884 [cs.CL]`.

[53] Bowen Shi et al. "Learning audio-visual speech representation by masked multimodal cluster prediction". In: *arXiv preprint arXiv:2201.02184* (2022).

[54] Yonglong Tian, Dilip Krishnan, and Phillip Isola. *Contrastive Multiview Coding*. 2020. arXiv: `1906.05849 [cs.CV]`.

[55] Keiichi Tokuda et al. "Speech parameter generation algorithms for HMM-based speech synthesis". In: *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100)* 3 (2000), 1315–1318 vol.3.

[56] Keiichi Tokuda et al. "Speech Synthesis Based on Hidden Markov Models". In: *Proceedings of the IEEE* 101.5 (2013), pp. 1234–1252. DOI: `10.1109/JPROC.2013.2251852`.

[57] Hung-Yu Tseng et al. "Regularizing Generative Adversarial Networks under Limited Data". In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 7917–7927.

[58] Jean-Marc Valin and Jan Skoglund. *LPCNet: Improving Neural Speech Synthesis Through Linear Prediction*. 2019. arXiv: `1810.11846 [eess.AS]`.

[59] Zhirong Wu et al. *Unsupervised Feature Learning via Non-Parametric Instance-level Discrimination*. 2018. arXiv: `1805.01978 [cs.CV]`.

[60] Chenfeng Xu et al. *PreTraM: Self-Supervised Pre-training via Connecting Trajectory and Map*. 2022. DOI: `10.48550/ARXIV.2204.10435`. URL: `https://arxiv.org/abs/2204.10435`.

[61] Yihao Xue, Kyle Whitecross, and Baharan Mirzasoleiman. "Investigating Why Contrastive Learning Benefits Robustness against Label Noise". In: *First Workshop on Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022*. 2022. URL: `https://openreview.net/forum?id=s436PHXRzMm`.

[62] Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. "CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92)". In: 2019.

[63] 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 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2019), pp. 6199–6203.

[64] Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. *Probability density distillation with generative adversarial networks for high-quality parallel waveform generation.* 2019. DOI: 10.48550/ARXIV.1904.04472. URL: https://arxiv.org/abs/1904.04472.

[65] Takayoshi Yoshimura et al. "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis". In: *6th European Conference on Speech Communication and Technology (Eurospeech 1999)* (1999).

[66] Shengyu Zhao et al. "Differentiable Augmentation for Data-Efficient GAN Training". In: *ArXiv* abs/2006.10738 (2020).

[67] Evgenii Zheltonozhskii et al. "Contrast To Divide: Self-Supervised Pre-Training for Learning With Noisy Labels". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. Jan. 2022, pp. 1657–1667.

[68] Fengda Zhu et al. *Vision-Language Navigation with Self-Supervised Auxiliary Reasoning Tasks.* 2019. DOI: 10.48550/ARXIV.1911.07883. URL: https://arxiv.org/abs/1911.07883.