Latent Diffusion Model for Audio: Generation, Quality Enhancement, and Neural Audio Codec

Haohe Liu **Wenwu Wang** **Mark D. Plumbley**
Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey

Abstract

In this demo, we explore the versatile application of Latent Diffusion Models (LDMs) in audio tasks, showcasing their capabilities across three state-of-the-art systems: AudioLDM-2 for text-to-audio generation, AudioSR for audio quality enhancement, and SemantiCodec for ultra-low bitrate neural audio coding. AudioLDM-2 employs an LDM to decode high-quality audio from intermediate Audio Masked Autoencoder (AudioMAE) features, which are generated using a continuous language model conditioned on textual input. AudioSR leverages an LDM to perform robust audio super-resolution, enhancing the quality of low-resolution audio across various types and bandwidths, from speech and music to general sounds. SemantiCodec utilizes an LDM to efficiently decode audio from semantically rich, low-bitrate representations, demonstrating effective audio compression. Together, these systems illustrate the broad utility of LDM as audio decoder for diverse audio generation, enhancement, and neural audio codec tasks. This report highlights the significance of these innovations and outlines our demo objectives.

1 Introduction

As shown in Figure 1, this demo presents three state-of-the-art systems that leverage latent diffusion models (LDMs) [10] for various audio tasks: AudioLDM-2 [7] for text-to-audio generation, AudioSR [4] for audio quality enhancement, and SemantiCodec [6] for neural audio coding. These models demonstrate the versatility of LDMs in different audio applications, addressing challenges in audio generation, super-resolution, and compression.

The demo will showcase the following systems:

- **AudioLDM-2**: A text-to-audio generation model that transforms textual descriptions into high-fidelity audio across a wide range of audio types. Besides using LDM, AudioLDM-2 is also the first to integrate continuous language modelling with GPT-2 [8] into a framework that utilizes self-supervised pretrained audio features for generating audio. At the time of publication, AudioLDM-2 demonstrated state-of-the-art performance across text-to-audio, text-to-speech, and text-to-music tasks, demonstrating its cutting-edge capabilities in the audio generation task.
- **AudioSR**: An LDM-based audio super-resolution model that enhances the quality of low-resolution audio by predicting high-frequency details, improving audio fidelity for various input sampling rates and types. AudioSR is the first system to perform audio super-resolution across flexible input bandwidths and audio types. AudioSR has also been demonstrated as an effective post-processing module by improving the output of various audio generation models, including FastSpeech-2 [9], AudioLDM [5], and MusicGen [2]. AudioSR has been adopted as the default post-processing system in state-of-the-art audio generation systems such as FireRedTTS [3].

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

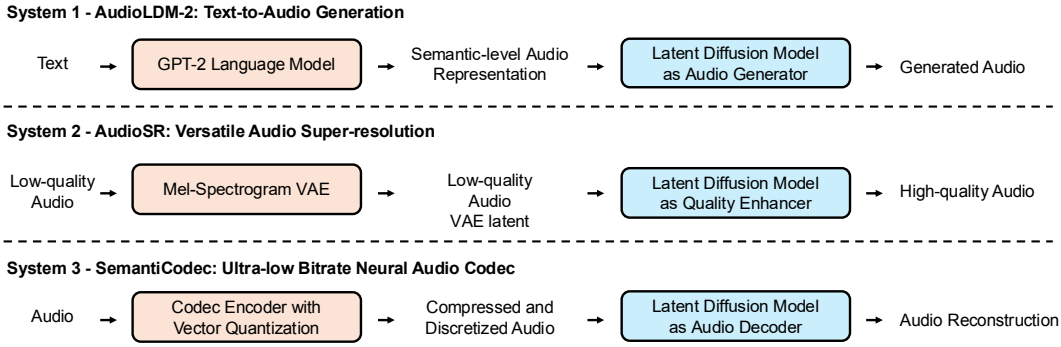

Figure 1: Overview of three audio processing systems using Latent Diffusion Models (LDMs). Each system utilizes LDMs differently: AudioLDM-2 for text-to-audio generation, AudioSR for enhancing low-quality audio through super-resolution, and SemantiCodec for efficient audio compression and reconstruction.

- **SemantiCodec**: A neural audio codec that efficiently compresses audio into a low-bitrate representation while retaining rich semantic information, potentially enhancing the performance of downstream tasks such as audio language modeling [1]. SemantiCodec is the first system to achieve ultra-low bit rate (0.3-1.4 kbps) and ultra-low token rate (supporting 25, 50, 100 tokens per second) audio compression for open-domain audio content. SemantiCodec is also the first system to employ an LDM as the decoder in a neural audio codec, showcasing the effectiveness of diffusion-based generative models in building neural audio codec.

These three systems will be deployed on HuggingFace for demonstration, where users can interact with the models online by generating audio from text, enhancing low-resolution audio, and compressing and reconstructing audio. Section 2 outlines the key objectives of the demo and the configuration settings available for each model.

2 Demo Objectives

The demo aims to showcase the following capabilities:

Text-to-Audio Generation with AudioLDM-2, allowing users to directly convert textual descriptions into a variety of audio outputs, such as environmental sounds, music, or speech. For example, by inputting a phrase like *A ghostly choir chanting hauntingly beautiful hymns*, the system can instantly generate the corresponding audio. By interacting with the system, users will experience the capabilities of AudioLDM-2, including how textual prompts influence model performance and the quality of audio generation, potentially sparking new ideas for audio creation tasks.

Audio Super-Resolution with AudioSR, which can handle flexible input sampling rates (e.g., from 2kHz to 16kHz bandwidth) and upsample audio to 24kHz bandwidth with a 48kHz sampling rate. In our demo, users can upload their audio or manually degrade an arbitrary audio file, then restore its quality using AudioSR. This interactive process allows users to understand how AudioSR can enhance audio quality, recognize any artefacts it might introduce, and identify conditions under which AudioSR works the best or might not perform optimally.

Efficient Audio Compression and Reconstruction with SemantiCodec, showcasing the efficiency of LDMs in neural audio coding by compressing audio into a semantically rich, low-bitrate representation and reconstructing the original audio with high fidelity. The systems support various bitrates between 0.3 kbps and 1.4 kbps and can work with multiple token rates, including 25, 50, and 100 tokens per second. Through interaction with our SemantiCodec demo, users can understand the trade-offs between token rate, bitrate, and audio reconstruction quality. This experience will also demonstrate how various encoding parameters influence the fidelity of the decoded audio, offering insights into the efficiency and effectiveness of neural audio codecs.

Acknowledgment

We acknowledge the support provided by the China Scholarship Council during a visit of Jisheng Bai to Nanyang Technological University. This research was partly supported by Engineering and Physical Sciences Research Council (EPSRC) Grant EP/T019751/1 "AI for Sound", a Research Gift from Adobe, and a PhD scholarship from the Centre for Vision, Speech and Signal Processing (CVSSP) at the University of Surrey and BBC R&D.

References

[1] Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. AudioLM: A language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 42:2523–2544, 2023.

[2] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Defossez. Simple and controllable music generation. In *Advances in Neural Information Processing Systems*, volume 36, pages 47704–47720, 2023.

[3] Hao-Han Guo, Kun Liu, Fei-Yu Shen, Yi-Chen Wu, Feng-Long Xie, Kun Xie, and Kai-Tuo Xu. FireRedTTS: A foundation text-to-speech framework for industry-level generative speech applications. *arXiv preprint arXiv:2409.03283*, 2024.

[4] Haohe Liu, Ke Chen, Qiao Tian, Wenwu Wang, and Mark D Plumbley. AudioSR: Versatile audio super-resolution at scale. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1076–1080, 2024.

[5] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. AudioLDM: Text-to-audio generation with latent diffusion models. *International Conference on Machine Learning*, 2023.

[6] Haohe Liu, Xuenan Xu, Yi Yuan, Mengyue Wu, Wenwu Wang, and Mark D Plumbley. SemantiCodec: An ultra low bitrate semantic audio codec for general sound. *arXiv preprint:2405.00233*, 2024.

[7] Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D. Plumbley. AudioLDM 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2871–2883, 2024.

[8] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 2019.

[9] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech. In *International Conference on Learning Representations*, 2021.

[10] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.