Mathematical Models for Audio Generation and Processing Effects in Interactive Media

Junhao Sun *

St Paul's School, London, UK

* Corresponding Author Email: 1305223426@qq.com

Abstract. This paper provides a systematic review of mathematical models for audio generation and processing effects in interactive media applications. It first analyzes the core challenges of interactive audio in terms of real-time performance, controllability, and adaptability. Building upon this foundation, the paper focuses on dissecting three major technical approaches. Traditional methods, represented by physical modeling and procedural audio, offer computational efficiency and intuitive interactivity. Sample-based synthesis techniques, such as wavetable and granular synthesis, enable rich real-time variations while maintaining sound quality. Cutting-edge deep generative models and neural audio processing models deliver unprecedented generative diversity and audio fidelity, achieving high-level semantic control through conditional generation and latent space manipulation. Current research is driving interactive audio toward real-time, conditional, and semantic capabilities. Future success hinges on balancing high-performance models with stringent real-time and low-resource constraints.

Keywords: Interactive Audio Generation, Neural Audio Processing, Real-Time Diffusion Models.

1. Introduction

The emergence of interactive media has put audio technology under highly intense pressure as compared to before, when traditional media was applied. In technologies like video games, VR/AR, and the metaverse, audio has ceased to be an element of the background and has become a vital part of immersion. It should be able to react instantaneously to each user action and feedback at zero latency. At the same time, audio content should be dynamic and it must intelligently react to the plot, scene or user action by changing its emotional tone and content. Moreover, the system would be able to accommodate various users and equipment and provide them with personalized audio experiences. The resulting trend has been the formation of intelligent audio systems with the ability to fulfill real-time, dynamic and adaptable needs simultaneously, as an essential factor in the development of interactive media.

Traditional digital signal processing and physical modeling methods rely on established mathematical models. Examples include physical modeling synthesis—which simulates instrument sound production by solving wave equations—and various digital audio effects based on linear system theory, delay lines, and filters. These approaches offer high computational efficiency and timely response, enabling stable and reliable audio stream processing. Consequently, they are widely integrated into professional audio software and game audio engines for virtual instruments and real-time sound effects. Traditional static audio fails to meet modern demands, necessitating intelligent audio generation and processing technologies responsive to interactive signals (user input, game states, physical environments, etc.). In recent years, deep learning and generative models have emerged as the most active research directions in this field. These technologies leverage deep neural network models to learn complex audio patterns directly from data. Diffusion models have emerged as a mainstream solution due to their capability to generate high-quality, diverse audio. These models can produce music, sound effects, and speech based on textual descriptions, labels, or other control information. In interactive applications, the key challenge lies in achieving "real-time" and "conditional" model operation—ensuring sufficiently fast generation speeds while dynamically adjusting outputs based on real-time user input. In interactive music, musical structures can dynamically evolve in response to audience actions or emotional states. Within immersive installation

art and live performances, audio interacts in real time with motion, gestures, or other sensor data, becoming an integral part of the overall experience. Furthermore, AI audio tools significantly enhance efficiency and creative possibilities in audio post-production and interactive media development [1] [2].

This paper systematically reviews the mathematical models underpinning interactive media audio, focusing on how generative models and processing effect models couple with interactive control signals. Finally, it identifies current technical challenges and outlines future research directions.

2. Mathematical Foundations and Core Challenges of Interactive Audio

2.1. Temporality and Low-Latency Constraints

Interactive media impose extremely stringent real-time requirements on audio systems. To ensure immersive interactions remain intact, the total latency from user input to auditory feedback typically must remain below 20 milliseconds. This strict constraint stems from the human auditory-motor perception system. Excessive latency causes a disconnect in control perception and may even induce vertigo. This fundamental constraint directly excludes models with high computational complexity that cannot complete inference within a single audio buffer cycle. Consequently, model design must balance expressive capability with computational efficiency. Many state-of-the-art models excelling in offline tasks (such as early autoregressive models) struggle to be directly applied to real-time interactive scenarios due to their serial nature in sequence generation. This drives research toward more efficient network architectures or specialized hardware acceleration.

2.2. Controllability and Interpretability

In artistic creation, the value of audio models extends beyond generative quality to encompass controllability and interpretability. Sound designers and artists typically operate using high-level perceptual attributes like “tension,” “spatiality,” and “roughness” for creation, rather than hidden variables or weights directly tied to the model's internal mechanisms. A core challenge thus lies in establishing an effective, stable mapping between these intuitive, semantic control parameters and the model's underlying mathematical parameters.

3. Mathematical Models for Interactive Audio Generation

The core task of interactive audio generation lies in dynamically synthesizing high-quality audio in response to real-time, non-deterministic interactive inputs.

3.1. Procedural Audio and Physical Modeling

Procedural audio abandons the playback of static audio files, instead synthesizing sound in real time by algorithmically simulating the physical processes of sound production. This approach is inherently suited for interaction, as altering parameters of the physical process directly and predictably modifies the generated sound.

Modal synthesis centers on solving differential equations describing object vibrations. By directly adjusting physical parameters like mass, stiffness, and damping, it intuitively controls timbre and decay characteristics, establishing a direct correspondence between parameters and perceptual attributes.

Waveguide Synthesis Mathematical models based on delay lines and digital filters synthesize sustained musical tones by simulating wave reflection and dissipation at medium boundaries. This approach resolves the computational complexity issues of physical instrument modeling. Its high computational efficiency enabled real-time synthesis of sounds like guitars and orchestral instruments on contemporary hardware, finding widespread use in electronic music and early game audio.

Particle models target more complex continuous acoustic events like shattering, friction, and flowing water. By assigning simple amplitude envelopes and spectral characteristics to each “particle”

while controlling emission density and rules, this model generates exceptionally rich, non-repetitive textural sounds. It resolves the limitations of traditional sample playback in handling prolonged, random sound variations, providing an efficient solution for environmental sound effects and foley in games.

Interactivity Analysis The interactivity of physics-based modeling approaches is highly intuitive. Users or programs drive the model by altering physical parameters—such as applied forces or material properties—or environmental conditions. The model then generates corresponding acoustic outputs based on physical laws. This causal relationship ensures highly predictable and interpretable control while maintaining strong computational efficiency.

3.2. Sample-Based Synthesis and Splicing

Sample-based synthesis and splicing methods utilize pre-recorded high-quality audio samples. By triggering, morphing, and combining these samples in real-time, they generate diverse sounds tailored to interactive contexts.

Wavetable synthesis relies on cyclically reading a single-cycle waveform lookup table, smoothly shifting the read pointer position through linear or higher-order interpolation to achieve continuous pitch changes. This method overcomes issues like quantization noise during pitch shifting or the need for massive sample libraries in early samplers, enabling synthesis across a wide pitch range from a single waveform source. It has become one of the mainstream technologies in digital music synthesis.

Granular synthesis treats the audio stream as a sequence of "granules" lasting tens of milliseconds, each shaped by a time-domain window function (e.g., Gaussian window) applied to a small audio segment. By asynchronously scheduling these granules in real time and independently controlling their playback rate (pitch), amplitude, and spatial position, it generates diverse effects ranging from pitch shifting and time stretching to complex cloud-like textures. This model addresses the lack of dynamic variability in static samples across temporal and spectral dimensions, offering unprecedented control over microstructural reorganization of sound. It significantly enriches the sonic landscapes of interactive music and sound effects.

Sample-based methods achieve control by mapping interactive signals to sample playback parameters (e.g., wavetable index, playback rate, particle density). Their advantage lies in generating rich variations from high-quality real recordings at low computational cost, making them dominant in video game audio middleware. However, their variation range is constrained by the underlying sample library, making it difficult to generate entirely new timbres not present in the source library.

3.3. Deep Generative Models

Deep generative models aim to learn the underlying probability distribution of audio signals from large-scale data, enabling them to generate novel, high-quality audio samples based on conditional inputs.

Autoregressive models addressed the unnatural sound quality of traditional parametric speech synthesis, producing speech that approached human-level quality at the time and demonstrating the feasibility of directly modeling raw waveforms. However, their sequential generation nature resulted in extremely slow inference speeds, severely limiting direct application in real-time interactive scenarios.

Generative Adversarial Networks (GANs) learn data distributions through an adversarial game between a generator and a discriminator. In the audio domain, Donahue et al. 's work (WaveGAN) demonstrated GANs' potential for unconditional generation of short audio clips. Its value lies in providing a generative approach with a single-step forward pass, enabling fast inference [3]. More importantly, Karras et al.'s StyleGAN demonstrated that its latent space possesses decoupling properties, making it possible to control the macro-level attributes of generated audio by manipulating "style vectors." This provides a structured interface for interactive control [4].

Diffusion models have emerged as the state-of-the-art in generative fields in recent years. These models progressively add noise to data through a forward process, then learn a reverse denoising

process. Their core interactive mechanism lies in guidance—such as classifier-guided or classifier-free guidance proposed by Dhariwal & Nichol—which directs the denoising process using conditional information (e.g., textual descriptions, spectral features) to control generated content. This addresses the challenge of precisely controlling generated audio semantic properties, providing a robust mathematical framework for high-level semantic-based interactions.

Stream Matching/Consistency Models represent an innovative approach to overcome the slow sampling speed of diffusion models [5].

Variational Autoencoders (VAEs) learn to encode input data into a continuous, low-dimensional latent space. In audio, Engel et al. 's NSynth work demonstrates VAE capabilities in learning and interpolating instrument timbres. This model enables smooth, controllable transitions between distinct sounds. Through linear interpolation or nonlinear exploration in the latent space, users can create continuous timbral variations absent in the original data [6].

The strength of deep generative models lies in their near-unlimited generative diversity and high fidelity. Their interactivity is primarily achieved through two interfaces. Conditional generation treats interactive signals as conditional inputs to the model (e.g., using interactive event embeddings as guidance in diffusion models), enabling end-to-end mapping from instructions to audio. Latent space exploration allows users to directly navigate and manipulate within a low-dimensional, structured space, achieving smooth, continuous control over sound properties. The core challenge currently lies in balancing the exceptional performance of large models with the demanding requirements of real-time and low-resource interactive applications.

4. Mathematical Models for Interactive Audio Processing Effects

The core of interactive audio processing lies in intelligently altering existing sound properties in real time to adapt to dynamically changing contexts.

4.1. Classic Effects Based on Digital Signal Processing

Classic digital signal processing (DSP) effects, based on differential equations and filters, form the algorithmic foundation of real-time audio processing. They directly modify audio samples through deterministic mathematical operations, offering exceptional computational efficiency and predictability.

The core of filtering and equalization lies in linear time-invariant system theory. The design of infinite impulse response (IIR) and finite impulse response (FIR) filters addresses the selective enhancement or attenuation of specific audio frequency bands. Their advantage lies in extreme computational efficiency—a second-order IIR filter (i.e., a dual-second-order filter) can implement a peak or shelf equalizer, enabling multi-band parametric equalization in resource-constrained interactive systems.

The physical foundation of delay and reverb simulates the reflection and mixing of sound waves in space. This model addresses the challenge of approximating complex room acoustics in the digital domain, offering the advantage of adding spatial presence and richness to dry audio with minimal computational overhead.

Modulation effects like amplitude modulation (AM) and frequency modulation (FM) operate on the mathematical principle of controlling one signal's parameters with another signal.

Nonlinear processing, such as wave shaping distortion, employs mapping functions to directly distort the input waveform. This technique addresses the challenge of adding harmonic saturation or aggressive distortion effects to sound. Its advantage lies in the ability to drastically alter a sound's tonal characteristics through simple mathematical transformations, providing powerful timbre-shaping tools for interactive audio.

The interactivity of classic DSP effects is direct and explicit. Their algorithmic parameters—such as filter cutoff frequency, delay feedback amount, or LFO rate—naturally serve as “knobs” for

interactive control. This one-to-one mapping enables artists and programmers to manipulate sound with precision and predictability.

4.2. Analysis-Based Intelligent Audio Processing

Analysis-based intelligent audio processing methods no longer directly manipulate samples. Instead, they first decompose audio into meaningful acoustic features, then modify these features and resynthesize them to achieve more sophisticated and intelligent processing effects.

Classic mathematical models of vocoders, such as channel vocoders, typically obtain the spectral envelope of an analyzed source signal through short-time Fourier transforms or filter banks. This envelope is then used to shape the spectrum of a carrier signal. This model solves the problem of imparting the timbre characteristics of one sound to another. Its advantage lies in achieving unprecedented timbre transformation capabilities, making it widely used in voice morphing, robotic speech synthesis, and music composition.

Source-Filter models represent speech and many instrumental sounds as the convolution of a glottal wave or noise with a resonant system. This approach resolves the challenge of independently controlling pitch (by modifying the excitation source) and timbre (by modifying the resonant peaks). Its advantage lies in enabling decoupled control over perceptual sound attributes, making it possible to alter a speaker's vowels or “shape” an instrument's timbre in real-time during interaction.

Analysis-based methods elevate interaction to a higher semantic level. Interaction targets can be decomposed into modulations of analytical parameters, such as reducing energy in specific frequency bands or increasing reverb time. This enables systems to perform adaptive, intelligent processing based on the audio content itself, rather than simple global parameter adjustments.

4.3. Neural Audio Processing

Leveraging the powerful nonlinear mapping capabilities of neural networks to replace or augment traditional DSP algorithms, enabling more complex audio transformations or higher-fidelity sound effect simulations.

Neural vocoders represent one of the most successful applications of neural audio processing. Follow-up work on van den Oord et al.'s WaveNet employs it as a vocoder, reconstructing high-quality original waveforms from low-dimensional spectral features [7]. This model addresses the course, unnatural sound quality issues of traditional source-filter vocoders. Its advantage lies in synthesizing near-original audio quality from highly compressed spectral features, making it an indispensable component in modern real-time text-to-speech (TTS) and voice conversion systems. It significantly enhances the realism of synthesized speech in interactive systems.

End-to-end effect modeling aims to directly learn input-output mappings of hardware effects or complex processing chains using neural networks. Research by Wright et al. demonstrated the ability of recurrent neural networks (RNNs) or convolutional neural networks (CNNs) to precisely emulate classic hardware equalizers, compressors, and guitar amplifier distortion. This approach addresses the difficulty of accurately modeling complex nonlinear effect chains with explicit mathematical formulas. Its advantage lies in the ability to “clone” the tonal characteristics of any complex system, providing novel tools for sound design [8].

Real-time audio synthesis networks, exemplified by Engel et al.'s differentiable digital signal processing (DDSP), embed classical, interpretable DSP modules—such as harmonic additive synthesizers, white noise sources, and tunable filters—into an auto-differentiation framework. A compact neural network predicts the parameters of these DSP modules based on input control signals (e.g., pitch). This framework addresses the poor controllability and high computational overhead of pure neural network models. Its advantage lies in combining the expressive power of neural networks with the controllability, efficiency, and real-time capability of DSP modules. This enables users to intuitively control a powerful synthesizer through high-level semantic features (e.g., “timbre brightness”), offering exceptional interactivity [9].

Interactive interfaces for neural audio processing models exhibit diverse approaches. Neural vocoders achieve rapid, controllable mapping from compact features to high-fidelity waveforms; end-to-end effect networks enable training and tuning based on "target sound" references; while DDSP-type models bridge high-level semantics to underlying synthesis parameters. Collectively, these models propel interactive audio processing toward smarter, more realistic, and more creative directions. However, their computational costs and "black-box" nature remain factors requiring careful consideration in practical applications.

5. Conclusion

This paper systematically investigates mathematical models for audio generation and processing technologies tailored for interactive media. To meet real-time, dynamic, and adaptive interaction demands, the field has evolved from traditional physical modeling and digital signal processing toward advanced techniques such as deep generative models and neural audio processing. By incorporating interactive signals as conditional inputs or for latent space exploration, these models achieve high-quality real-time audio generation and intelligent processing. Interactive audio technology is rapidly advancing toward real-time, conditional, and semantic capabilities. Nevertheless, there are still severe issues with the deployment of such advanced models. To begin with, a rigid trade off should be implemented between the very high performance of the model and the high requirements of the interactive application in terms of real-time and low-latency. Second, the model interpretability and controllability should be strengthened, to correspond to artists intuitive capabilities of creative semantics, and to create high-level semantic instructions smoothly turned into model parameters. Lastly, the intensity of computational resources needs to be greatly minimized in order to allow such intricate intelligent audio systems work on the edge computerized environments, such as VR/AR headsets and mobile devices, and thus large-scale commercial implementation of the technology.

References

[1] Liu Yao. Research on Real-Time Audio Processing Technology. Deep Research, 2025 (02): 1 - 15.

[2] Andreev A. Generative Models for Speech Enhancement: From Statistical Heuristics to Data-Driven Sampling. HSE Summary Papers, 2024: 1 - 12.

[3] Donahue C, Mcauley J, Puckette M. Adversarial Audio Synthesis. 2018.

[4] Karras T., Laine S., Aila T. A Style-Based Generator Architecture for Generative Adversarial Networks. Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019: 4401 – 4410.

[5] Dhariwal P, Nichol A. Diffusion models beat GANs on image synthesis, Advances in Neural Information Processing Systems 34. 2021: 8780 - 8794.

[6] Engel J, Resnick C, Roberts A, et al. Neural audio synthesis of musical notes with WaveNet autoencoders, International Conference on Machine Learning. PMLR, 2017: 1068 – 1077.

[7] Van den Oord A, Dieleman S, Zen H, et al. WaveNet: A Generative Model for Raw Audio, Proc. 9th ISCA Speech Synthesis Workshop. 2016: 125.

[8] Wright A, Välimäki V. Real-time black-box modelling with recurrent neural networks: Learning guitar amplifiers and distortion circuits, Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). IEEE, 2019: 268 - 272.

[9] Engel J, Resnick C, Roberts A, et al. DDSP: Differentiable Digital Signal Processing, International Conference on Learning Representations. 2020.