Road Construction Machinery Audio Analysis: A GAN-Based Synthesis Method

Jingteng Chen [1]

[1] State Grid Fujian Electric Power Co., Ltd. Putian Power Supply Company, Putian , Fujian, China.

Abstract. This study tackles the issue of limited construction machinery audio data, which restricts the development of early warning systems for protecting underground utilities during road construction. We propose a novel method for synthesizing audio data using GANs. By combining HIFI-GAN and BiLSTM networks, we develop a specialized model (H-GB) for audio synthesis. The model achieves a stable loss of 0.21, demonstrating high performance. Our experiments confirm the successful creation of high-fidelity audio data. This work enhances synthesized audio quality and the model's ability to learn construction-related audio features, providing a strong basis for advancing construction monitoring applications.

Keywords: Generative Adversarial Network，Road Construction，Audio Data Synthesis，BiLSTM.

1. Introduction

Road construction is crucial for urban development, impacting transportation and infrastructure. However, it poses risks to underground facilities, particularly power distribution networks, which are vital for urban electricity supply. Damage during construction can cause power outages and safety accidents, making the protection of these networks a key issue in urban planning and management [1].

To address this, researchers have explored machine noise classification for early warning systems. These systems classify construction machinery sounds to predict potential damage to underground facilities. However, traditional audio data collection is time-consuming, labor-intensive, and costly, limiting dataset diversity and system reliability [2].

With advancements in generative artificial intelligence, this study proposes using Generative Adversarial Networks (GANs) [4] to synthesize diverse audio data for road construction machinery, eliminating the need for extensive field collection. Additionally, a Bidirectional Long Short-Term Memory (BiLSTM) network [5] is integrated to enhance feature extraction, improving data quality and diversity through data augmentation. This approach not only overcomes data scarcity but also enables accurate and efficient noise monitoring, providing robust support for construction safety management and protecting underground power networks .

2. Related Work

Audio synthesis has advanced through three main approaches: signal processing techniques, autoregressive neural networks[6], and non-autoregressive models. The WORLD vocoder (2016) [7] excels in text-to-speech by converting acoustic parameters to raw audio signals. WaveNet [8] and WaveRNN ([9]) set industry standards for high-quality audio synthesis by leveraging temporal characteristics. Non-autoregressive models like WaveGlow (2019)[10] enable rapid synthesis but require extensive data and computational resources.

GANs have also revolutionized audio synthesis. Models like MelGAN [11] introduced innovative structures, while HIFI-GAN (2020) [12]set new benchmarks for efficiency and natural-sounding voices. In our research, we integrate HIFI-GAN [12] with a BiLSTM network [5] to handle complex audio data from construction machinery. This approach addresses limited datasets through data augmentation, ensuring robustness and reliability in our findings and providing advanced means for audio monitoring in road construction machinery.

3. Methods

3.1 Data Collection and Preprocessing

In this research, we developed a Python-based web crawler to collect raw audio data from various construction machinery online. We gathered over 100 audio samples (totaling 3GB) from diverse sources. To ensure data quality, we preprocessed the samples by cropping, resampling, normalizing volume, detecting endpoints, filtering frequencies, and converting them into Mel spectrograms (Figure 1). We also standardized the audio length to 4 seconds. These steps helped us build a comprehensive dataset of over 5,000 samples, providing a strong foundation for audio synthesis and analysis.

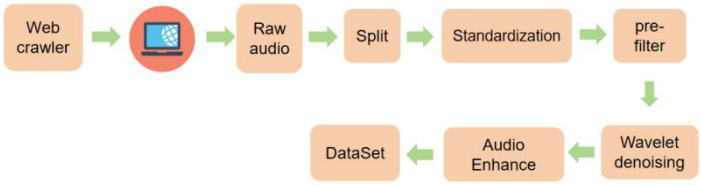

Fig. 1 Data Collection and Preprocessing

3.2 HIFI-GAN

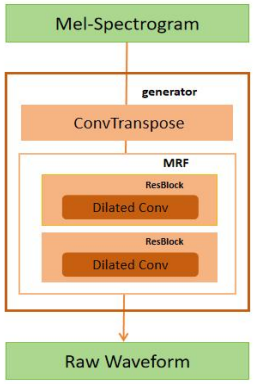

Fig. 2 HiFi-GAN Generator Network Architecture

HiFi-GAN [12] is a state-of-the-art GAN for audio synthesis, outperforming other open-source models and achieving near-human quality. In this study, we deploy HiFi-GAN using PyTorch, which includes a generator and two discriminators: Multi-Scale Discriminator (MSD) and Multi-Period Discriminator (MPD) [12].As illustrated in Figure 2, the role of the generator is to transform Mel-spectrogram features into actual waveforms. It uses a hybrid loss function that combines least squares, feature matching, and Mel-spectrogram losses, as specified by the subsequent equations:

$$L(\theta) = \frac{1}{N}\sum_{i=1}^{N}(y_i - y'_i)^2, \quad (1)$$

$$L_{Adv}(G; D) = E_s[(D(G(s))) - 1)^2], \quad (2)$$

$$L_{Mel}(G) = E_{(x,s)}[||\varphi(x) - \varphi(G(s))||_1], \quad (3)$$

$$L_{FM}(G; D) = E_{(x,s)}[\sum_{i=1}^{T}\frac{1}{N_i}||D^i(x) - D^i(G(s))||_1], \quad (4)$$

$$L_G = L_{Adv}(G; D) + \omega_{fm}L_{FM}(G; D) + \omega_{Mel}L_{Mel}(G), \quad (5)$$

Where N is the number of samples, is the true value of the i-th sample, is the predicted value of the i-th sample, and L is the average loss over the entire dataset, s represents the mel-spectrogram, x is the real input data, D(x) is the discriminator's score for the input, G(x) is the generator's output based on the mel-spectrogram, λ denotes the weights corresponding to the loss functions, and the total loss function for the generator is the sum of the three types of loss functions.

The HiFi-GAN principle underpins the construction of two discriminator networks, including the Multi-Period Discriminator (MPD) [12]. The MPD consists of stacked convolutional modules. It samples the original one-dimensional waveform at various intervals, converting it into a two-dimensional format, and then applies row-by-row convolution operations. Figure 3 shows the detailed process and architecture of the MPD network.

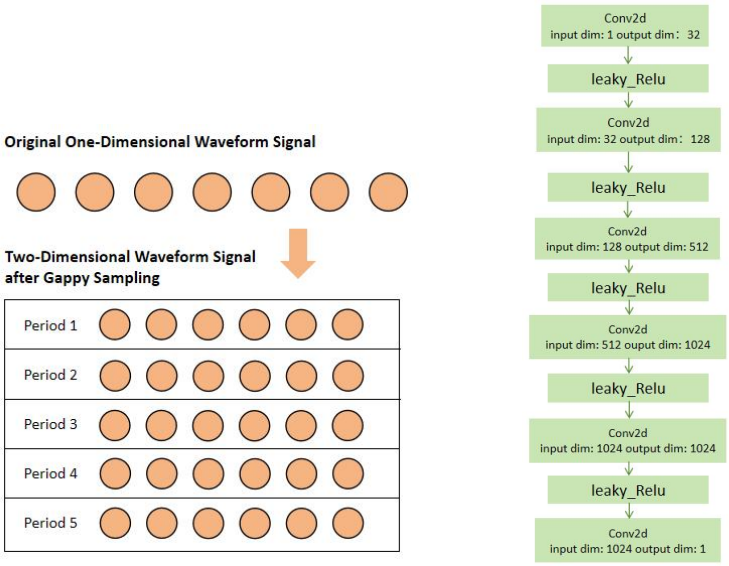

Fig. 3 Multi-Period Interval Sampling Illustration (Left) Network Architecture (Right)

The Multi-Period Discriminator (MPD) [12] captures waveform signals with interval sampling, yet it omits long-range dependencies due to the lack of continuous sampling. To counteract this, the Multi-Scale Discriminator integrates average pooling to distill data into a more concentrated form, succeeded by convolutional layers that refine feature extraction. The architecture is depicted in Figure 4.

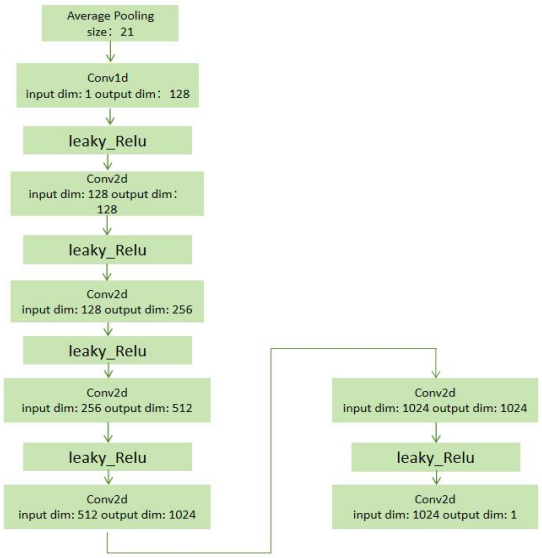

Fig. 4 Multi-Scale Discriminator (MSD) Network Architecture Diagram

The loss functions for both discriminators are consistent, with the objective of discerning the authenticity of generated data and real data. Therefore, it is merely necessary to simply sum the losses of the real audio and the generated audio. The formula is as follows:

$$L_{Adv}(D;\ G)\ =\ E_{(x,\ s)}[(D(x)\ -\ 1)^2\ +\ (D(G(s)))^2] \tag{6}$$

Where s represents the mel-spectrogram, x is the real data, D(x) is the discriminator's scoring function for the input, and G(s) is the output generated by the generator.

3.3 BiLSTM

Our study employs the BiLSTM[5] network, which, with its bidirectional processing capability, effectively addresses the challenge that traditional RNNs face in capturing long-range dependencies in sequential data analysis. The BiLSTM is highly regarded for its sensitive capture of contextual information in text analysis and time series forecasting, enabling a more comprehensive understanding of datasets. As shown in Figure 5, the BiLSTM network we constructed includes an input layer, two hidden layers, and a linear output layer, with specific parameter settings detailed in Table 1. This network demonstrates exceptional efficiency in extracting features from preprocessed Mel spectrograms, significantly enhancing our model's ability to recognize and learn audio features.

Table 1. Parameter List

Parameter	Values
input_dim	80
num_layers	2
hidden_dim	128
output_dim	32

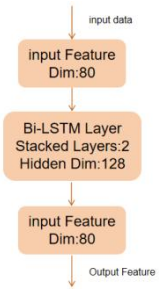

Fig. 5 BILSTM Feature Extraction Network Architecture

4. Results

4.1 Training Results

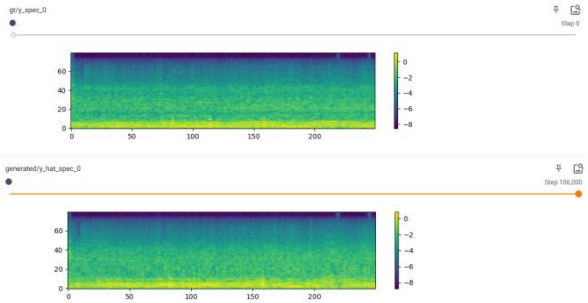

Fig. 6 Original Audio Frequency Spectrum (Top) Generated Audio Frequency Spectrum (Bottom)

Through extensive hyperparameter tuning and 100,000-step adversarial training, the model significantly reduced overall generator loss and Mel-spectrogram loss. It learned to generate realistic road construction machinery audio, closely matching the original audio's frequency spectrum (Figure 6). The discriminator's accuracy in distinguishing synthesized from genuine audio also improved. The generator's total loss stabilized at 30 (Figure 7), with the key Mel-spectrogram loss consistently at 0.21 (Figure 8).

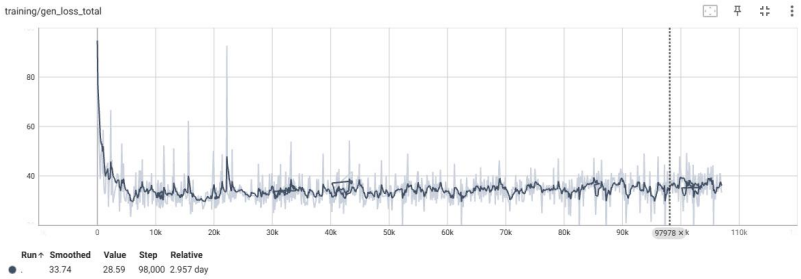

Fig. 7 Generator Loss Variation

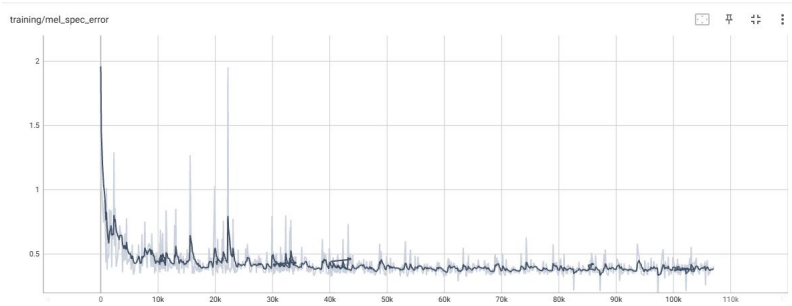

Fig. 8 Average Loss Variation of Mel-Spectrogram

4.2 Training Results

In this paper, we only demonstrate the inference results of the audio synthesis model for one of the most common types of machinery used in road construction, the tractor equipment. By loading the model parameters and importing the original audio data, the corresponding equipment audio data can be inferred and generated. We utilize the third-party library matplotlib to visualize the waveform of the original input audio data and the generated audio data, with the results shown in Figure 9.

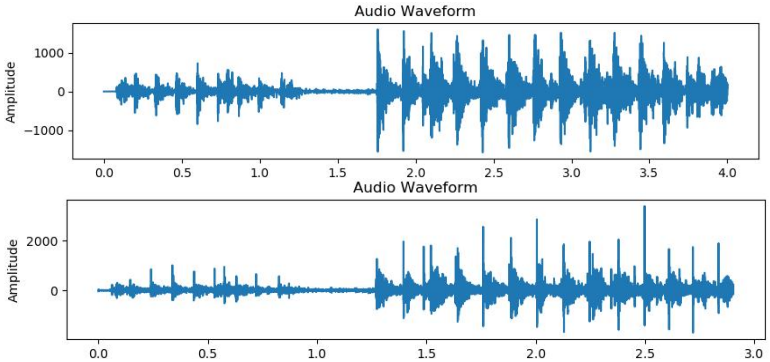

Fig 9 Original Audio Waveform (Top) Synthesized Audio Waveform (Bottom)

5. Summary

This study proposes a novel method for generating road construction machinery audio data using GANs combined with BiLSTM-based feature extraction. It addresses the challenge of limited audio data for training, offering a new approach for construction noise identification, hazard detection, and early warning systems. Future work will focus on improving the model's architecture to enhance audio diversity and quality, expanding its application to broader datasets, and exploring advanced generative AI algorithms to achieve higher fidelity and varied audio outputs.

Acknowledgments

Science and Technology Project of State Grid Fujian Electric Power Co., Ltd., "Research on Intelligent Analysis Technology for External Damage to Distribution Network Lines" (521320230005);The specific research fund of The Innovation Platform for Academician of Hainan Province (No.YSPTZX202145).

References

[1] Zhiqing Li, Chunhui Li, Zhe Zhang, etc Intelligent warning protection device for preventing external damage of transmission lines [J]. Electric World, 2023, 64 (02): 27-29

[2] Xiangyuan Chen, Wei Qin, Yanchi Liu, etc Audio recognition method for belt conveyor roller faults by integrating convolutional neural network and linear regression [J/OL]. Coal Science and Technology: 1-9 [2022-07-07] http://kns.cnki.net/kcms/detail/11.2402.TD.20240701.1327.001.html.

[3] Aamir Wali, Zareen Alamgir, Saira Karim, Ather Fawaz, Mubariz Barkat Ali, Muhammad Adan, Malik Mujtaba,Generative adversarial networks for speech processing: A review[J].Computer Speech & Language[J],2022,72(101308),0885-2308

[4] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., & Bengio, Y. (2014). Generative Adversarial Nets. Neural Information Processing Systems.

[5] Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF Models fo r Sequence Tagging. ArXiv, abs/1508.01991.

[6] Shi, Z. (2021). A Survey on Audio Synthesis and Audio-Visual Multimodal Processing. ArXiv, abs/2108.00443.

[7] MORISE, M., YOKOMORI, F., and OZAWA, K. World: A vocoder-based high-quality speech synthesis system for real-time applications. IEICE Transactions on Information and Systems,E99.D(7):1877 – 1884, 2016. doi: 10.1587/transinf.2015EDP7457.

[8] Van Den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N.,Senior, A. W., and Kavukcuoglu, K. Wavenet: A generative model for raw audio. SSW, 125, 2016.Amjady N. Short-term hourly load forecasting using time series modeling with peak load estimation capability. IEEE Transactions on Power Systems, 2001, 16(4): 798-805.

[9] Kalchbrenner, N., Elsen, E., Simonyan, K., Noury, S., Casagrande, N., Lockhart, E., Stimberg, F.,Oord, A. v. d., Dieleman, S., and Kavukcuoglu, K. Efficient neural audio synthesis. arXiv preprint arXiv:1802.08435, 2018.

[10] Prenger, R., Valle, R., and Catanzaro, B. Waveglow: A flow-based generative network for speech synthesis. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3617 – 3621. IEEE, 2019.Fangfang. Research on power load forecasting based on Improved BP neural network. Harbin Institute of Technology, 2011.

[11] Kumar, K., Kumar, R., Boissière, T.D., Gestin, L., Teoh, W.Z., Sotelo, J.M.,Brébisson, A.D., Bengio, Y., & Courville, A.C. (2019). MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis. Neural Information Processing Systems.

[12] Kong, J., Kim, J., & Bae, J. (2020). HiFi-GAN: Generative Adversarial N etworks for Efficient and High Fidelity Speech Synthesis. ArXiv, abs/2010. 05646.