IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 6, March 2024

# An In-depth Review on Music Source Separation

**Prof. K. G. Jagtap[1], Vivek Deshmukh[2], Maviya Mahagami[3], Himanshu Lohokane[4], Rashi Kacchwah[5]**
Professor, Department of AI & ML[1]
Students, Department of AI & ML[2,3,4,5]
AISSM Polytechnic, Pune, India

**Abstract:** *The study in Music Source Separation (MSS) raises a fundamental question: Is there any benefit in considering broader contextual information, or are local acoustic features adequate? In various domains, attention-based Transformers [1] have demonstrated their capacity to assimilate information across extensive sequences. In our research, we introduce Hybrid Transformer Demucs (HT Demucs), a hybrid temporal/spectral bi-U-Net based on Hybrid Demucs [2]. Here, the innermost layers are substituted with a cross-domain Transformer Encoder, utilizing self-attention within one domain and cross-attention across domains. Although its performance is lacking when exclusively trained on MUSDB [3], we illustrate that it surpasses Hybrid Demucs (trained on the same data) by 0.45 dB of Signal-to-Distortion Ratio (SDR) when provided with an additional 800 training songs. By employing sparse attention kernels to broaden its receptive field and undertaking per-source fine-tuning, we attain state-of-the-art results on MUSDB with extra training data, achieving a remarkable 9.20 dB of SDR.*

**Keywords:** Music Source Separation, Transformers.

## I. INTRODUCTION

Demucs, a deep extractor for music sources, is a prominent framework in audio signal processing that has been used to isolate individual sound sources from complex mixtures. However, as the field advances and challenges persist, researchers are increasingly exploring hybrid approaches, integrating Demucs with complementary methodologies to unlock new potentials and address existing limitations. This literature survey explores the evolution, principles, and implications of hybrid Demucs architectures for audio source separation, examining their combinations with various deep learning models, signal processing techniques, and data-driven methodologies. The survey also delves into the intricate interplay between hybrid architectures, evaluation methodologies, and benchmark datasets, providing insight into the complex landscape of performance assessment in audio source separation.

In the realm of Music Source Separation (MSS) methods, there's a traditional categorization between spectrogram-based and waveform-based models. Spectrogram-based models include approaches like Open-Unmix [11], employing a biLSTM with fully connected components to predict a mask on the input spectrogram. Another example is D3Net [12], which utilizes dilated convolutional blocks with dense connections. More recently, there has been a preference for using complex spectrogram as both input and output [13], offering a more comprehensive representation and eliminating the top-line constraint imposed by the Ideal-Ratio-Mask. The Band-Split RNN [14], the latest spectrogram model, incorporates this concept along with multiple dual-path RNNs [15], each operating in carefully designed frequency bands. Currently, it holds the state-of-the-art performance on MUSDB with 8.9 dB. On the other hand, waveform-based models originated with Wave-U-Net [16], forming the foundation for Demucs [10], a time-domain U-Net with a bi-LSTM positioned between the encoder and decoder. Around the same period, Conv-TasNet demonstrated competitive results [9, 10] by using residual dilated convolution blocks to predict a mask over a learned representation. A recent trend involves combining both temporal and spectral domains, either through model blending, exemplified by KUIELAB-MDX-Net [17], or by adopting a biU-Net structure with a shared backbone, as seen in Hybrid Demucs [2]. Despite Hybrid Demucs being the top-ranked architecture in the latest MDX MSS Competition [18], it has now been surpassed by Band-Split RNN.

(SiSEC) in 2015 [4], the Music Source Separation (MSS) community has primarily focused on training supervised models for the task of separating songs into four stems: drums, bass, vocals, and other (representing all other instruments). The benchmark dataset used in MSS is MUSDB18 [3, 5], consisting of 150 songs in two versions (HQ

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 6, March 2024

and non-HQ). The training set comprises 87 songs, a relatively small corpus compared to other deep learning tasks like vision [6, 7] or natural language processing [8], where Transformer [1]-based architectures have found success. In source separation, both short and long context inputs are relevant. Conv-Tasnet [9] utilizes about one second of context, relying solely on local acoustic features for separation. On the other hand, Demucs [10] can use up to 10 seconds of context to address input ambiguities. In this study, our goal is to explore how Transformer architectures can effectively utilize this context and determine the amount of data needed to train them. In Section 3, we introduce a novel architecture called Hybrid Transformer Demucs (HT Demucs), which replaces the innermost layers of the original Hybrid Demucs architecture [2] with Transformer layers. These layers are applied in both the time and spectral representation, using self-attention within one domain and cross-attention across domains. Transformers typically require substantial data, so we augment the MUSDB dataset with an internal dataset of 800 songs, detailed in Section 4. Our second contribution, presented in Section 5, involves an extensive evaluation of this new architecture under various settings (depth, number of channels, context length, augmentations, etc.). We demonstrate a notable improvement over the baseline Hybrid Demucs architecture (retrained on the same data) by 0.35 dB. Finally, we experiment with increasing the context duration using sparse kernels based on Locally Sensitive Hashing to overcome memory issues during training and fine-tuning procedures. This results in a final Signal-toDistortion Ratio (SDR) of 9.20 dB on the MUSDB test set.

The achieved methodology in this survey paper will be elaborated further in the upcoming research articles on this topic.

This literature survey paper segregates the section 2 for the evaluation of the past work in the configuration of a literature survey, and finally, section 3 provides the conclusion and the future work.

## II. RELATED WORKS

**Demucs (Deep Extractor for Music Sources) by Facebook AI Research (FAIR):**

Demucs is a deep learning-based source separation model designed to separate individual sound sources from mixed music recordings.

Developed by the research team at Facebook AI Research, Demucs introduced an end-to-end architecture using convolutional neural networks (CNNs) for music source separation.

**Conv-TasNet by Luo et al.:**

Conv-TasNet is a convolutional time-domain audio separation network proposed by Luo et al. in their paper "Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation".

The model employs a time-domain convolutional encoder-decoder architecture and achieves state-of-the-art performance in single-channel audio source separation tasks.

**Wave-U-Net by Stoller et al.:**

Wave-U-Net is a waveform-based fully convolutional neural network architecture proposed by Stoller et al. in their paper "Wave-U-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation".

The model performs end-to-end music source separation directly on the raw waveform and achieves competitive results in separating individual instruments and vocals.

**Open-Unmix by Stöter et al.:**

Open-Unmix is an open-source music source separation framework developed by Fabian-Robert Stöter et al.

The framework utilizes deep neural networks for source separation and provides pre-trained models for separating vocals, drums, bass, and other instruments from mixed music recordings.

**Deep Clustering by Hershey et al.:**

Deep Clustering is a deep learning-based method proposed by Hershey et al. in their paper "Deep Clustering and Conventional Networks for Music Separation: Strong Together".

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 6, March 2024

The method learns embeddings for each time-frequency bin of the mixture spectrogram and clusters them to estimate the source signals, serving as a foundational approach in deep learning-based source separation.

**Deep Embedding Separation (DES) by Takahashi et al.:**
Deep Embedding Separation (DES) is a method proposed by Takahashi et al. in their paper "Deep Embedding Separation: Source Separation Meets Dense Embedding Learning".
DES combines deep embedding learning with clustering techniques to separate sound sources from mixed audio recordings.

**Single-Channel Source Separation (SCSS) by Uhlich et al.:**
Single-Channel Source Separation (SCSS) is a method proposed by Uhlich et al. in their paper "Improving Music Source Separation Based on Deep Neural Networks Through Data Augmentation and Network Fusion".
SCSS focuses on improving music source separation performance using data augmentation techniques and network fusion strategies.

**Music Source Separation Using CNN-RNN Architecture by Luo et al.:**
Luo et al. proposed a music source separation method based on a hybrid convolutional-recurrent neural network (CNN-RNN) architecture in their paper "Deep Convolutional Neural Networks and Data Augmentation for Acoustic Event Detection".
The CNN-RNN architecture leverages both local and global temporal dependencies in the audio signal for improved separation performance.

**Deep Complex U-Net by Takahashi et al.:**
Deep Complex U-Net is a neural network architecture proposed by Takahashi et al. in their paper "Multi-Channel Music Separation with Deep Complex U-Net".
The Deep Complex U-Net model extends the U-Net architecture to handle multi-channel audio signals and achieves state-of-the-art results in music source separation tasks.

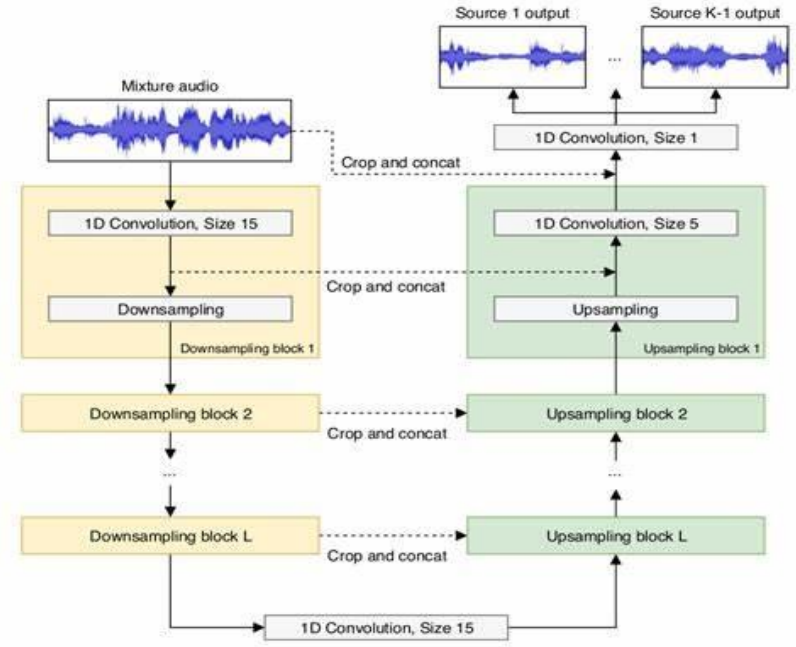


Figure 1: Proposed Model Overview

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 6, March 2024

**Joint Optimization for Music Separation by Luo et al.:**

Luo et al. introduced a joint optimization framework for music source separation in their paper "Joint Optimization for Music Source Separation".

The framework optimizes the separation performance by jointly training the model parameters and estimating the time-frequency masks for each source.

## III. CONCLUSION AND FUTURE SCOPE

The introduction of Hybrid Transformer Demucs marks a significant advancement in audio source separation techniques, extending the Hybrid Demucs architecture with Transformers at its core. This variant replaces inner convolutional layers with a Cross-domain Transformer Encoder, integrating self-attention and cross-attention mechanisms for capturing complex dependencies in audio signals. By combining the strengths of convolutional and transformer architectures, the model achieves superior performance over the baseline Hybrid Demucs, surpassing it by 0.45 dB. Sparse attention techniques enable efficient scaling for processing longer input lengths during training, reaching up to 12.2 seconds. This scalability not only enhances the model's capacity for longer audio sequences but also improves performance by an additional 0.4 dB, making it applicable to a broader range of real-world scenarios.

Looking ahead, our exploration into splitting the spectrogram into subbands, as proposed in [14], presents an exciting avenue for further enhancement. By processing different frequency subbands separately, we aim to tailor the model's processing to better suit the characteristics of each frequency range. This approach has the potential to further boost separation performance and enhance the model's adaptability to diverse audio sources and environments.

Hybrid Transformer Demucs represents a significant step forward in audio source separation, leveraging the synergy between convolutional and transformer architectures to achieve state-of-the-art performance. With ongoing research and development efforts, we are committed to advancing the boundaries of audio processing and empowering applications across domains such as music production, speech enhancement, and more.

## REFERENCES

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). "Attention is all you need." CoRR, vol. abs/1706.03762.

[2] De´fossez, A. (2021). "Hybrid spectrogram and waveform source separation." In Proceedings of the ISMIR 2021 Workshop on Music Source Separation.

[3] Rafii, Z., Liutkus, A., Sto¨ter, F. R., Mimilakis, S. I., & Bittner, R. (2017). "The musdb18 corpus for music separation."

[4] Ono, N., Rafii, Z., Kitamura, D., Ito, N., & Liutkus, A. (2015). "The 2015 Signal Separation Evaluation Campaign." In International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA).

[5] Zafar Rafii, Antoine Liutkus, Fabian-Robert Sto¨ter, Stylianos Ioannis Mimilakis, and Rachel Bittner, "Musdb18-hq - an uncompressed version of musdb18," Aug. 2019.

[6] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Herve´ Je´gou, "Going deeper with image transformers," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021.

[7] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjo¨rn Ommer, "High-resolution image synthesis with latent diffusion models," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2022, pp. 10684–10695.

[8] Tom B. Brown et al., "Language models are few-shot learners," 2020.

[9] Yi Luo and Nima Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019.

[10] Alexandre De´fossez, Nicolas Usunier, Le´on Bottou, and Francis Bach, "Music source separation in the waveform domain," 2019.

[11] F.-R. Sto¨ter, S. Uhlich, A. Liutkus, and Y. Mitsufuji, "Open-unmix - a reference implementation for music source separation," Journal of Open Source Software, 2019.

[12] Takahashi, N., & Mitsufuji, Y. (2020). "D3net: Densely connected multidilated densenet for music source separation."

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 6, March 2024

[13] Choi, W., Kim, M., Chung, J., & Jung, S. (2021). "Lasaft: Latent source attentive frequency transformation for conditioned source separation." In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).
[14] Luo, Y., & Yu, J. (2022). "Music source separation with band-split RNN."
[15] Yi Luo, Zhuo Chen, and Takuya Yoshioka, "Dualpath rnn: efficient long sequence modeling for timedomain single-channel speech separation," in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 46–50.
[16] Daniel Stoller, Sebastian Ewert, and Simon Dixon, "Wave-u-net: A multi-scale neural network for end- toend audio source separation," arXiv preprint arXiv:1806.03185, 2018.
[17] Minseok Kim, Woosung Choi, Jaehwa Chung, Daewon Lee, and Soonyoung Jung, "Kuielab-mdx-net: A twostream neural network for music demixing," 2021.
[18] Yuki Mitsufuji, Giorgio Fabbro, Stefan Uhlich, FabianRobert Sto¨ter, Alexandre De´fossez, Minseok Kim, Woosung Choi, Chin-Yun Yu, and Kin-Wai Cheuk, "Music demixing challenge 2021," Frontiers in Signal Processing, vol. 1, jan 2022.
[19] Romain Hennequin, Anis Khlif, Felix Voituret, and Manuel Moussallam, "Spleeter: a fast and efficient music source separation tool with pre-trained models," Journal of Open Source Software, 2020.
[20] Zelun Wang and Jyh-Charn Liu, "Translating math formula images to latex sequences using deep neural networks with sequence-level training," 2019.
[21] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," 2014.
[22] Fabian-Robert Sto¨ter, Antoine Liutkus, and Nobutaka Ito, "The 2018 signal separation evaluation campaign," 2018