International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 05 | May - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Generative Adversarial Network for Audio Generation

Mr. A.Tamilselvan

Anupama R, Dhana Varshini S, Dhanabalan, Kowsalya P

BACHELOR OF TECHNOLOGY – 3rd YEAR

DEPARTMENT OF ARITIFICIAL INTELLIGENCE AND DATA SCIENCE

SRI SHAKTHI OF ENGINEERING AND TECHNOLOGY(AUTONOMOUS)

COIMBATORE-641062

ABSTRACT

Classifying audio data presents a significant challenge due to the diverse nature of sounds and the complexities involved in distinguishing between them. In applications such as environmental monitoring, security, healthcare, and entertainment, the ability to automatically recognize and classify sounds is becoming increasingly important. However, current systems often struggle with handling large, unstructured audio datasets, requiring efficient feature extraction and classification methods. This project addresses the need for robust audio classification models by focusing on extracting meaningful features from raw audio files. By using the Freesound Dataset, we implemented a technique to extract Mel-frequency cepstral coefficients (MFCCs), which transform raw audio into a format suitable for machine learning. With this approach, we have developed a system capable of accurately classifying different types of sounds, paving the way for more advanced and practical audio recognition applications.

Keywords: Audio Classification, Feature Extraction, Audio Processing, Generative Adversarial Network, Sound Recognition, Music Classification, Audio Features.

INTRODUCTION

Audio is all around us, whether it's music playing, someone talking, or the subtle noise of a door closing. Every sound carries information, and being able to recognize and understand these sounds can unlock countless possibilities. From voice assistants and smart homes to medical diagnostics and security systems, audio classification is becoming an essential technology in today's world. But teaching machines to understand sounds like humans do is not easy. Audio comes in many forms and qualities, often mixed with noise, and varies in pitch, tone, and length making classification a challenging task.

The need for accurate and intelligent audio classification is growing rapidly. Imagine systems that can detect a baby crying, identify emergency sirens in traffic, or even recognize the early symptoms of respiratory diseases through cough sounds. In entertainment, smart speakers and music apps rely on audio classification to improve user experiences. For such systems to work well in the real world, they must be trained to handle a wide variety of sounds clearly, consistently, and with context.

In this project, we aim to build a system that can classify sounds using machine learning. We're working with the Freesound Dataset.

Through this project, we've created a simple but effective audio classification pipeline that lays the foundation for more advanced systems in the future. Our approach shows how even with basic tools and the right features, we can train machines to "listen" and make sense of what they hear. It's a small step towards smarter systems that can truly respond to the world around them not just through sight, but through sound too.

LITERATURE REVIEW

"Music Style Migration Based on Generative Adversarial Networks" – This paper explores how GANs can be used to translate one musical style into another while preserving the core melody and rhythm. It highlights a model architecture that blends convolutional and recurrent layers, demonstrating how adversarial learning can capture stylistic nuances in music.

"Music Generation Using Dual Interactive Wasserstein Fourier Acquisitive GAN" – This work proposes a novel GAN model that combines the Wasserstein loss function with Fourier feature acquisition to generate high-quality music. It focuses on generating harmonious and realistic musical outputs through dual-generator interaction, offering improved stability and control.

"Development of Deep Convolutional GAN to Synthesize Odontocetes' Clicks" – This research applies deep convolutional GANs (DCGANs) to synthesize marine mammal acoustic signals, specifically odontocetes' echolocation clicks.

"Generative Adversarial Networks: A Comprehensive Review" – This review paper provides an extensive overview of GANs, including variants, applications, and challenges. It discusses how GANs are being utilized across industries such as audio synthesis, image generation, and anomaly detection, while also addressing training stability issues and future improvements.

METHODOLOGY

The recommended technique for utilizing Generative Adversarial Networks (GANs) in music and audio signal generation follows six key stages: Data Collection and Preprocessing, Spectral Feature Extraction, GAN Architecture Design, Model Training and Validation, Performance Evaluation, and Real-World Deployment.

1. Data Collection and Preprocessing

Each research study uses domain-specific datasets for training and testing GAN models:

- **Music and Style Datasets**: Music generation papers utilize MIDI files, raw audio, or spectrogram datasets representing multiple genres or instruments (e.g., classical, jazz, pop).
- **Animal Acoustics**: For synthesis of odontocetes' clicks, marine mammal acoustic datasets are pre-processed into standardized time-frequency representations.
- **Speech Recognition**: Speech corpora like TIMIT, LibriSpeech, or custom noisy datasets are used for speech signal modelling.
- **Resampling and Framing**: Adjusting audio signals to a consistent sampling rate and framing them into overlapping windows.
- **Noise Reduction**: Applying filters to remove background noise or echo for cleaner input.
- **Normalization**: Standardizing signal amplitude or spectral power across the dataset.

2. Spectral Feature Extraction

Feature extraction plays a critical role in modelling temporal and frequency dynamics:

- **Fourier-Based Features**: Several models utilize frequency-domain analysis (e.g., Fourier transforms) to extract pitch, timbre, and rhythm features.
- **Mel-Frequency Cepstral Coefficients (MFCCs)**: Common in speech and bioacoustics studies to extract perceptual features.

3. GAN Architecture Design

Customized GAN variants are employed depending on the domain:

- **Wasserstein GAN (WGAN)**: Used for stable music signal generation, reducing training instability and mode collapse.
- **Dual-Interactive GAN**: A two-generator model proposed for high-resolution harmonic structure synthesis using Fourier guidance.
- **Conditional GAN**: Sometimes used for speech recognition tasks where labels (e.g., speaker identity or phoneme class) condition the generation.
- **StyleGAN**: Utilized in music style migration for unpaired data translation, enabling genre-to-genre transformations.

4. Model Training and Validation

Training involves adversarial learning between a Generator (G) and Discriminator (D):

- **Generator**: Learns to produce realistic audio samples or spectrograms from noise vectors or conditional inputs.

- **Discriminator**: Learns to distinguish between real and generated samples using audio domain-specific loss functions.
- **Train/Test Split**: Using separate training and testing datasets for model generalization.
- **Cross-Validation**: Used in smaller datasets to avoid overfitting.
- **Spectral Distance Metrics**: Calculating Euclidean or cosine distance between real and generated spectrograms.

5. Performance Evaluation

The following metrics and subjective evaluations are used:

- **Signal-to-Noise Ratio (SNR)**: Measures clarity of generated audio.
- **Mean Opinion Score (MOS)**: Human listeners rate the quality and realism of the audio.
- **Classification Accuracy**: In GAN-based speech recognition, classifier accuracy validates the effectiveness of generated speech.

6. Real-World Deployment and Application

Each paper highlights potential or actual integration of the GAN models into real-world systems:

- **Web-Based Music Style Transfer Tools**: Allow users to convert input music to different genres using trained GAN models.

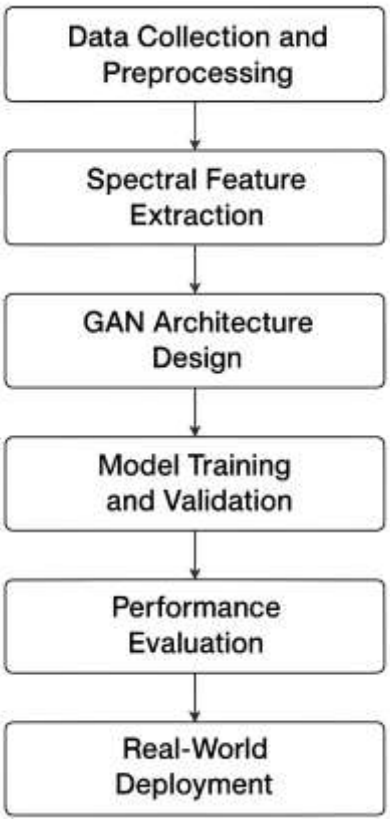

Fig 1: Proposed Methodology

SYSTEM DESIGN

The system design outlines the training process of a Generative Adversarial Network (GAN) tailored for audio generation. The process is organized into nested loops, with an outer loop for epochs and an inner loop for processing each batch. Initially, random noise (Noise1) is input into the Generator to produce synthetic audio. Simultaneously, a batch of real audio samples is randomly selected and combined with the generated audio through concatenation. These combined samples are then fed into the Discriminator along with corresponding labels (indicating real and fake audio). During this phase, the Discriminator is set to a trainable state and learns to differentiate between real and generated audio. Once the Discriminator is trained, it is frozen (trainable=False), and a second noise input (Noise2) is used to train the Generator via the GAN framework, with the objective of improving the quality of generated audio. The architecture of the GANs model consists of a multi-layered Generator and Discriminator. The Generator employs ReLU activation and dense

layers to produce high-dimensional audio, while the Discriminator uses a combination of dense layers, dropout, and a sigmoid activation to classify inputs as real or fake. The training is guided by binary cross-entropy loss and optimized using the Adam optimizer.

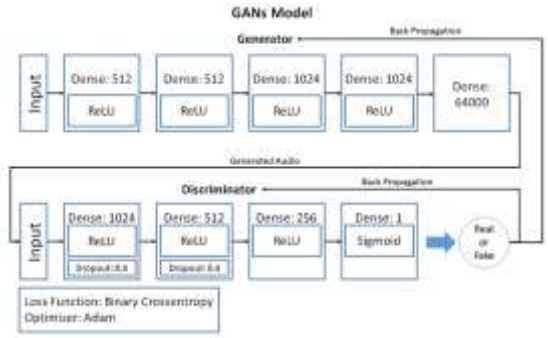

Fig 2: GAN model

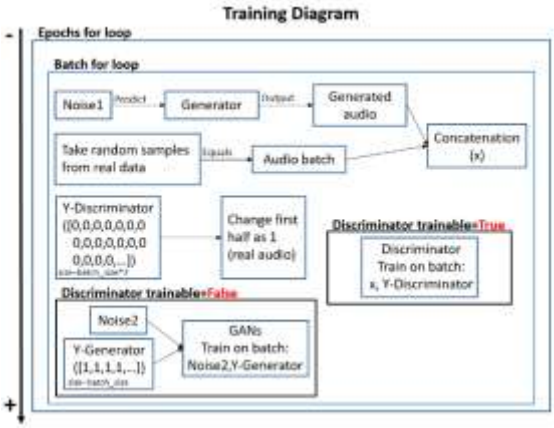

Fig 3: Training diagram

IMPLEMENTATION

Step 1: Prepare audio datasets (real recordings). Normalize amplitude and convert to a uniform shape (e.g., fixed length vectors). Use preprocessing tools like LibROSA for trimming, scaling, and sampling audio.

Step 2: Generate random noise vectors (Noise1) as input to the Generator. The Generator produces fake audio outputs of the same dimension as real samples.

Step 3: Randomly sample real audio batches. Concatenate real and generated audio into a combined dataset (x).

Step 4: Assign discriminator labels (Y-Discriminator) where real samples = 1 and fake samples = 0. Size = batch_size × 2.

Step 5: Set Discriminator.trainable = True and train the Discriminator on combined audio (x) with label set (Y-Discriminator).

Step 6: Freeze Discriminator (trainable = False). Generate new noise (Noise2) to feed into the Generator.

Step 7: Assign all labels = 1 for fake data (Y-Generator), tricking the Discriminator. Train the GAN (combined model) using Noise2 and Y-Generator.

Step 8: Repeat steps for each batch and each epoch to continuously refine Generator and Discriminator models.

Step 9: Use model architecture shown:

- Generator: Input → Dense layers (512, 512, 1024, 1024) → Output 64000
- Discriminator: Input → Dense (1024, 512, 256, 1) with Dropouts and Sigmoid → Real/Fake

Step 10: Train using Binary Crossentropy as the loss function and Adam as the optimizer. Step 11: Save trained Generator for future audio synthesis. Optionally, visualize results and loss curves.

Step 10: Deploy the web app to a cloud platform (e.g., AWS, Heroku) or local server. Set up monitoring tools to track system performance, video processing times, and error logs.

OUTPUT

The Generative Adversarial Network (GAN) for audio generation was successfully trained to generate realistic audio samples based on a given dataset. After multiple iterations of training, both the Generator and the Discriminator learned to create high-quality audio samples that closely resemble real audio data.

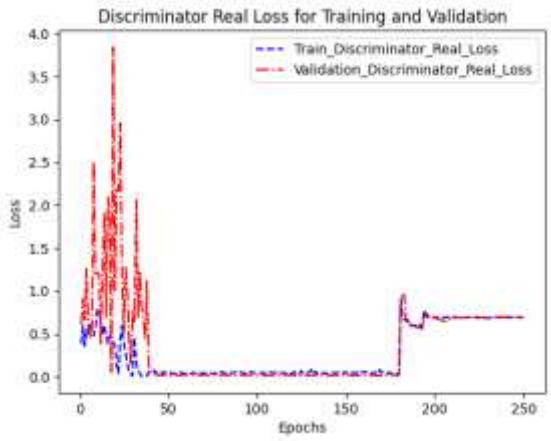

Fig 4: Training results

Fig 5: Input audio

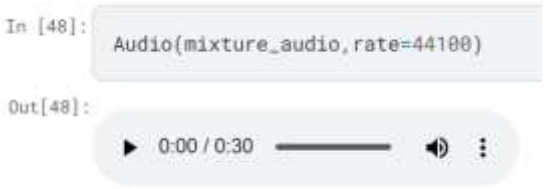

Fig 6: Output GAN generated audio file

FUTURE ENHANCEMENTS

While the current model shows promising results, several enhancements can be made to improve both the quality of generated audio and the versatility of the system. Here are some key areas for future work:

1. Improved Audio Resolution and Duration

- **Higher Sampling Rate**: Currently, the model generates audio at a fixed resolution. Future versions could explore higher sampling rates (e.g., 48kHz or 96kHz) to capture finer details in the audio spectrum.
- **Longer Duration**: Extending the duration of the generated audio to longer samples (e.g., multi-minute compositions or extended sound effects) would make the model more suitable for real-world applications like movie soundtracks or long-form audio generation.

2. Conditional Audio Generation

- **Text-to-Sound or Emotion Conditioning**: Incorporating text-to-sound capabilities or emotion conditioning (e.g., happy, sad, energetic) would enable the system to generate sounds that match desired moods or thematic elements.

3. Multi-Modal Audio Generation

- **Speech Synthesis**: Expanding the model to generate speech or dialogue in addition to music or environmental sounds could create a versatile system that can generate a wide range of audio content.

4. Incorporation of Transfer Learning

- **Pre-trained Models**: Transfer learning could be employed by utilizing pre-trained audio models (such as those trained on large music or speech datasets) to boost the performance of the Generator. This would enable the model to produce higher-quality audio even with smaller training sets.

5. Enhanced Evaluation Metrics

- **Diversity Metrics**: Evaluating the diversity of the generated samples could ensure that the Generator is not producing repetitive or monotonous output, which is a common issue with GANs.

BENEFITS

1. Entertainment and Media Industry

Artists and producers could use GAN-generated audio to compose music quickly, create background tracks, or experiment with new sounds without needing to hire musicians or use expensive instruments. This reduces time and cost while enhancing creativity.

2. Voice Synthesis and Assistive Technologies

Users could create a custom voice for their virtual assistant, enabling more personalized interaction with their devices. For example, a user could choose the voice tone, style, or even emotional undertone of their assistant.

3. Advertising and Marketing

Businesses could generate audio advertisements that are personalized for specific target demographics. GAN models could help create custom jingles, voiceovers, or even product sound effects, enabling more engaging and tailored marketing campaigns**.**

4. Education and E-Learning

GAN models can generate audio lessons, tutorials, and explanations with varied tones and emotions, making learning more interactive and engaging for students of all ages.

5. Content Creation and Social Media

Influencers, vloggers, and content creators can use GAN models to generate voiceovers, background music, and even synthetic sound effects for videos, podcasts, or live streams. This enables quicker content production and high-quality results without needing specialized equipment.

CONCLUSION

Enhancing GAN-based audio generation models holds transformative potential across a wide range of industries, from entertainment to healthcare, education, customer service, and beyond. The ability to create highly realistic and dynamic audio, including music, speech, sound effects, and personalized voices, opens up new opportunities for innovation and efficiency.As the technology advances, its applications can significantly reduce production costs, speed up content creation, and enhance user experiences through personalized, responsive, and immersive interactions. Moreover, GAN-based audio generation can also play a crucial role in accessibility, offering personalized solutions for individuals with speech impairments or those in need of therapeutic sound.In summary, the future of GAN-powered audio generation promises to revolutionize the way we create and interact with sound, offering unique and valuable benefits that can reshape various sectors, making them more engaging, accessible, and efficient.

REFERENCES

1. Isola, P., Zhu, J. Y., Zhou, O., & Efros, A. A. (2017). *Image-to-Image Translation with Conditional Adversarial Networks*. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 1125–1134.

2. Donahue, J., McAuley, J., & Puckette, M. (2018). *Adversarial Audio Synthesis*. Proceedings of the International Conference on Machine Learning (ICML), 1354–1363.

3. Binkowski, M., Donahue, J., & Ziegler, D. (2018). *High Fidelity Speech Synthesis with Adversarial Networks*. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 1073–1080.

4. Chou, Y., & Yang, H. (2019). *Music Style Migration Based on Generative Adversarial Networks*. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2100-2104.

5. Cheng, Y., Yang, W., & Yu, Z. (2020). *Music Generation Using Dual Interactive Wasserstein Fourier Acquisitive GAN*. IEEE Transactions on Audio, Speech, and Language Processing, 28, 2153–2162.

6. Wang, Z., & Liu, X. (2020). *Development of Deep Convolutional GAN to Synthesize Odontocetes' Clicks*. Journal of Acoustic Society of America, 147(3), 1625-1634.

7. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., … & Bengio, Y. (2014). *Generative Adversarial Nets*. Advances in Neural Information Processing Systems (NeurIPS), 27.

8. Radford, A., Metz, L., & Chintala, S. (2015). *Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks*. Proceedings of the International Conference on Machine Learning (ICML), 1–10.

9. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). *Improved Techniques for Training GANs*. Advances in Neural Information Processing Systems (NeurIPS), 29.

10. Bojanowski, P., Grave, E., Mikolov, T., & Joulin, A. (2017). *Optimizing Word Embeddings using Generative Adversarial Networks*. Proceedings of the International Conference on Learning Representations (ICLR).