Harmony AI-Music Remastering Using Deep Learning Models

Sujatha P
Senior Grade Lecturer,

Rubina
Senior Grade Lecturer

Department of Computer Science & Engg,
Government Polytechnic, Bagalkot, Karnataka

1. Abstract

In today's digital age, where technology permeates every aspect of our lives, the demand for high-quality audio experiences has skyrocketed. From music streaming services like Spotify and Apple Music to video conferencing platforms like Zoom and Microsoft Teams, clear and immersive audio is essential for both entertainment and professional purposes. However, many audio recordings, whether captured through professional equipment or consumer-grade devices, suffer from various imperfections that detract from the listening experience. These imperfections can include background noise, limited frequency range, and unwanted distortions. HarmonyAI is an Poposed work that leverages the power of AI to enhance audio recordings, providing users with a richer and more enjoyable listening experience. By combining advanced signal processing techniques with deep learning models, HarmonyAI aims to address common audio imperfections and elevate the overall audio quality. The Proposed work utilizes a dataset of high-quality and low-quality audio pairs to train AI models, enabling them to learn the complex relationships between distorted and clean audio signals.

Keywords: Deep Learning, Artificial Intelligence, wavenet, music generation

2. Introduction

The pursuit of high-quality audio has led to the development of various audio enhancement techniques. Traditional methods, such as Wiener filtering and spectral subtraction, have been widely used to address noise reduction and other audio imperfections. However, recent advancements in artificial intelligence (AI), particularly in deep learning, have opened up new possibilities for audio enhancement, surpassing traditional methods in performance and flexibility. HarmonyAI is an innovative project that leverages the power of AI to enhance audio recordings, providing users with a richer and more enjoyable listening experience. By combining advanced signal processing techniques with deep learning models, HarmonyAI aims to address common audio imperfections and elevate the overall audio quality. The proposed work utilizes a dataset of high-quality and low-quality audio pairs

to train AI models, enabling them to learn the complex relationships between distorted and clean audio signals. The core motivation behind HarmonyAI is to bridge the gap between imperfect recordings and the desired high-quality audio experience. Whether it's removing background noise from a live concert recording, enhancing the clarity of speech in a podcast, or restoring a vintage music track, HarmonyAI aims to provide users with the tools to improve their audio content.

3. Literature Survey

The field of audio enhancement has undergone a significant transformation, progressing from traditional signal processing techniques to more sophisticated deep learning-based approaches. This survey explores the key milestones in this journey, emphasizing the transformative impact of deep learning on audio enhancement.

[1] Santiago Pascual st.al, 2017 proposes a generative adversarial network (GAN) for speech enhancement, where the generator network learns to produce enhanced speech while the discriminator network distinguishes between real and enhanced speech.The SEGAN model achieves state-of-the-art performance in speech enhancement tasks, demonstrating the potential of GANs in this domain.[2] Yong Xu et.a;., 2014 Explores the use of Convolutional Neural Networks (CNNs) for speech enhancement, demonstrating their ability to capture local spectral and temporal features in speech signals. The proposed CNN-based model achieves significant improvements in speech quality and intelligibility compared to traditional methods.[3] Aaron van den Oord et.al 2016 introduces WaveNet, a deep learning model based on dilated convolutions that can generate high-quality audio waveforms.WaveNet has been successfully applied to various audio-related tasks, including speech synthesis, music generation, and audio enhancement.[4] Daniel Stoller et.al.,2018 investigates the use of deep learning models for music source separation in the waveform domain, aiming to extract individual instruments or vocals from a music recording.

The proposed model, based on a U-Net architecture, achieves promising results in separating different music sources.

4. Proposed Methodology

HarmonyAI proposes an innovative methodology that seamlessly integrates Digital Signal Processing (DSP) techniques with cutting-edge deep learning models to address the multifaceted challenges of audio enhancement. This hybrid approach is designed to leverage the strengths of both traditional signal processing methods and modern artificial intelligence, creating a powerful synergy that pushes the boundaries of audio quality improvement. Fig 1.1 shows the flow diagram

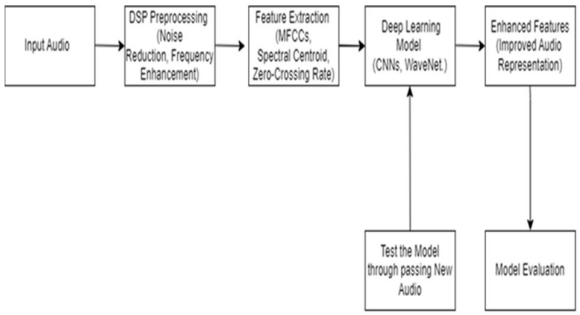

Fig1.1 Flow Diagram

HarmonyAI leverages the power of AI to enhance audio recordings, aiming to improve their clarity, richness, and overall listening experience. It achieves this through a combination of two main stages:

DSP Preprocessing:

- Noise reduction using Wiener filtering.
- Frequency enhancement using a Butterworth low-pass filter

Deep Learning Model:

- A CNN model for feature extraction and enhancement.
- A WaveNet model for capturing temporal dependencies.

Components of the Proposed Solution

1. **Digital Signal Processing (DSP) Module**

- This module applies preprocessing techniques to the input audio recordings to address basic imperfections and prepare the data for the deep learning model. It includes the following components:

Noise Reduction: Employs Wiener filtering to reduce background noise and static, improving the signal-to-noise ratio.

Frequency Enhancement: Utilizes a Butterworth low-pass filter to enhance the frequency range of the audio, ensuring a richer representation of the original sound.

2. **Deep Learning Module**

This module is the core of HarmonyAI, responsible for learning complex patterns and relationships between distorted and clean audio signals to perform advanced enhancement. It includes the following components:

1. CNN Model: The CNN model identifies intricate spectral and temporal features in audio spectrograms, focusing on patterns like pitch, harmonics, and noise artifacts. Its architecture incorporates residual blocks to enhance feature propagation, reducing the risk of gradient vanishing during training. Attention mechanisms further refine the model's focus on critical areas of the spectrogram, ensuring precise enhancement of audio details.

2. WaveNet Model: WaveNet employs dilated convolutions to capture long-term temporal relationships in audio signals. This autoregressive model processes data sequentially, ensuring temporal coherence in the enhanced audio. Its ability to model complex dependencies makes it particularly effective for tasks like speech and music enhancement, where maintaining the natural flow is crucial.

3. Combined Model: The integration of CNN and WaveNet models leverages their strengths in spectral and temporal analysis, respectively. The combined model aligns their outputs to create a richer and more balanced audio representation, addressing imperfections

across all dimensions. The proposed work is divided into following phases.

Phase 1: Data Acquisition and Preparation.

The dataset is organized into directories for each data split: train, val, and test. Each directory contains audio files representing both high-quality and low-quality recordings across various genres, instruments, and recording conditions. This diversity ensures the AI model can generalize to different real-world audio conditions. Data cleaning, DSP preprocessing such as noise reduction and frequency enhancement using Butterworth low-pass filtering techniques are applied in this phase. The following features are extracted from audio file.

- **Mel-Frequency Cepstral Coefficients (MFCCs)**: Captures the power spectrum of the audio signal, commonly used in audio and speech processing.
- **Spectral Centroid**: Measures the "brightness" of the sound by determining the center of mass of the spectrum.
- **Zero-Crossing Rate**: Indicates the number of times the audio signal changes sign, providing a measure of noisiness or percussiveness.
- **Chroma**: Analyzes the harmonic content of the audio by representing the 12 pitch classes in music.
- **Spectral Contrast**: Measures the difference in amplitude between peaks and valleys in the audio spectrum, providing information about timbral texture.

The extracted features are concatenated together and normalized to ensure consistency in the feature set. The length of the feature set is standardized to a fixed length of 100 using librosa.util.fix_length, ensuring uniformity across all samples. The data is split into train,test and validate samples.

Phase 2: Model Development and Training.

Model Selection: For the task of audio remastering, we chose deep learning architectures that are particularly effective at handling the spatial and temporal characteristics of audio signals. Specifically, we selected **Convolutional Neural Networks (CNNs)** and **WaveNet models** due to their proven performance in tasks involving sequential data such as audio.

The architecture of the **CNN model** uses residual blocks and attention mechanisms, while the **WaveNet model** employs causal convolutions with dilated convolutions to capture long-range temporal dependencies.

The model architecture, named CNN, is tailored to enhance audio recordings by leveraging:

- Residual connections to avoid vanishing gradients.
- Attention mechanisms for focusing on important features.
- Fully connected and reshaped output layers for task-specific predictions.

WaveNet Model: The WaveNet model uses causal convolutions with dilated convolutions to capture long-range dependencies. Residual connections are incorporated, ensuring gradient flow throughout the layers. The model outputs a set of features that are reshaped and used for

further processing. The combination of dilated convolutions and residual connections helps the model to effectively capture both short- and long-term dependencies in the audio.

- The combined model merges the outputs from both the CNN and WaveNet networks, effectively leveraging both spatial and temporal patterns in the audio data. This combined approach improves the quality of feature extraction and enhances the overall performance of the model.

Phase3: System Evaluation and optimization

Objective Evaluation: We'll use objective metrics like Signal-to-Noise Ratio (SNR) and Peak Signal-to-Noise Ratio (PSNR) to measure how much we've improved the audio quality. These metrics give us a concrete way to assess the reduction in noise and distortion.

Subjective Evaluation: We'll also conduct listening tests with real people to get their feedback on the enhanced audio. This will help us understand how our system affects the overall listening experience.

- **Signal-to-Noise Ratio (SNR)** is a measure of the strength of a desired signal relative to the background noise present in a system. It quantifies how much useful information (signal) is contained in the data compared to unwanted distortions (noise). A higher SNR indicates better signal quality.

$$\text{SNR (in dB)} = 10 \cdot \log_{10}\left(\frac{P_{\text{signal}}}{P_{\text{noise}}}\right)$$

- **Peak Signal-to-Noise Ratio (PSNR)** is a metric used to measure the quality of a reconstructed or compressed signal, often used in image and audio processing. It compares the original signal with the distorted or compressed version to evaluate the degradation introduced.

$$\text{PSNR (in dB)} = 10 \cdot \log_{10}\left(\frac{\text{MAX}^2}{\text{MSE}}\right)$$

Phase 4: Deployment and Integration

Deployment: Deploy it on a platform where people can access it. This could be a web server, or a cloud-based service.

- **Graphical User Interface:** The Harmony-AI Music Remastering web application enables users to upload audio files and enhance them using AI-based techniques. The app provides features to play and control the original and enhanced audio files through a waveform display, volume slider, and time stamps **User Interface**: The frontend of the web application is developed using HTML, CSS, and JavaScript, with special emphasis on user experience and ease of use. The user interface is designed to be simple and intuitive. It features an audio player for both the original and enhanced audio, allowing users to compare the two. A waveform visualization is also provided for both the original and remastered tracks, offering a visual representation of the audio signals. The app ensures that the user can easily navigate between the original and enhanced versions, with basic audio controls like play, pause, stop, and volume adjustments.
- **Waveform Visualization**: The waveform visualization is implemented using the WaveSurfer.js library, which dynamically generates visual representations of the audio

files. This allows users to visually analyze the differences between the original and enhanced tracks. The visual aspect makes the process more engaging and helps users understand how the AI enhancement affects the audio signal.

- **File Upload and Processing**: Users can upload audio files in standard formats such as MP3 or WAV. Once the file is uploaded, the backend AI models process the file using the pre-trained deep learning models (such as CNN and WaveNet). The processing involves noise reduction (via Wiener filtering), frequency enhancement (via Butterworth low-pass filtering), and feature extraction. The system then returns the enhanced audio file, ready for playback.
- **Real-time Processing Feedback**: The application also provides real-time feedback during the enhancement process. As the AI model processes the uploaded audio file, users can view progress indicators showing the status of the enhancement. This feature helps users stay informed about the ongoing operation and gives them an estimate of the remaining processing time.
- **Backend and Model Integration**: The backend of the web application is responsible for handling the file upload, model inference, and file delivery to the frontend. The backend is built using Python and Flask, with the AI models integrated into this server environment. When an audio file is uploaded, the server processes the file, passes it through the deep learning models for enhancement, and sends back the improved audio file to the frontend for user playback.
- **Responsive Design**: The web application is built to be responsive, ensuring that users can seamlessly access and use the platform. The design is friendly, adapting the layout and controls for optimal use.

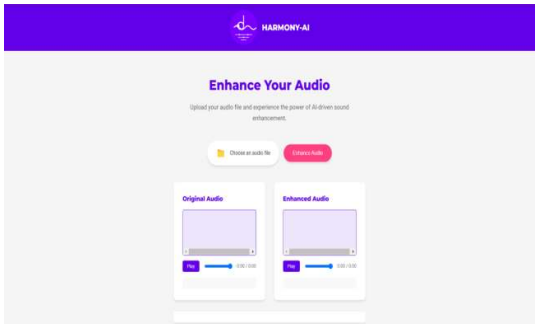

Fig 1.2 **User Interface**

5. Experimentation and Results

HarmonyAI is an AI-powered audio enhancement project that utilizes a diverse dataset of high-quality and low-quality audio recordings to train deep learning models. The project employs DSP techniques for pre-processing and feature extraction, while a combined CNN and WaveNet model is used for audio enhancement. The system's performance is evaluated using objective metrics like SNR and PSNR, as well as subjective listening tests. The project aims to achieve significant improvements in both objective and perceptual audio quality.

Sample Testing

Before implementing in user interface we fed the model with the low quality music and got results, which are as follows in the figure 1.3.

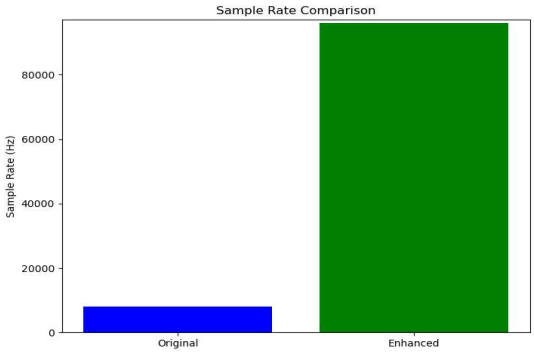

Fig 1.3 Sample rate comparison between Original and Enhanced audio

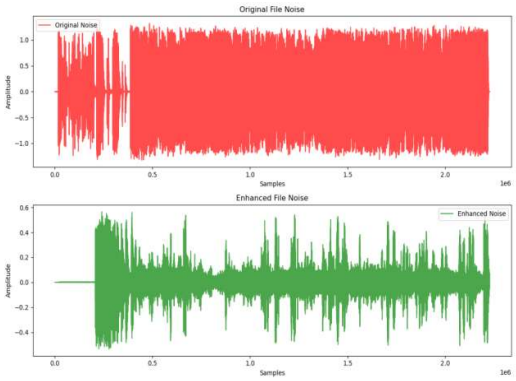

Fig 1.4 Sample Testing using User Interface

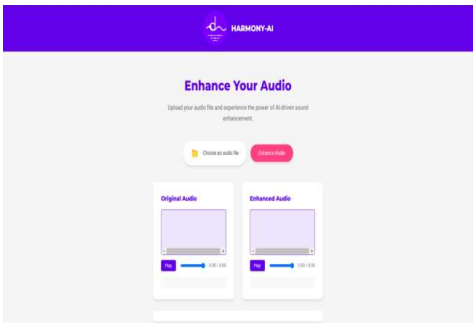

Fig 1.5 User Interface

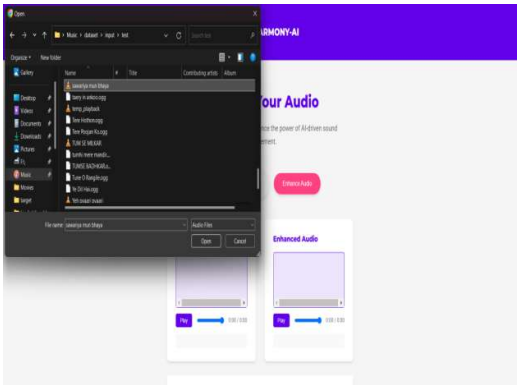

Fig 1.6 Choosing the file to be enhanced

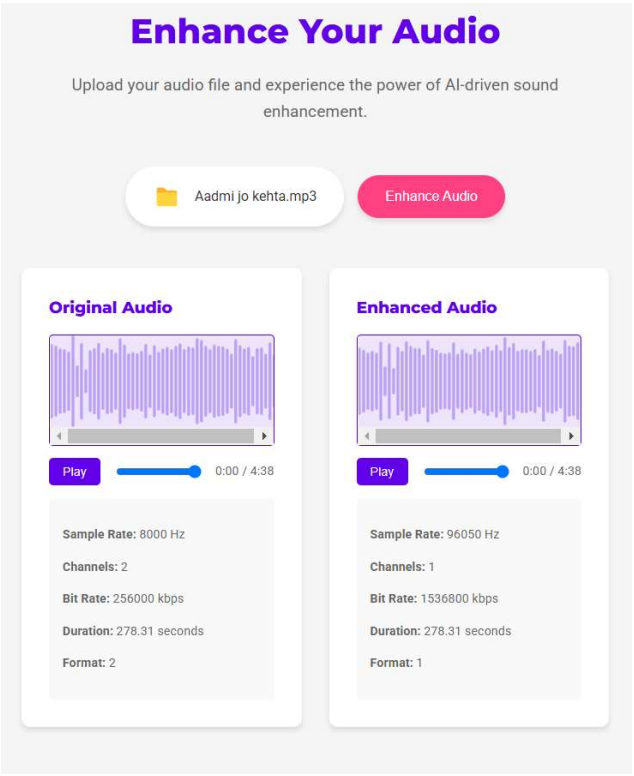

Fig 1.7 Completion of audio enhancement

6. References

[1] Santiago Pascual, Antonio Bonafonte, and Joan Serrà, (2017) " SEGAN: Speech Enhancement Generative Adversarial Network"

[2] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee (2014.) "A Convolutional Neural Network Approach for Speech Enhancement"

[3] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu (2016.) "Wavenet: A Generative Model for Raw Audio"

[4] Daniel Stoller, Sebastian Ewert, and Simon Dixon((2018) "Music Source Separation in the Waveform Domain"

[5] K. Chen, W. Zhang, S. Dubnov, G. Xia and W. Li, "The Effect of Explicit Structure Encoding of Deep Neural Networks for Symbolic Music Generation,"(2019) International Workshop on

Multilayer Music Representation and Processing (MMRP), 2019, pp. 77-84, doi: 10.1109/MMRP.2019.00022.

[6] Michael Furner, Md Zahidul Islam, Chang-Tsun Li. (2019) Knowledge Discovery and Visualisation Framework using Machine Learning for Music Information Retrieval from Broadcast Radio Data. Expert Systems with Applications 89, pages 115236

[7] Daniel Rivero, Iván Ramírez-Morales, Enrique Fernandez-Blanco, Norberto Ezquerra, Alejandro Pazos. (2019) Classical Music Prediction and Composition by Means of Variational Autoencoders. Applied Sciences 10:9, pages 3053.

[8] Wang Shuo, Mu Ming. (2019) Exploring online intelligent teaching method with machine learning and SVM algorithm. Neural Computing and Applications 20.

[9] Joo-Wha Hong, Qiyao Peng, Dmitri Williams. (2020) Are you ready for artificial Mozart and Skrillex? An experiment testing expectancy violation theory and AI music. New Media & Society 19, pages 146144482092579.