State-of-the-Art Models for Generating Synthetic Sounds

The models that are commonly used for generating synthetic sounds, including bird sounds, include:

- WaveNet
- Generative Adversarial Networks (GANs) like WaveGAN and SpecGAN
- Diffusion Models
- Variational Autoencoders (VAEs)
- Recurrent Neural Networks (RNNs) based models like LSTMs and GRUs

Overview and Performance of Generative Models

Overview:

- **WaveNet**: Developed by DeepMind, WaveNet is a deep generative model that generates raw audio waveforms, known for producing high-quality, natural-sounding audio. It is computationally expensive and slow to generate audio, especially for long sequences.
- **GANs (Generative Adversarial Networks)**: GANs, particularly Conditional GANs (cGANs), have been adapted for audio generation. They are faster than some models like WaveNet for inference, but training can be challenging and unstable.
- **Diffusion Models**: These probabilistic models create realistic samples by simulating a diffusion process. They have shown impressive results in generating high-quality audio but are computationally intensive to train.
- **VAEs (Variational Autoencoders)**: VAEs are easier to train than GANs and can provide good quality outputs. They may not be as sharp or high fidelity as GANs or WaveNet but offer a balance of quality and complexity.
- **RNN-based Models (e.g., LSTM, GRU)**: Recurrent Neural Networks are good at capturing temporal dependencies but may not perform as well for generating high-fidelity audio compared to models like WaveNet or GANs.

Performance:

- For high-quality and natural sound generation, **WaveNet** and **Diffusion Models** are ideal due to their ability to model raw audio waveforms and probabilistic sampling, respectively.
- **GANs** provide a balance of quality and computational efficiency, suitable for quicker outputs with slightly compromised quality.
- **VAEs** offer an easier training process and reasonable output quality, making them a practical choice for generating variations of known sounds.
- **RNNs** are suitable for tasks requiring understanding of temporal dependencies but are less effective for high-fidelity audio synthesis.

Articles on Performance of Generative Models for Synthetic Audio

Given links are for more detailed information on the performance of different generative models in the context of Audio:

1. Synthesizing Soundscapes: Leveraging Text-to-Audio Models for Environmental Sound Classification
   This article explores the use of various generative models, including AudioLDM2 and AudioGen, for generating synthetic environmental sounds.
2. Adversarial Audio Synthesis
   This paper discusses the use of GANs like WaveGAN and SpecGAN for audio synthesis, highlighting their effectiveness and challenges in generating high-quality synthetic audio.
3. Sound Model Factory: An Integrated System Architecture for Generative Audio Modelling
   This research covers an integrated approach combining GANs and RNNs for synthetic audio generation, providing insights into the benefits of using multiple model architectures.
4. A Comprehensive Survey on Diffusion Models and Their Applications
   This survey provides an overview of diffusion models, their theoretical foundations, and their application to audio synthesis among other fields.
5. Deep Generative Models for Musical Audio Synthesis
   This article discusses deep generative models, such as WaveNet, for audio synthesis and their ability to produce high-fidelity sound through conditioning strategies.