Stem Separation and Audio Enhancement using Deep learning

Prerak Joshi
Department of Artificial Intelligence and Data Science
Thakur College of Engineering and Technology
Mumbai
joshiprerak.123@gmail.com

Huzaifa Khan
Department of Artificial Intelligence and Data Science
Thakur College of Engineering and Technology
Mumbai
workwithhuzi@gmail.com

Ajit Singh
Department of Artificial Intelligence and Data Science
Thakur College of Engineering and Technology
Mumbai
ajitsinghds02@gmail.com

Ms. Niki Modi
Department of Artificial Intelligence and Data Science
Thakur College of Engineering and Technology
Mumbai
niki.modi@tcetmumbai.in

Abstract: **This paper presents advancements in deep learning for audio processing, emphasizing two key areas: stem separation and audio enhancement. Stem separation, vital for music production and audio restoration, has been significantly improved through models like HT Demucs, which leverages hybrid transformer architectures to capture both local and global audio features. HT Demucs surpasses traditional models by operating directly in the waveform domain, optimizing both accuracy and real-time performance.**

In audio enhancement, we propose a Convolutional Recurrent Neural Network (CRNN) model integrated with Time-Attention Transformers, designed to enhance speech clarity while mitigating noise and reverberation. Unlike conventional frequency-domain methods, our model processes raw waveforms, delivering more efficient and artifact-free outputs, ideal for real-time applications in mobile devices and streaming platforms.

Experimental validation on the MUSDB-18 dataset and additional benchmarks confirm the superiority of our approaches in both separation and noise suppression. Future work will incorporate real-time processing using the JUCE framework and extend the system to support six-channel stem separation with online API capabilities.

Keywords: Hybrid Demucs, Stem Separation, CRNN Architecture, Audio Enhancement, Denoising, Real-Time Stem Separation.

I. INTRODUCTION

Audio processing has seen significant advancements with the introduction of deep learning techniques, particularly in the fields of stem separation and audio enhancement. Stem separation, the task of isolating individual instrumental or vocal tracks from a mixture, is essential in various applications, including music production, remixing, and audio restoration. Recent approaches leverage deep learning models to achieve greater accuracy and efficiency. Convolutional networks applied in the time-domain have led to substantial improvements in the quality of musical source separation, enabling its deployment in real-time systems.[2] These advancements allow real-time audio interaction, which was previously challenging with traditional methods due to computational constraints.

In parallel, Audio enhancement focuses on improving the perceptual quality of degraded audio by reducing noise and other artifacts. This is particularly important in communication systems, broadcasting, and hearing aids, where audio clarity is crucial. Conventional signal processing techniques often fail in challenging environments with complex noise characteristics.[3] demonstrate that deep neural networks (DNNs) have shown remarkable success in addressing these issues, surpassing traditional algorithms. The need for real-time audio enhancement solutions is growing, driven by the widespread adoption of mobile devices and live streaming platforms, which require immediate audio processing with minimal latency.

However, real-time processing in deep learning systems poses significant challenges due to the high computational demands. To address this, propose structured pruning techniques that optimize the size and complexity of neural networks, allowing for more efficient memory usage and faster inference times.[1] These methods are crucial in ensuring that both stem separation and audio enhancement can be performed in real-time, making them applicable for live applications, such as interactive audio systems, mobile devices, and streaming platforms. The integration of deep learning with real-time processing for stem separation and audio enhancement opens new possibilities for improved audio quality and user experience. This paper explores the state-of-the-art methods in these areas and highlights the challenges and potential solutions for achieving real-time performance.

II. RELATED WORK

Music source separation has been extensively explored using deep learning methods,One notable approach with several architectures emerging as state-of-the-art in handling multi-source audio. is Spleeter, a tool designed for ease of use, high performance, and speed. Based on TensorFlow, Spleeter provides pre-trained models capable of separating audio into 2, 4, or 5 stems with a single command line. Its models deliver separation results that closely match the state-of-the-art, making it one of the best-performing publicly available tools for 4-stem separation on the widely used MusDB18 benchmark. Spleeter also distinguishes itself through its remarkable processing speed, achieving stem separation up to 100 times faster than real-time on a single GPU. [2]

Another prominent advancement is Demucs, which focuses on separating four known sources: drums, bass, vocals, and other accompaniments. Traditional methods rely on predicting soft masks over mixture spectrograms, but waveform-based approaches have lagged behind in terms of performance. The introduction of Demucs, a convolutional and recurrent model, demonstrates significant improvements in this area by outperforming the Wave-U-Net model by 1.6 points of signal-to-distortion ratio (SDR). In addition, Demucs incorporates a novel weakly supervised learning scheme, remixing stems from unlabeled tracks with isolated sources from a supervised dataset, thereby enhancing its capability to process waveform inputs comparably to spectrogram-based methods.[4]

In mobile speech enhancement, background noise suppression is critical for clear communication, particularly in noisy environments. Traditional systems often integrate speech enhancement algorithms with multiple microphones in mobile devices. A novel approach for real-time speech enhancement on dual-microphone phones employs a densely-connected convolutional recurrent network for dual-channel complex spectral mapping.[5] This model, optimized through structured pruning, provides a memory-efficient, low-latency solution suitable for real-time processing. Experimental results show that this dual-microphone system consistently outperforms earlier deep learning-based methods, including beamformers, demonstrating its efficacy in mobile communication scenarios. Speech dereverberation has also benefited from advancements in deep learning. Izotope's work on non-stationary noise removal through blind source separation has been adapted to the problem of dereverberation by training a bidirectional recurrent neural network (BRNN).[6] This approach contrasts with traditional spectral subtraction-based methods, showing superior performance in both qualitative and quantitative evaluations. The deep learning-based approach significantly improves speech clarity by effectively mitigating reverberation, further advancing the field of speech enhancement.

III. DATA SET

TWe curated an internal dataset composed of several thousand songs from around 200 different artists, representing a diverse array of musical genres. Each song’s individual stems were classified into one of four groups based on labels provided by the music producers (such as "vocals," "fx," or "sub"). However, these labels were often subjective and, at times, ambiguous. To enhance the dataset's accuracy, we manually reviewed and corrected the automated labeling for a subset of these tracks, discarding any stems with unclear labels. Our initial Hybrid Demucs model was trained using the MUSDB dataset in combination with this manually verified subset of songs.[1]

To further refine the dataset, we applied specific preprocessing steps. We kept only the stems where all four sources were active for at least 30% of the track's duration. A segment was classified as silent if its volume was below a defined threshold. For each song, categorized into sources such as drums, bass, vocals, and other instruments, we analyzed the output of the Hybrid Demucs model for each separated stem.[2] Ideally, if the stems were perfectly labeled and the model performed flawlessly, the separated outputs would align exactly with the original stems.

This section draws on the methodology used in the training process of the Hybrid Demucs model and highlights how the dataset was curated and prepared for training, including the measurement of stem volume over time to ensure consistency.

IV. PROPOSED ARCHITECTURE

A. Stem Separation Model

For our proposed system we opted to use Hybrid Transformer for Demucs (HT Demucs) (Add reference) over other existing stem separation architectures.

HT Demucs uses an encoder-decoder architecture for its processing.

The model uses an encoder-decoder architecture where the input audio is first broken down into smaller time-frequency representations by the encoder, allowing the model to capture intricate details in the audio signal. The transformer layers are responsible for modeling long-range dependencies within the audio sequence, offering improved context-awareness when separating different stems like vocals, bass, drums, and other instruments. This hybrid approach enables HT

Demucs to handle both local and global structures in music effectively.[4]

The use of fade-ins and fade-outs within the model helps to smooth transitions at the boundaries of the separated audio signals. These fades reduce potential artifacts or abrupt changes in the reconstructed signals, particularly when reassembling stems after separation. Internally, the decoder reconstructs the separated sources from the encoded representations, with attention mechanisms in the transformers ensuring that each stem retains its distinct temporal and harmonic properties.[7] The hybrid nature of HT Demucs, with convolutional layers handling short-term features and transformers capturing longer-term dependencies, allows for more precise and musically coherent separations compared to traditional methods.

The implementation of HT Demucs was done through PyTorch for this research by using libraries like torchaudio to facilitate the audio processing tasks.

HT Demucs gives a performance edge over Spleeter and Demucs because of the use of its hybrid transformer. It is an improvement made on existing Demucs model because the original model was only capable of working on direct waveforms which was useful for more rhythmic and temporal based separations however it wasn't too accurate for Harmonic content. Hybrid transformer essentially allows the Demucs model to do both, Waveform based and Frequency based separation by using STFT.[8] Consequently, this enables HT Demucs to be powerful for separating harmonic content from the original mix file.[9]

B. Audio Enhancement Architecture

The proposed noise suppression system is built upon a sophisticated model architecture designed for real-time audio enhancement, integrating multiple neural network layers and components to address different aspects of audio processing.[10] At its core, the model utilizes a Convolutional Recurrent Neural Network (CRNN) that merges the strengths of a DenseNet-based convolutional structure and a Recurrent Neural Network (RNN) for temporal sequence modeling. The CRNN efficiently processes segmented audio frames, typically ranging from 20 ms to 40 ms in length, capturing both spatial and temporal features of the input signal.[11] This allows the model to separate speech from noise by performing spectral mapping across the frames. The densely connected convolutional layers help extract fine-grained spatial features, while the integrated RNN enhances the model's ability to capture temporal dependencies, making it well-suited for dynamic environments. The architecture is trained on a paired dataset of clean and noisy audio, using techniques such as dilated convolutions to improve temporal context and the model's ability to preserve speech characteristics during noise suppression. To further refine speech quality, the model incorporates a Time-Attention Transformer (TAT)[7], or a similar de-reverberation mechanism, into the architecture. This component specializes in reducing the effects of reverberation by isolating the direct speech signal from reflections and echoes, which can degrade speech clarity. The transformer-based attention mechanism allows the system to focus on critical temporal segments, enhancing speech intelligibility and removing reverberant artifacts.[7][10] Additionally, the model integrates a time-domain enhancement stage, which addresses missing frequency content by utilizing a 1D Convolutional Encoder-Decoder architecture. This stage reconstructs lost frequencies in the time domain, restoring a natural and complete speech waveform.[12] The decoder component refines the spectral gaps by mapping the enhanced feature representations back into the waveform, ensuring a smooth, high-quality output. The overall architecture is optimized for low-latency operation, making it suitable for deployment on mobile and embedded systems where real-time processing is crucial, offering a robust and efficient solution for real-world noise suppression tasks.

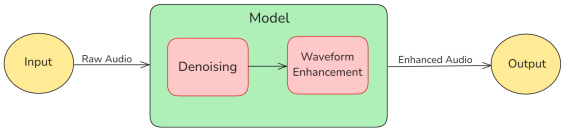
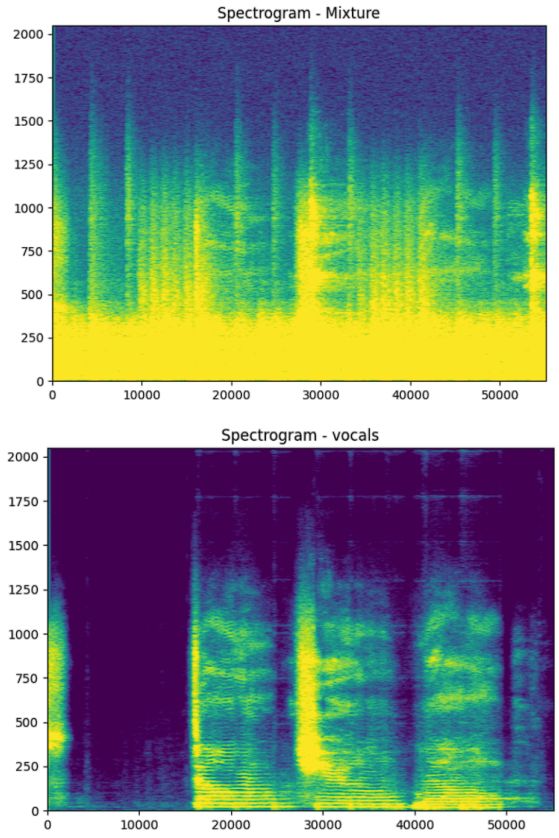


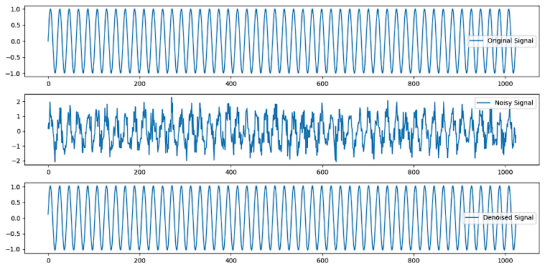
Fig. 1. Audio Enhancement Diagram

V. EXPERIMENTAL RESULTS

Using the MUSDB-18 dataset, we developed a stem separation model with four output channels. The decision to utilize four channels stemmed from the scarcity of publicly accessible datasets for six-stem separation and the substantial increase in computational complexity that a six-channel configuration would impose. A six-stem model would lead to higher memory consumption and potential performance degradation due to the enlarged dataset size.



In our audio enhancement experiments, we explored various methodologies and parameter settings. Through extensive testing, we determined that frequency-domain approaches, particularly those leveraging the Short-Time Fourier Transform (STFT), were more computationally intensive and slower compared to time-domain approaches based on the raw waveform. Furthermore, while we tested different batch sizes for the STFT-based approach, our findings indicated that transformer models, which process the entire waveform, provided superior performance in terms of both efficiency and speed. These insights guided our decision to prioritize waveform-based techniques for the enhancement task.



VI. COMPARATIVE ANALYSIS

A. Stem Separation

Research comparing HT Demucs and Spleeter highlights the advantages of HT Demucs in both performance and speed, especially in music source separation tasks. HT Demucs, a hybrid temporal/spectral model with Transformer encoders, shows state-of-the-art results on the MUSDB18 dataset, achieving 9.20 dB of SDR for vocals, outperforming traditional models like Spleeter, which primarily uses spectrogram-based U-Net architectures. HT Demucs benefits from its cross-domain Transformer structure, which captures long-range contextual information, essential for high-quality separation. It processes in the waveform domain, avoiding the limitations associated with spectrogram methods (such as phase reconstruction issues), providing cleaner outputs. Additionally, HT Demucs is optimized for faster real-time processing due to its efficient use of sparse attention kernels[7], making it more suitable for mobile or edge-device applications compared to Spleeter[2], which is slower and less precise in complex scenarios.

Model	Extra	Drums	Bass	Other	Vocals	Average
Spleeter (25k)	5.91	6.71	5.51	4.55	6.86	5.91
D3Net (1.5k)	6.68	7.36	6.2	5.37	7.8	6.68
Demucs v2 (150)	6.79	7.58	7.6	4.69	7.29	6.79
Hybrid Demucs (800)	8.34	9.31	9.13	6.18	8.75	8.34

Fig. 2. SDR scores are in decibels (dB).

Spleeter, while popular for its accessibility and decent performance, lags behind HT Demucs in terms of overall accuracy and SDR. Spleeter's strength lies in its simplicity and ease of use but is now considered outperformed by more advanced models like Demucs.[1][4]

B. Audio Enhancement

The proposed system offers several advantages over traditional noise suppression methods by avoiding the frequency domain conversion of Short-Time Fourier Transform (STFT), resulting in a more resource-efficient architecture. This simplification makes the system particularly suitable for mobile

applications, as it relies on single-channel input rather than the dual-channel input typically required for spatial noise separation in noise suppression systems.[13] The integration of a Time-Attention Transformer (TAT) enhances long-term temporal modeling, enabling the model to selectively focus on critical speech segments and improving both dereverberation and noise suppression. This, in combination with a Bidirectional Recurrent Neural Network (BRNN)[14], further refines the system's ability to handle long-term dependencies in the waveform domain, effectively enhancing speech clarity and mitigating reverberation.[15]

In contrast, conventional noise suppression systems, which employ Convolutional Recurrent Neural Networks (CRNN) combined with RNN layers, capture long-term temporal dependencies primarily in the frequency domain, focusing on non-stationary noise suppression.[16] While effective for noise reduction, these methods often require complex post-processing, such as inverse STFT (iSTFT), to convert the processed spectrogram back into a waveform. This step introduces the risk of artifacts if phase estimation is inaccurate. By operating directly in the waveform domain, the proposed system avoids such artifacts and outputs an enhanced waveform without the need for additional post-processing. Furthermore, its focus on reverberation and missing frequency restoration makes it particularly effective in scenarios where speech is degraded by reverberation or lacks certain frequency components, while traditional noise suppression systems are more specialized for dealing with stationary and non-stationary noise.[15][17]

CONCLUSION

This research presents a comprehensive study on advancing audio processing through deep learning techniques, focusing on stem separation and audio enhancement. The integration of deep learning models, such as HT Demucs for stem separation and sophisticated CRNN architectures for audio enhancement, demonstrates significant improvements in both performance and real-time processing capabilities.

The HT Demucs model, leveraging a hybrid transformer architecture, achieves superior stem separation results compared to traditional methods like Spleeter. Its ability to capture both local and global audio features enhances the accuracy and quality of the separation, making it a robust solution for complex audio tasks. The model's real-time processing capability further amplifies its applicability in mobile and edge-device environments.

In audio enhancement, the proposed CRNN-based noise suppression system, complemented by a Time-Attention Transformer (TAT) and a bidirectional recurrent network, offers a more efficient and effective approach compared to conventional methods. By operating directly in the waveform domain and incorporating advanced temporal modeling, the system mitigates common issues such as artifacts from phase estimation and provides clearer, more natural audio enhancements.

FUTURE SCOPE

In the future, we plan to significantly enhance the system by integrating real-time inference capabilities through the JUCE framework in C++, overcoming Python's inherent limitations in execution speed and ensuring efficient, low-latency processing for real-time applications. This transition will enable the system to handle real-time audio processing more effectively on various platforms, especially mobile devices. Furthermore, we aim to develop a unified package combining both C++ and Python, offering pre-built functions that can be easily used for audio processing tasks. This modular design not only enhances scalability but also promotes open-source contributions, fostering community-driven improvements. Additionally, we plan to increase the number of stem separation channels from the current four to six by incorporating custom high-fidelity audio, thereby expanding the system's capability to handle more complex and nuanced audio environments. Moreover, rather than limiting the system to offline processing, we will explore providing an API for real-time online applications, such as conference calls, allowing the system to be seamlessly integrated into communication platforms, improving audio clarity during live interactions.

REFERENCES

[1] Défossez, Alexandre, et al. "Demucs: Deep extractor for music sources with extra unlabeled data remixed." arXiv preprint arXiv:1909.01174 (2019).

[2] Hennequin, Romain, et al. "Spleeter: a fast and efficient music source separation tool with pre-trained models." Journal of Open Source Software 5.50 (2020): 2154.

[3] Xu, Yong, et al. "A regression approach to speech enhancement based on deep neural networks." IEEE/ACM transactions on audio, speech, and language processing 23.1 (2014): 7-19.

[4] Rouard, Simon, Francisco Massa, and Alexandre Défossez. "Hybrid transformers for music source separation." ICASSP 2023-2023 IEEE International

Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023.

[5] dos Santos, Arthur, Pedro de Oliveira, and Bruno Masiero. "A retrospective on multichannel speech and audio enhancement using machine and deep learning techniques." Proceedings of the 24th International Congress on Acoustics. 2022.

[6] Ashraf, Mohsin & Abid, Fazeel & Ud Din, Ikram & Rasheed, Jawad & Yesiltepe, Mirsat & Yeo, Sook Fern & Ersoy, Merve. (2023). A Hybrid CNN and RNN Variant Model for Music Classification. Applied Sciences. 13. 1476. 10.3390/app13031476.

[7] Vaswani, A. "Attention is all you need." Advances in Neural Information Processing Systems (2017).

[8] Défossez, Alexandre, et al. "Music source separation in the waveform domain." arXiv preprint arXiv:1911.13254 (2019).

[9] Hsu, Jia-Lien & Chang, Shuh-Jiun. (2021). Generating Music Transition by Using a Transformer-Based Model. Electronics. 10. 2276. 10.3390/electronics10182276.

[10] Wen, Cuihong & Zhu, Longjiao. (2022). A Sequence-to-Sequence Framework Based on Transformer With Masked Language Model for Optical Music Recognition. IEEE Access. PP. 1-1. 10.1109/ACCESS.2022.3220878.

[11] Perotin, Lauréline, et al. "CRNN-based multiple DoA estimation using acoustic intensity features for Ambisonics recordings." IEEE Journal of Selected Topics in Signal Processing 13.1 (2019): 22-33.

[12] Fu, Xinyu, et al. "CRNN: a joint neural network for redundancy detection." 2017 IEEE international conference on smart computing (SMARTCOMP). IEEE, 2017.

[13] Kim, Younsik, et al. "Deep learning-based statistical noise reduction for multidimensional spectral data." Review of Scientific Instruments 92.7 (2021).

[14] Zhu, Qiannan, et al. "Knowledge base reasoning with convolutional-based recurrent neural networks." IEEE Transactions on Knowledge and Data Engineering 33.5 (2019): 2015-2028.

[15] Roda, Antonio, et al. "Audio documents restoration as a documentary source in the linguistic research comparison of instruments." Speech audio archives: preservation, restoration, annotation, aimed at supporting the linguistic analysis. Roma: Accademia Nazionale dei Lincei (2018): 117-138.

[16] Dignam, Christopher. (2024). Harmonies on the String: Exploring the Synergy of Music and STEM. International Journal of Technology in Education and Science. 8. 491-521. 10.46328/ijtes.571.

[17] He, Sujie & Li, Yuxian. (2024). Research on Music Classification Technology Based on Integrated Deep Learning Methods. ICST Transactions on Scalable Information Systems. 11. 10.4108/eetsis.4954.

[18] Mahrishi, Mehul, et al., eds. Machine learning and deep learning in real-time applications. IGI global, 2020.

[19] Stefani, Domenico, Simone Peroni, and Luca Turchet. "A comparison of deep learning inference engines for embedded real-time audio classification." Proceedings of the International Conference on Digital Audio Effects, DAFx. Vol. 3. MDPI (Multidisciplinary Digital Publishing Institute), 2022.

[20] Chowdhury, Jatin. "Rtneural: Fast neural inferencing for real-time systems." arXiv preprint arXiv:2106.03037 (2021).