Enhancement in audio recording using diffusion model

A Project Based Learning Report Submitted in partial fulfilment of the requirements for the award of the degree

of

Bachelor of Technology

in The Department of AI and DS

Deep Learning 23AD2205A

Submitted by

2310080081: Ziya Parvez
2310080026: Pavan Kumar
2310080086: Rashmika G
2310080078: Eelapanti Mythri
2310080076: Sonali Pradhan

Under the guidance of

Dr. Gangamohan Paidi

Department of Artificial Intelligence and Data Science

Koneru Lakshmaiah Education Foundation, Aziz Nagar

Aziz Nagar – 500075

FEB - 2025.

Introduction

Sound improvement in audio recording is an indispensable process that attempts to enhance quality, clarity, and overall audio fidelity. Irrespective of music production, podcasting, voice communication, or archiving, high-fidelity audio contributes to successful transmission and listener experience. Conventional sound improvement tools like noise elimination, equalizing, and compressing have enjoyed extensive application as a means to counteract ambient noise, distortion, and varied frequency response. Nonetheless, these techniques tend to come short in addressing highly complex or severe audio degradations.

Progress in artificial intelligence and machine learning has brought novel techniques into audio enhancement. One of these is diffusion models, which have received extensive interest due to their capacity to produce high-quality audio through successive refinement of noisy or degraded signals. These models learn the latent architecture of clean audio and are well able to restore recordings that have been degraded by a wide range of types of noise and artifacts. Using vast amounts of data and advanced algorithms, diffusion models provide a new hope for obtaining even higher quality audio, even in very adverse circumstances. This introduction discusses the potential for these new techniques to transform the area of audio record improvement.

Literature Review/ Application Survey

Deep Generative Replay with Denoising Diffusion Probabilistic Models for Continual Learning in Audio Classification

Hyeon-Ju Lee and Seok-Jun Buu's work tackles the issue of constant learning for audio classification, where deep learning models tend to experience catastrophic forgetting when updating to novel audio classes. To resolve this, the authors introduce a Diffusion-Driven Generative Replay (DDGR) framework that utilizes Denoising Diffusion Probabilistic Models (DDPM) to create high-quality replay data for old tasks and a triplet network classifier to separate class-specific features. Experiments on the Audio MNIST and ESC-50 datasets illustrate the efficacy of the method, obtaining 95.45% and 72.99% mean incremental accuracy, respectively, and drastically decreasing catastrophic forgetting relative to baseline models such as GANs and VAEs. The DDGR framework provides an efficient solution for dynamic audio classification environments, which can be extended to areas of speech recognition, healthcare, and surveillance. Yet, the performance of the method is sometimes restricted by the suboptimal data generation quality. As future direction, improving data utilization with generated data and considering sophisticated representation learning methods could improve the performance of the model further. Overall, the work outlines an interesting approach to continual learning that optimizes between adaptability and knowledge preservation for audio classification tasks. [1]

Investigating the design space of diffusion models for speech enhancement

The paper by Philippe Gonzalez et al. discusses the application of diffusion models, which have proven themselves in image synthesis, to speech enhancement. The paper systematically examines the design space of diffusion models with regard to elements such as neural network preconditioning, weighting of the training loss, stochastic differential equations (SDEs), and levels of stochasticity in the inverse process. Employing datasets such as VoiceBank+DEMAND and a tailored MultiCorpus dataset, the authors compare their systems to baselines such as SGMSE and Conv-TasNet. Main takeaways are that the quality of diffusion-based speech enhancement is independent of the drift from noisy to clean speech, and optimized SDEs and sampling techniques (e.g., Heun-based sampler) lower computational expenses by a factor of four without compromising quality. The systems suggested outperform current diffusion-based and discriminative ones on performance metrics such as PESQ, ESTOI, and SNR gains. This work fills the gap in applications of diffusion models to speech versus images, providing some insights into building effective, high-performance speech enhancement systems. Nevertheless, there are still issues such as increased computational requirements and restricted generalizability to mismatched situations. Future research may investigate unsupervised training techniques and insensitivity to arbitrary distortions, further pushing the applicability of diffusion models in speech enhancement.[2]

Multi-Aspect Conditioning for Diffusion-Based Music Synthesis: Enhancing Realism and Acoustic Control

Ben Maman et al.'s work resolves the drawbacks of conventional music synthesis techniques, which tend to compromise between realism and fine-grained acoustic control. The authors introduce a diffusion-based music synthesis framework that improves realism and offers fine-grained control over acoustics, timbre, and style using multi-aspect conditioning. By conditioning on musical sheet (notes, instruments) and version-specific features (acoustics, performance settings) with Feature-wise Linear Modulation (FiLM) layers, the model has fine-grained control and high-fidelity synthesis. Trained on large-scale multi-instrument datasets with MIDI-aligned scores, the model surpasses current methods such as Hawthorne et al.'s T5 model both in qualitative listening tests and quantitative measures such as Fréchet Audio Distance (FAD). It also provides acoustic coherence and seamless transitions over extended musical passages. The method fills the gap between control and realism, and thus it is appropriate for use in music production, film scoring, and interactive audio systems. The drawback is high computational expense and limited generalization to novel musical genres. Future research may investigate the extension of the framework to genres such as jazz and pop, human singing voice synthesis with lyric conditioning, and more advanced controls for expressive performance aspects such as vibrato and dynamics. This work demonstrates the promise of diffusion models in pushing the boundaries of music synthesis.[3]

Blind audio bandwidth extension: A diffusion-based zero-shot approach

The paper by Eloi Moliner, Filip Elvander, and Vesa Välimäki presents BABE (Blind Audio Bandwidth Extension), a zero-shot method for high-frequency content restoration of audio recordings with unknown lowpass degradation parameters. Following up on the task of blind audio bandwidth extension (ABE), BABE uses pre-trained diffusion models to estimate degradation parameters iteratively and restore missing high frequencies through diffusion posterior sampling. Rated via objective measures (LSD, FD) and subjective listening trials (MUSHRA), BABE surpasses current state-of-the-art blind ABE systems and matches performance with informed (oracle) schemes. It is shown to generalise well to real historical records, greatly enhancing audio quality and yielding "good" quality restoration. As a zero-shot blind ABE solution, BABE avoids task-dependent training and is more interpretable by directly modelling degradation. This method has great promise for historical audio restoration, making past recordings sound more natural. Its future applications could include improving robustness, dealing with more complicated degradations, and expanding its use to other than music. BABE is a new and powerful use of diffusion models in audio restoration, closing the gap between generative AI and real-world audio enhancement tasks.[4]

Causal Diffusion Models for Generalized Speech Enhancement

The paper "Causal Diffusion Models for Generalized Speech Enhancement" introduces a causal speech enhancement system based on generative diffusion models to process a wide range of speech corruptions, including noise, reverberation, clipping, packet loss, bandwidth reduction, and codec distortion. Unlike conventional denoising methods, the model is real-time possible with a 20 ms algorithmic latency by adopting a U-Net-like architecture with causal convolutions, cumulative group normalization, and strided convolutions for downsampling. The model is trained with a task-adapted diffusion process, where a forward stochastic process progressively adds noise, and a reverse process, conditioned on a neural score model, generates clean speech with a denoising score matching loss. The authors construct a MultiCorruption dataset, with a variety of real-world degradations, and evaluate the model with intrusive (PESQ, POLQA, ESTOI, SI-SDR, LSD) and non-intrusive (DNSMOS, Wav-to-Vec MOS) metrics. Experiments compare specialized vs. generalized models and causal vs. non-causal setups, showing that specialized and non-causal models are slightly better but that the generalized causal model is competitive, making it suitable for real-world deployment. Although with high computational complexity, future optimizations—such as fewer diffusion steps—may be possible for real-time deployment.[5]

Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models

Make-An-Audio is a prompt-enhanced diffusion-based text-to-audio generation model to counter challenges such as limited high-quality text-audio pairs and complicated audio modeling. It proposes pseudo prompt enhancement for generating diverse concept compositions and applies a spectrogram autoencoder for effective audio representation with high-level semantic understanding. The model uses contrastive language-audio pretraining (CLAP) for profound language understanding, resulting in high-fidelity audio generation. Make-An-Audio obtains cutting-edge performance in both objective and subjective assessments, delivering better audio quality and precise text-audio alignment. It generalizes across various user-specified inputs, such as text, audio, images, and video, to support diverse audio content generation. This allows for customized transfer and fine-grained control, which boosts creative potential. The model is a leader in high-definition, high-fidelity audio synthesis in various modalities and is an extremely effective device for producing realistic and varied audio experiences. Through closing the natural language-complex audio signal gap, Make-An-Audio creates new benchmarks for text-to-audio technology and paves the way for groundbreaking applications in multimedia content generation.[6]

Diffusion-based Unsupervised Audio-Visual Speech Enhancement

The paper "Diffusion-based Unsupervised Audio-Visual Speech Enhancement" presents a new approach that mixes diffusion models with non-negative matrix factorization (NMF) for unsupervised audio-visual speech enhancement (AVSE). The model starts by pre-training a diffusion model on clean speech conditioned on matching video features, learning a generative speech distribution. During testing, this model is combined with an NMF-based noise model, where an iterative update mechanism adjusts the noise parameters using a posterior sampling approach within the reverse diffusion process. A new efficient inference algorithm (UDiffSE+) is presented, with fewer iterations than the existing state-of-the-art methods, and much better efficiency. The system is evaluated on the TCD-TIMIT dataset in matched (known types of noise) and mismatched (unseen types of noise) conditions, with generalization improvements over supervised generative AVSE approaches such as FlowAVSE. Metrics such as SI-SDR, PESQ, and ESTOI confirm that the inclusion of visual lip movements enhances speech quality and intelligibility, particularly under low-SNR scenarios. UDiffSE+ also provides a good performance-computational efficiency trade-off, being 5× faster than existing diffusion-based models, and thus more practical for real-world deployment.[7]

AudioLDM: Text-to-Audio Generation with Latent Diffusion Models

The article "AudioLDM: Text-to-Audio Generation with Latent Diffusion Models" introduces an improved text-to-audio (TTA) generation model with improved generation quality and computational efficiency. In contrast to earlier models that need large-scale audio-text paired data, AudioLDM learns latent audio representations using a variational autoencoder (VAE) and synthesize audio conditioned on Contrastive Language-Audio Pretraining (CLAP) embeddings. The model is trained on audio embeddings without the use of text-audio pairs. Latent diffusion model (LDM) works on a compressed space, with the advantage of being more efficient without compromising audio quality. AudioLDM enables zero-shot text-guided audio manipulations such as style transfer, super-resolution, and inpainting. Performance on the AudioCaps dataset indicates that AudioLDM performs better than state-of-the-art TTA models such as DiffSound and AudioGen, on objective (Frechet Distance, IS, KL) and subjective (human rating) scores, with state-of-the-art performance using one GPU. The research claims that training with audio-only embeddings is more effective than training with text-audio pairs, showing a stable and efficient method for TTA generation and audio manipulation.[8]

Audio Generation with Multiple Conditional Diffusion Model

The paper "Audio Generation with Multiple Conditional Diffusion Model" proposes a multi-conditional text-to-audio (TTA) generation model with more control over generated audio in addition to text. Current TTA models have the capability to model fine-grained audio features only from text; hence, this paper utilizes timestamp, pitch contour, and energy contour as control conditions in addition to text. These are encoded by a trainable control condition encoder, supported by a large language model (LLM), while the fusion is achieved by a Fusion-Net into a latent diffusion model (LDM) without altering the pre-trained TTA model's weights. The authors propose a new dataset with audio, text, and control conditions and evaluation metrics for temporal order, pitch, and energy control. Experimental results indicate that the method improves the control accuracy of audio generation significantly and outperforms baseline models like AudioLDM and Tango in terms of semantic coherence preservation and fine-grained control assistance. The study confirms that multiple control condition training yields better performance and parameter efficiency, making the model suitable for application in virtual reality, video editing, and interactive systems.[9]

From Discrete Tokens to High-Fidelity Audio Using Multi-Band Diffusion

The article "From Discrete Tokens to High-Fidelity Audio Using Multi-Band Diffusion" introduces Multi-Band Diffusion (MBD), a novel diffusion-based model for high-fidelity audio synthesis from low-bitrate discrete representations. In contrast to traditional GAN-based models, which introduce artifacts and distortions, MBD processes audio in separate frequency bands separately, reducing error addition and improving perceptual quality. The technique includes a frequency equalization (EQ) processor, which equalizes energy between bands, and a power noise scheduler, which improves training efficiency. MBD is tested on speech, music, and environmental sounds and outperforms existing state-of-the-art models like HiFi-GAN, PriorGrad, and EnCodec in quality. The research compares the model using subjective human ratings (MUSHRA) as well as objective measurements (ViSQOL, Mel-SNR), confirming its superiority in producing natural, artifact-free audio. The article also explores text-to-speech (TTS) and text-to-music (TTM) tasks, using MBD as a decoder with models like Bark and MusicGen, leading to improved clarity and realism. While its computational cost is greater, MBD outperforms GAN-based methods in fidelity and is thus a superior choice for speech synthesis, music synthesis, and general audio modeling.[10]

REFERENCES:

[1] H.-J. Lee and S.-J. Buu, “Deep Generative Replay with Denoising Diffusion Probabilistic Models for Continual Learning in Audio Classification,” *IEEE Access*, 2024.

[2] P. Gonzalez, Z.-H. Tan, J. Østergaard, J. Jensen, T. S. Alstrøm, and T. May, “Investigating the design space of diffusion models for speech enhancement,” *IEEE/ACM Trans Audio Speech Lang Process*, 2024.

[3] B. Maman, J. Zeitler, M. Müller, and A. H. Bermano, “Multi-Aspect Conditioning for Diffusion-Based Music Synthesis: Enhancing Realism and Acoustic Control,” *IEEE/ACM Trans Audio Speech Lang Process*, 2024.

[4] E. Moliner, F. Elvander, and V. Välimäki, “Blind audio bandwidth extension: A diffusion-based zero-shot approach,” *IEEE/ACM Trans Audio Speech Lang Process*, 2024.

[5] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, T. Peer, and T. Gerkmann, “Causal Diffusion Models for Generalized Speech Enhancement,” *IEEE Open Journal of Signal Processing*, 2024.

[6] R. Huang *et al.*, “Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models,” in *International Conference on Machine Learning*, 2023, pp. 13916–13932.

[7] J.-E. Ayilo, M. Sadeghi, R. Serizel, and X. Alameda-Pineda, “Diffusion-based Unsupervised Audio-visual Speech Enhancement,” *arXiv preprint arXiv:2410.05301*, 2024.

[8] H. Liu *et al.*, “Audioldm: Text-to-audio generation with latent diffusion models,” *arXiv preprint arXiv:2301.12503*, 2023.

[9] Z. Guo *et al.*, “Audio Generation with Multiple Conditional Diffusion Model,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, pp. 18153–18161.

[10] R. San Roman, Y. Adi, A. Deleforge, R. Serizel, G. Synnaeve, and A. Défossez, “From discrete tokens to high-fidelity audio using multi-band diffusion,” *Adv Neural Inf Process Syst*, vol. 36, pp. 1526–1538, 2023.