THE HONG KONG POLYTECHNIC UNIVERSITY

CAPSTONE PROJECT

Comparison of the Performance of GAN and Diffusion Models in the Vocoder Field

Author:
XUE, Bingxin

Supervisor:
Dr. Jiang, Binyan

A thesis submitted in fulfillment of the requirements
for the degree of Bachelor of Science
in the
Data Science and Analytics

May 12, 2024

A Signal

A.1 Phase Estimation

Phase specifies the location or timing of a point within a wave cycle of a repetitive waveform. It is critical to signal reconstruction because it contains timing information about the waveform. The reconstructed signal may lack important details without accurate phase information, resulting in distorted audio. (University of Rhode Island and Inner Space Center, n.d.)

A.2 Short-Time Fourier Transform

The signal is obtained through sampling; it is a discrete-time sequence signal, so we need to use the Discrete Time Fourier Series (DFT) to process the signal accordingly.

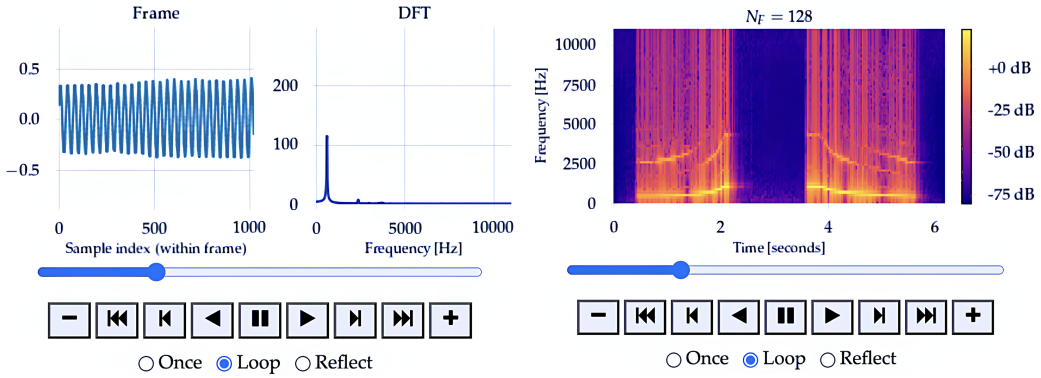

Figure 11: Outputs of audio data processing.
with DFT and STFT (McFee, 2023a)

STFT is the process of taking small pieces of the signal (length L) and analyzing the DFT of each piece. Compared to DFT, STFT can provide time and frequency resolution. The spectrogram is a visualization of STFT, displaying the frequency changes over time in the audio signal. (Xuan, 2023)

A.3 Spectral Leakage

When the signal's waveform is not a complete period, performing a DFT or Fast Fourier Transform (FFT) will produce discontinuities, which lead to leakage of frequency energy from the final spectrum obtained to other frequencies. This phenomenon is called spectral leakage. (McFee, 2023a)

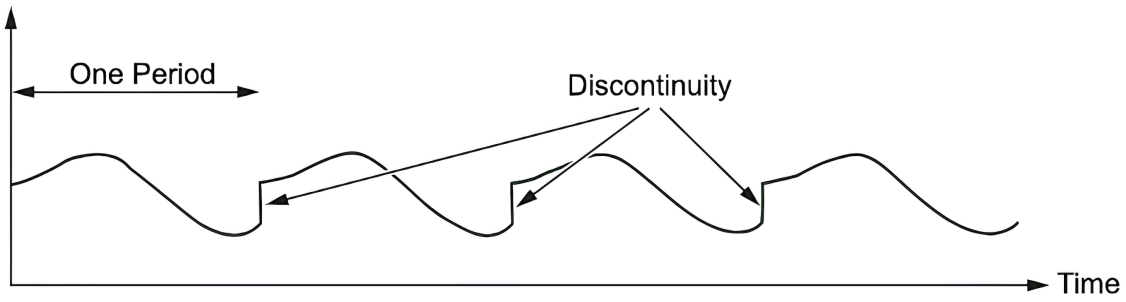

Figure 12: Discontinuity of signal (NI, 2023)

However, spectral leakage cannot be avoided in general. We can only control leakage to direct the leaked energy in various ways, i.e. windowing. (McFee, 2023b)

A.4 Window

Window is multiplying the signal $x[n]$ by another signal $w[n]$ of the same duration, resulting in the windowed DFT: $\hat{X} \leftarrow \text{DFT}(x[n] \cdot w[n])$ (McFee, 2023b)

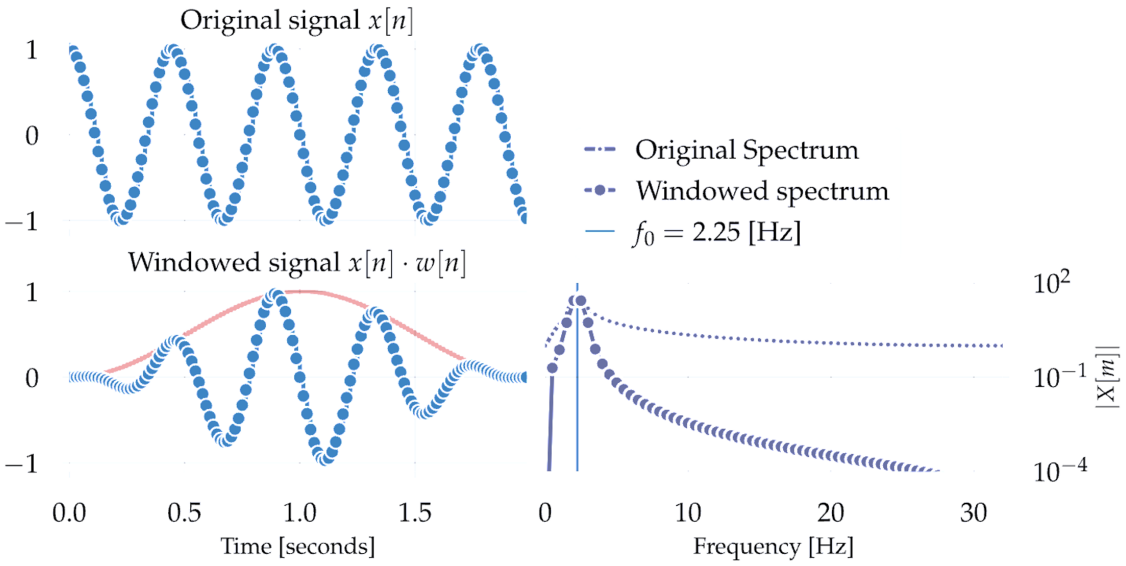

Figure 13: Window (McFee, 2023b)

A.5 Amplitude Spectrum

The amplitude spectrum is a simple transformation of the DFT. (Taboga, n.d.)

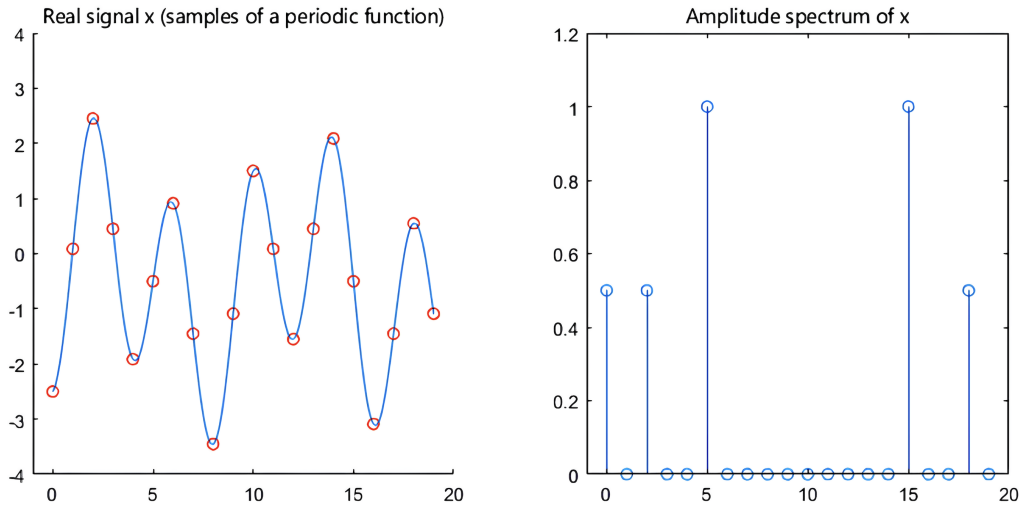

Figure 14: Amplitude Spectrum Example (Taboga, n.d.)

A.6 Mel-Spectrogram

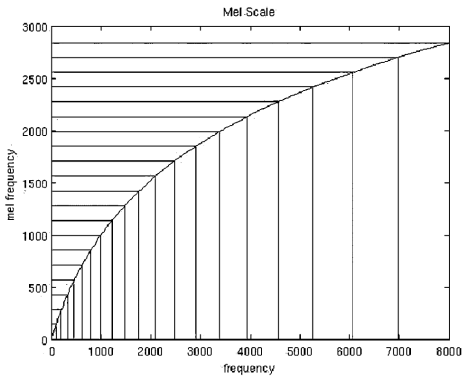

Figure 15: Mel Scale (Pei, 2010)

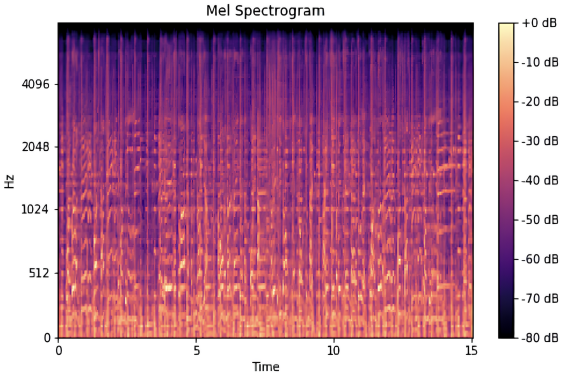

Figure 16: Mel-Spectrogram (Roberts, 2020)

Human perception of sound frequency is limited, and we are more sensitive to differences in low-frequency sounds but less sensitive to differences in high-frequency sounds. In 1937, Stevens, Volkmann, and Newmann proposed a unit of pitch called the Mel scale, which solved this problem. The Mel scale spectrogram converts frequencies into spectrograms based on the Mel scale. It simulates the characteristics of the human auditory system by expanding the frequency intervals in the high-frequency region, making differences in the high-frequency region easier to perceive. Therefore, it is widely used in speech processing and

audio signal analysis. Mel-spectrogram is also used as the input for the vocoder in this experiment. (Roberts, 2020)

B Model

B.1 Markov Chain

A Markov chain is a stochastic process with the Markov property, meaning that the conditional distribution of future states given the current state depends only on the current state and is independent of past states. As the following formula:

$$\mathbb{P}(X_{n+1} = s_{n+1} | X_n = s_n, X_{n-1} = s_{n-1}, X_{n-2} = s_{n-2}, ...) = \mathbb{P}(X_{n+1} = s_{n+1} | X_n = s_n) \tag{12}$$

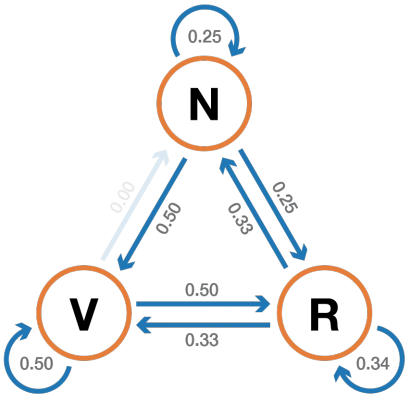

Figure 17: Graph representation of the Markov chain modelling our fictive TDS reader behaviour. (Rocca, 2019)

It is often used to model sequences of random events occurring in discrete time and discrete state space. (Rocca, 2019)

B.2 Receptive Field

A receptive field is a specific area that the nerves perceive and process. In biology, the receptive field is the area processed by each visual nerve. We apply biological concepts to deep learning; the area of data processed by each nerve in a neural network is called the receptive field. For example, in image processing, the receptive field is the area covered by each convolution kernel moving across the input image. (Katta, 2024)

B.3 Sampling

Sampling can be categorised into upsampling and downsampling.

After upsampling, the dimension of the output feature space is larger than the dimension of the input feature map, which can enrich the details of the sample. Assuming the sample is a picture, we can increase the image's resolution by upsampling.

After downsampling, the dimension of the feature space is smaller than the dimension of the feature map at input. It removes redundant information and allows us to notice important features in the sample, thus capturing the feature distribution of the sample. (Chaubey, 2020)

B.4 Convolution

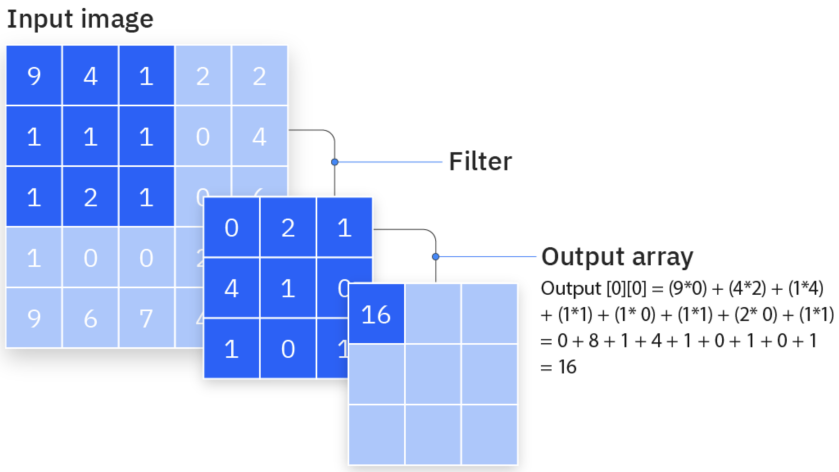

Figure 18: Convolution Operation

Convolution is a downsampling operation，can extract features from samples. The convolutional layer is the core component of the neural network where most of the computation occurs. (Anwar, 2020)

$$o = \frac{i + 2p - k}{s} + 1 \tag{13}$$

, where i = input, k = kernel, p = padding, and s = stride, o = the size of the output feature map.

Assuming we perform convolution on a colour image, the input data comprises three channels corresponding to the three primary colours of RPG. Each channel is a two-dimensional matrix representing the pixel values of the corresponding

channel in the image. The filter, also called the kernel, is a small weight matrix that slides over the image, performs an elemental multiplication operation with the current portion of the input, and then aggregates all the results into a pixel. This process is called convolution. Eventually, we will generate a new output image through convolution.

B.5 Transposed Convolution

Transposed convolution is an inverse convolution operation process, an upsampling operation. It can enrich the details of the sample and recover the original size of the sample. (Anwar, 2020)

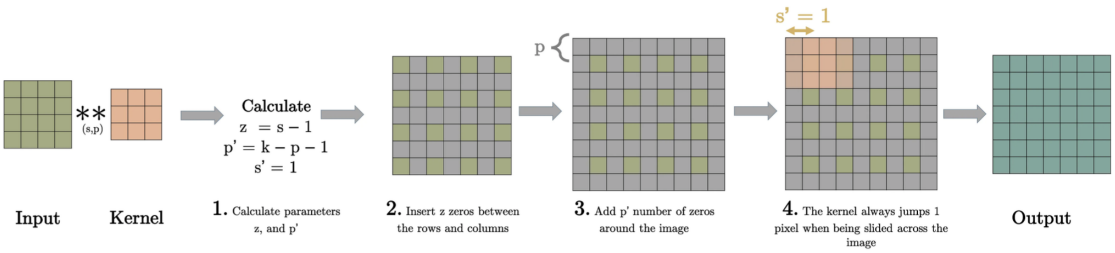

Figure 19: Transposed Convolution Process. (Anwar, 2020)

$$o = (i - 1) \times s + k - 2p \tag{14}$$

, where i = input, k = kernel, p = padding, and s = stride, o = the size of the output feature map.

Comparison					
Conv Type	**Operation**	**Zero Insertions**	**Padding**	**Stride**	**Output Size**
Standard	Downsampling	0	p	s	(i+2p-k)/s +1
Transposed	Upsampling	(s - 1)	(k-p-1)	1	(i-1)*s+k-2p

Figure 20: Summary of Convolution and Transposed Convolution. (Anwar, 2020)

B.6 Residual Blocks

If x is the input and $H(x)$ is the true distribution, then the residue is the difference between them: $R(x) = \text{Output} - \text{Input} = H(x) - x$.

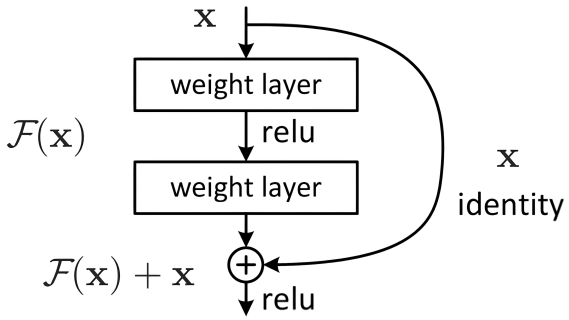

Figure 21: Single Residual Block (He et al., 2016)

Adding a residual block allows the network to learn the true distribution more efficiently and mitigate the vanishing gradient problem. (Sahoo, 2018)

B.7 Activation Functions

An activation function is a mathematical function applied to the output of a neuron in a neural network, which introduces non-linearity so that the neural network can model complex relationships between inputs and outputs. In a neural network, the output of a neuron is calculated by multiplying the inputs with the corresponding weights, summing them, adding a bias term, and then transforming the result into a nonlinear form through an activation function. Without an activation function, the neural network would be limited to modelling a linear relationship between the inputs and outputs.

For example,the Rectified Linear Unit (ReLU) is one of the activation functions.The function is defined as $f(x) = \max(0, x)$. (Rallabandi, 2023)

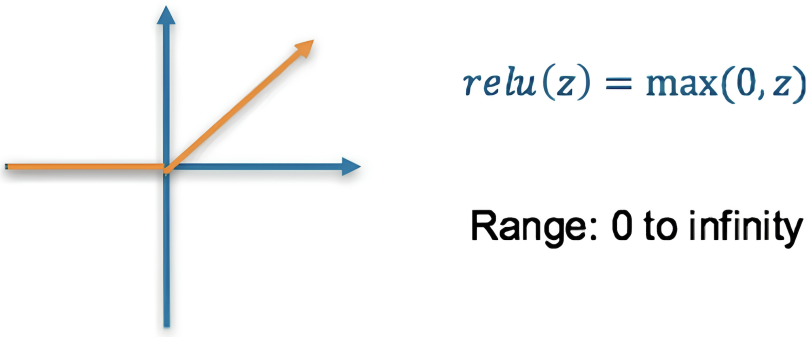

Figure 22: The ReLU Function (Rallabandi, 2023)

B.8 Spectral Normalization

Spectral normalisation is a method of normalising the weights of each layer by using a spectral norm $\sigma(W)$ so that the Lipschitz constant is equal to 1 for each layer and the entire network. It can renormalise the weights every time they are updated, thus mitigating the gradient explosion problem and improving training stability. (Hui, 2020)

B.9 Average Pooling

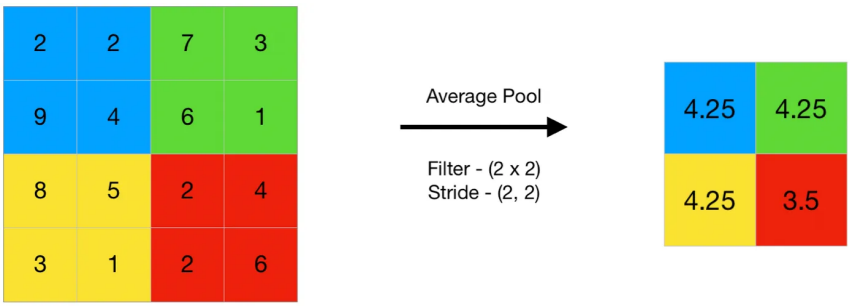

Figure 23: (Jain, 2024)

Pooling is downsampling the feature map while retaining important information. Average pooling is a type of pooling layer that computes the average value of the elements in the region covered in the feature map by the filter. (Jain, 2024)