Enhancing Neural Vocoders for High-Quality Speech Synthesis

ZHOU Hanbo *BTGY7K*, YANG Deyu *E44MLP*, XIA Xihang *BX7F0X*

Dec. 2024

1 Introduction and Related Work

Introduction

Neural vocoders are essential in modern text-to-speech (TTS) systems, converting intermediate acoustic representations, such as mel-spectrograms, into high-quality speech waveforms. Models like HiFi-GAN and AutoVocoder have significantly improved audio fidelity and computational efficiency, but challenges persist in achieving robustness to speaker variability, handling environmental noise, and maintaining real-time synthesis performance.

This project explores neural vocoder architectures, focusing on HiFi-GAN and AutoVocoder, to enhance quality, speed, and generalization. By training and testing these vocoders on datasets like LJSpeech (single-speaker) and VCTK (multi-speaker), the aim is to synthesize natural, high-quality speech that generalizes well across diverse speakers and noise conditions.

Related Work References

- **HiFi-GAN**: Kong et al. (2020) introduced HiFi-GAN, a GAN-based vocoder with a multi-scale discriminator and generator optimized using mel-spectrogram and feature-matching losses. It achieves high fidelity and real-time synthesis performance.
  - Reference: Kong, J., Kim, J., & Bae, J. (2020). *HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis*. NeurIPS. https://github.com/jik876/hifi-gan
- **AutoVocoder**: Designed to streamline vocoder functionality by focusing on computational efficiency and robustness. AutoVocoder introduces architectural simplifications while ensuring competitive perceptual quality.
  - Reference: https://github.com/hcy71o/AutoVocoder

2 Methods of Training

Type of Neural Network

- **HiFi-GAN**:
  - Generator: A convolutional architecture with dilated and residual layers designed to model long-range dependencies in audio.
  - Discriminator: Multi-scale and multi-period discriminators evaluate the realism of generated audio at various resolutions and periodicities.
- **AutoVocoder**:
  - Enhancements include attention mechanisms for better temporal modeling and additional periodic discriminators for robustness.

Datasets

- *LJSpeech*: Single-speaker dataset (24 hours of clean speech).
- *VCTK*: Multi-speaker dataset (109 speakers with diverse accents and noise conditions).

Hyperparameter

- Learning Rate: 0.0002.
- Learning Rate Decay: 0.999.
- Batch Size: 16.
- Optimizer: Adam ($\beta_1 = 0.8, \beta_2 = 0.99$).
- N FFT: 1024.
- Hop Size: 256.
- Win Size: 1024.
- Sampling Rate: 22050.

Loss Functions

- **HiFi-GAN**:
  - Adversarial Loss: Drives realism by training the generator to fool the discriminator.
  - Mel-Spectrogram Loss: Ensures alignment between generated and target spectrograms.
  - Feature Matching Loss: Encourages perceptual similarity in intermediate features.
- **AutoVocoder**:
  - Additional attention and periodic losses for better pitch and temporal alignment.

Training Setup

- Duration: 210000 steps.
- Hardware: NVIDIA 4090D GPU.
- Tools: PyTorch, `librosa` for preprocessing, and `torchaudio`.

3 Evaluation

Evaluation Metrics

- **PESQ (Perceptual Evaluation of Speech Quality)**

  Function: Predicts speech quality by comparing the synthesized audio with a reference signal, simulating human auditory perception.
  Formula:

  $$Q_{\text{PESQ}} = f\big(D_{\text{dist}}, D_{\text{clean}}\big)$$

  where:
  - D_{dist}: Perceived distortion in degraded speech.

– D_{clean}: Perceived distortion in clean reference speech.

Output: A scalar score ranging from -0.5 to 4.5.

- **MCD (Mel Cepstral Distortion)** :

  Function: Measures the spectral distortion between synthesized and reference audio using mel-cepstral coefficients (MCCs).
  Formula:

$$\text{MCD} = \frac{10}{\ln(10)} \sqrt{2 \sum_{i=1}^{D} (c_i^{\text{gen}} - c_i^{\text{ref}})^2}$$

  where:

  – c_i^{gen}: Generated mel-cepstral coefficient.

  – c_i^{ref}: Reference mel-cepstral coefficient.

  – D: Dimensionality of mel-cepstral coefficients.

  Output: Measured in decibels (dB), lower values indicate better quality.

- **Robustness**:

  – Performance on unseen speakers and noisy datasets. A test by using our own voice recorded in a cafe.

Results

Metric	**HiFi-Gan**	**AutoVocoder**
Average generation time per file	0.02 seconds	0.01 seconds
Average GPU memory usage per file	0.19 MB	0.18 MB

Table 1: Summary of Average Time and GPU Memory Usage for HiFi-Gan and AutoVocoder

Model	Average PESQ Score	Average MCD Score
Autovocoder	3.69	3.57
Hifi-GAN	2.73	4.61

Table 2: Comparison of Average PESQ and MCD Scores for Autovocoder and Hifi-GAN

Visualizations

First test file: *ms gt 5*

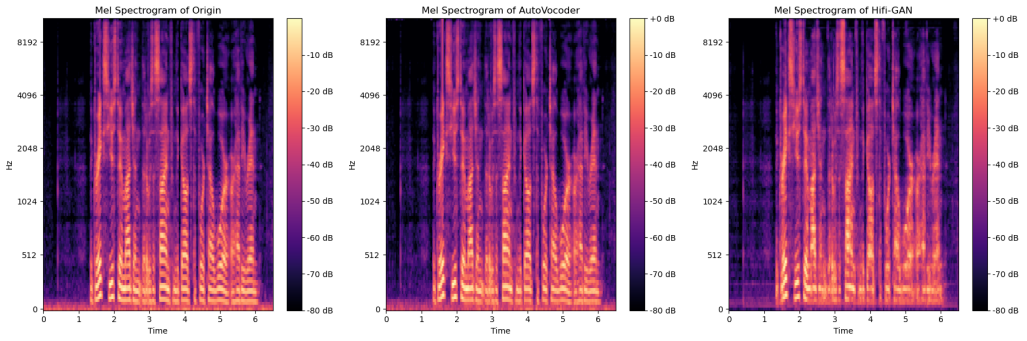

Figure 1: Comparison of Mel Spectrograms for Origin, Hifi-GAN, and AutoVocoder

Figure 1 provides a comparison of the Mel spectrograms for the original audio (‘Origin‘), the AutoVocoder-generated spectrogram, and the Hifi-GAN-generated spectrogram.

- **Origin vs. AutoVocoder:** The AutoVocoder spectrogram shows strong preservation of low-frequency components (below 2048 Hz), which are critical for capturing human speech features. However, slight smoothing and loss of detail are noticeable in the higher frequency bands, reducing fidelity to the original.
- **Origin vs. Hifi-GAN:** The Hifi-GAN spectrogram captures the general structure of the original but exhibits more distortions, particularly

in the mid- to high-frequency range. This could lead to less accurate reproduction of certain phonetic details, affecting the overall quality.

- **Overall Comparison:** AutoVocoder demonstrates better preservation of the lower frequencies compared to Hifi-GAN, making it more aligned with the original in critical areas for human speech intelligibility. However, both models show room for improvement in reconstructing higher frequencies.

The comparison highlights the differences in reconstruction quality between the Hifi-GAN and AutoVocoder models, with the latter performing closer to the original in this example.

Second test file: *my voice*

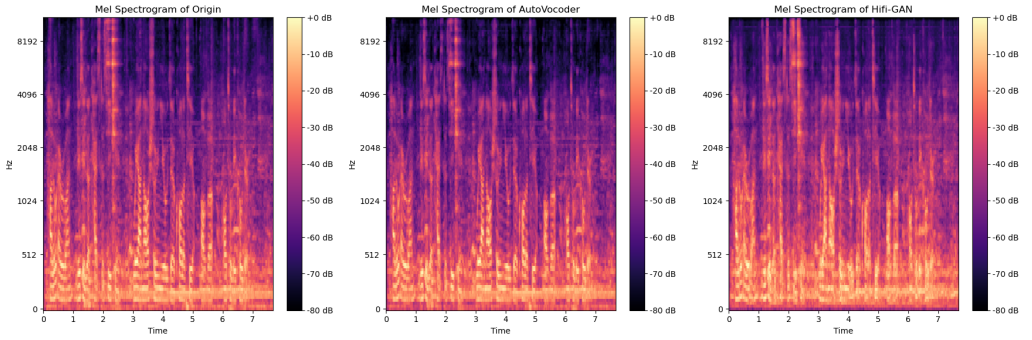

Figure 2: Comparison of Mel Spectrograms for Origin, Hifi-GAN, and AutoVocoder

Figure 2 compares the Mel spectrograms for the original audio (‘Origin‘), the AutoVocoder-generated audio, and the Hifi-GAN-generated audio.

- **Origin vs. AutoVocoder:** The AutoVocoder spectrogram closely resembles the original in the lower frequency range (below 2048 Hz), capturing essential details needed for human speech intelligibility. However, minor smoothing and slight artifacts can be observed in the higher frequency bands, indicating some loss in high-frequency detail.

- **Origin vs. Hifi-GAN:** The Hifi-GAN spectrogram captures the overall structure of the original but shows more distortions in the mid- and high-frequency ranges. This results in less accurate reproduction of fine-grained details, which could affect audio quality.
- **Overall Comparison:** AutoVocoder demonstrates better alignment with the original spectrogram in preserving key speech features, especially in the critical frequency range. Hifi-GAN, while maintaining a similar structure, struggles to reconstruct high-frequency components effectively.

4 Conclusion

Summary

The project implemented and compared neural vocoders for high-quality speech synthesis. HiFi-GAN demonstrated strong single-speaker performance, while AutoVocoder shows good adaptability and robustness across multi-speaker datasets.