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

References

Agrawal, N. (2023). *Decoding the Symphony of Sound: Audio Signal Processing for Musical Engineering*. https://towardsdatascience.com/decoding-the-symphony-of-sound-audio-signal-processing-for-musical-engineering-c66f09a4d0f5

Alencar, M. S., & da Rocha Jr, V. C. (2022). Speech coding. In *Communication systems* (pp. 97–134). Springer.

Anwar, A. (2020). *What is Transposed Convolutional Layer?* https://towardsdatascience.com/what-is-transposed-convolutional-layer-40e5e6e31c11

Brownlee, J. (2020). *Understand the Impact of Learning Rate on Neural Network Performance*. Machine Learning Mastery. https://machinelearningmastery.com/understand-the-dynamics-of-learning-rate-on-deep-learning-neural-networks/

Cai, L. (2012). Latent variable modeling. *Shanghai archives of psychiatry*, *24*(2), 118.

Chang, L.-C., & Hung, J.-W. (2022). A preliminary study of robust speech feature extraction based on maximizing the probability of states in deep acoustic models. *Applied System Innovation*, *5*(4), 71.

Chaubey, A. (2020). *Downsampling and Upsampling of Images — Demystifying the Theory*. Medium. https://medium.com/analytics-vidhya/downsampling-and-upsampling-of-images-demystifying-the-theory-4ca7e21db24a

Chen, N., Zhang, Y., Zen, H., Weiss, R. J., Norouzi, M., & Chan, W. (2020). Wavegrad: Estimating gradients for waveform generation. *arXiv preprint arXiv:2009.00713*.

Di Giorgi, B., Levy, M., & Sharp, R. (2022). Mel spectrogram inversion with stable pitch. *arXiv preprint arXiv:2208.12782*.

Dremio. (n.d.). *What is Learning Rate?* https://www.dremio.com/wiki/learning-rate/

Federal Agencies Digitization Guidelines Initiative. (n.d.). *Sampling rate (audio)*. https://www.digitizationguidelines.gov/term.php?term=samplingrateaudio

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, *27*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Henderson, E. (2019). *How to Evaluate Speech Technology Providers — 4 Key Considerations*. Medium. https://medium.com/speechmatics/how-to-evaluate-speech-technology-providers-4-key-considerations-3ddfdbdfdc5d

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems*, *33*, 6840–6851.

Hui, J. (2020). *GAN — Spectral Normalization*. Medium. https://jonathan-hui.medium.com/gan-spectral-normalization-893b6a4e8f53

Ito, K., & Johnson, L. (n.d.). *The LJ Speech Dataset*. Medium. https://keithito.com/LJ-Speech-Dataset/

Jain, A. (2024). *Pooling and their types in CNN*. Medium. https://medium.com/@abhishekjainindore24/pooling-and-their-types-in-cnn-4a4b8a7a4611

Ji, C., Bamunu Mudiyanselage, T., Gao, Y., & Pan, Y. (2021). A review of infant cry analysis and classification. *EURASIP Journal on Audio, Speech, and Music Processing*, *2021*, 1–17. 10.1186/s13636-021-00197-5

Jungil Kong, J. B., Jaehyeon Kim. (2020). *hifi-gan*. GitHub. https://github.com/jik876/hifi-gan

Katta, R. (2024). *Convolutional Neural Networks In Depth*. Medium. https://medium.com/@kattarajesh2001/convolutional-neural-networks-in-depth-c2fb81ebc2b2

Kong, J., Kim, J., & Bae, J. (2020). Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, *33*, 17022–17033.

Kumar, K., Kumar, R., De Boissiere, T., Gestin, L., Teoh, W. Z., Sotelo, J., De Brebisson, A., Bengio, Y., & Courville, A. C. (2019). Melgan: Generative adversarial networks for conditional waveform synthesis. *Advances in neural information processing systems*, *32*.

Le Roux, J., Wisdom, S., Erdogan, H., & Hershey, J. R. (2019). Sdr–half-baked or well done? *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 626–630.

Li, A., You, S., Yu, G., Zheng, C., & Li, X. (2022). Taylor, can you hear me now? a taylor-unfolding framework for monaural speech enhancement. *arXiv preprint arXiv:2205.00206*.

McFee, B. (2023a). Defining the stft. In *Digital signals theory*. CRC Press. https://brianmcfee.net/dstbook-site/content/ch09-stft/STFT.html?highlight=stft#defining-the-stft

McFee, B. (2023b). Spectral leakage and windowing. In *Digital signals theory*. CRC Press. https://brianmcfee.net/dstbook-site/content/ch06-dft-properties/Leakage.html#spectral-leakage-and-windowing

NI. (2023, September). Speech coding. In *Ni-scope*. Emerson. https://www.ni.com/docs/en-US/bundle/ni-scope/page/spectral-leakage.html

Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., & Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

Pei, J. (2010). Automatic speech recognition. *Vision Open Working Group, B rst edition*.

priyanka, P. (2023). *Evaluation Metrics for Speech(Audio) Signal Processing*. Medium. https://medium.com/@poudelnipriyanka/audio-metrics-their-importance-and-their-necessity-417950b0d848

Rallabandi, S. (2023). *Activation functions: ReLU vs. Leaky ReLU*. Medium. https://medium.com/@sreeku.ralla/activation-functions-relu-vs-leaky-relu-b8272dc0b1be

Rec, I. (2005). P. 862.2: Wideband extension to recommendation p. 862 for the assessment of wideband telephone networks and speech codecs. *International Telecommunication Union, CH–Geneva*, *41*, 48–60.

Roberts, L. (2020). *Understanding the Mel Spectrogram*. Medium. https://medium.com/analytics-vidhya/understanding-the-mel-spectrogram-fca2afa2ce53

Rocca, J. (2019). *Introduction to Markov chains*. Medium. https://towardsdatascience.com/brief-introduction-to-markov-chains-2c8cab9c98ab

Sahoo, S. (2018). *CResidual blocks — Building blocks of ResNet*. Medium. https://towardsdatascience.com/residual-blocks-building-blocks-of-resnet-fd90ca15d6ec

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., & Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. *International conference on machine learning*, 2256–2265.

Taal, C. H., Hendriks, R. C., Heusdens, R., & Jensen, J. (2011). An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Transactions on audio, speech, and language processing*, *19*(7), 2125–2136.

Taboga, M. (n.d.). Discrete fourier transform. In *Matrix algebra*. Statlect. https://www.statlect.com/matrix-algebra/discrete-Fourier-transform-amplitude-power-phase-spectrum

Tan, X., Qin, T., Soong, F., & Liu, T.-Y. (2021). A survey on neural speech synthesis. *arXiv preprint arXiv:2106.15561*.

University of Rhode Island and Inner Space Center. (n.d.). Introduction to phase. In *Discovery of sound in the sea*. https://dosits.org/science/advanced-topics/phase/

Vovk, I. (2020). *WaveGrad*. GitHub. https://github.com/ivanvovk/WaveGrad

Xuan, O. Z. (2023). *Exploring the Short-Time Fourier Transform: Analyzing Time-Varying Audio Signals*. Medium. https://medium.com/@ongzhixuan/exploring-the-short-time-fourier-transform-analyzing-time-varying-audio-signals-98157d1b9a12

Zvornicanin, E. (2024). *Relation Between Learning Rate and Batch Size*. https://www.baeldung.com/cs/learning-rate-batch-size