Sveučilište u Rijeci
University of Rijeka
http://www.uniri.hr

Polytechnic: Journal of Technology Education, Volume 1, Number 1 (2017)
Politehnika: Časopis za tehnički odgoj i obrazovanje, Volumen 1, Broj 1 (2017)

Politehnika
Polytechnica
http://www.politehnika.hr/journal
cte@uniri.hr

Transformers in music source separation

Grgur Živković
Prirodoslovno-matematički fakultet
Sveučilište u Splitu
Ruđera Boškovića 33
gzivkovic@pmfst.hr

Abstract

The field of Music Source Separation (MSS) has witnessed remarkable progress in its pursuit of isolating individual sound sources from complex audio mixtures. A central focus of the MSS community has been the separation of music into four key stems: drums, bass, vocals, and other instruments. MSS, driven by its potential applications in diverse domains, has attracted considerable attention, especially in speech enhancement and speech separation. Deep learning has significantly propelled the development of MSS, enabling the creation of more accurate and efficient source separation techniques. A substantial body of research has been dedicated to refining these techniques, relentlessly pursuing state-of-the-art results. Every couple of years, researchers converge to present their latest contributions and advances at different conferences. This paper provides a comprehensive overview of the MSS landscape, emphasizing the pivotal role of the Transformer model in contemporary source separation research.

Keywords: *Music source separation, Deep learning, Transformers*

1 Introduction

In the field of Music Source Separation (MSS), significant advancements have been made to address the challenge of separating individual sound sources from a mixture. The MSS community has mainly focused on separating songs into 4 stems: drums, bass, vocals and other (all other instruments). Benchmarking is almost universally done on MUSDB18 dataset [1, 2] (both hq and non-hq versions) which consists of 150 full lengths music tracks (~10h duration) of different genres along with their isolated stems. The field has garnered considerable attention due to its potential applications in various domains. For example, it has been used to enhance automatic music transcription, lyric and music alignment, musical instrument detection, lyric recognition, automatic singer identification and vocal activity detection [3]-[8]. Furthermore, the advances in MSS can be a good basis for improvements in related fields such as speech enhancement and speech separation [9]-[13]. Over the years, the development of deep learning methods has revolutionized this field, allowing for more accurate and efficient separation techniques. The most represented in the research include: CNNs, RNNs, LSTMs, masking-based methods and deep clustering methods. A great deal of research and focus has been devoted to advancing these methods, with the aim of achieving state-of-the-art results. Conferences such as SiSec [14] and MDX [15] have become prominent platforms for researchers to showcase their advancements in the field. Among the recent breakthroughs, the transformer model has emerged as a significant player in MSS, currently holding state-of-the-art results on the MUSDB-18 dataset. Its importance has grown steadily over the years, and it has found its application in MSS due to its ability to capture complex and long-range dependencies within audio signals.

2 Related Work

While early techniques in the field of MSS had many limitations and lacked quality and robustness, they paved the way for more advanced deep learning methods that exist today. Independent Component Analysis (ICA) is a statistical technique used for separating mixed signals into statistically independent components which are assumed to be a linear combination of the source signals [16]. Non-Negative Matrix Factorization (NMF) assumes that the spectrogram of a music mixture can be factorized into a set of basis spectrograms and their corresponding activation coefficients [17]. HMM-based predictions [18] and segmentation techniques [19] were used for learning a soft/binary mask over power spectrograms.

In 2014 came initial works using deep learning methods on speech source separation [20] which were followed by fully connected networks over few spectrogram frames [21], LSTMs [22] and multi scale convolutional/recurrent networks [23,24]. Initially, once deep learning has gained momentum, spectrogram-based models were more popular and yielded better results than waveform-based models. The former includes models like: U-Net [25] which uses magnitude spectrogram as the input and predicts a binary mask for each source, MMDenseLSTM [26] which combines multi-scale multi-level dense connections and long short-term memory (LSTM), D3Net [27] which uses dilated convolutional blocks with dense connections and Open-Unmix [28], a biLSTM with fully connected layers that predicts a mask on the input spectrogram.

The latest spectrogram model, Band-Split RNN [29], explicitly splits the spectrogram of the mixture into sub bands and perform interleaved band-level and sequence-level modelling. It currently achieves the state-of-the-art on MUSDB with 8.23 dB (if you only include models without extra training data). Waveform based models started with Wave-U-Net [30] an adaptation of the U-Net architecture that repeatedly resamples feature maps to compute and combine features at different time scales. It served as a basis for Demucs [31] which added a bi-LSTM between the encoder and decoder.

Alongside the exclusive time-domain and frequency-domain approaches, there exists a third approach that blends the two. Recently most prominent of these are KUIELABMDX-Net [32] which has a time-frequency branch and a time-domain branch, where each branch separates stems respectively, and Hybrid Demucs which uses a bi-U-Net structure with a shared backbone. The latter was the first ranked architecture at the 2021 MDX MSS Competition [15]. The current state-of-the-art model on MUSDB18 is HT Demucs [33] which adds a transformer to a Hybrid Demucs architecture. This paper will focus on the role of the transformer in the field of MSS and more specifically, in the Hybrid Demucs architecture.

3 Literature review

In this section, a comprehensive review of the literature on music source separation is presented, following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) methodology. The review aims to identify and analyse relevant studies published in the field, providing an overview of the current state of knowledge in music source separation techniques.

First, a comprehensive search of electronic databases was conducted, including IEEE Xplore, arXiv and Google Scholar, using relevant keywords such as "music source separation," "audio source separation," "audio signal processing," and "music signal processing." Additionally, the references of the selected papers were manually searched to identify any additional relevant studies. Between 2014 and 2023, 1325 published articles in total were found before applying exclusion criteria shown on table 1. After removing duplicates and studies not in English, 1304 articles remained for screening. Having discarded articles which only have abstracts and applying the exclusion criteria, 58 studies were included in this review.

For each selected study, key information was extracted, including the authors, publication year, research objectives, methodologies, datasets used, and key findings. The studies were then categorized based on their primary approach or technique in music source separation. Most of the studies focused on the following approaches:

1. Traditional methods: Several studies proposed techniques based on the analysis of the spectral content of audio signals, such as Wiener filtering, non-negative matrix factorization (NMF) and independent component analysis (ICA). These methods aimed to separate audio sources based on the assumption that different sources exhibit distinct spectral characteristics.
2. Deep learning-based methods: With the advent of deep learning, numerous studies explored the use of neural networks, particularly convolutional neural networks (CNNs) and recurrent neural networks

(RNNs). Recent successes of the transformer architecture in NLP have also merited interest from the MSS community. These approaches leveraged large, annotated datasets and achieved remarkable performance improvements compared to traditional methods.

3. Hybrid methods: Some studies proposed hybrid approaches that combined multiple techniques to improve the separation performance. For instance, combining spectral-based methods with temporal-based methods or incorporating prior information from musical scores.

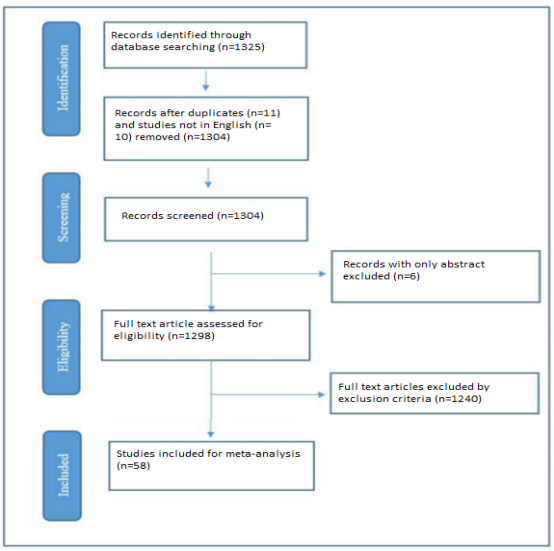

Figure 1. Flow diagram of database searching using PRISMA

Overall, the reviewed studies demonstrate significant advancements in music source separation techniques. Deep learning-based methods have shown remarkable success in separating complex audio sources. These methods leverage the power of neural networks and large annotated datasets to learn intricate representations of audio signals and achieve state-of-the-art performance. However, despite the advancements, challenges remain in achieving desired quality of separated stems which still have artifacts and considerate levels of distortion. Future research should focus on addressing these challenges, exploring novel approaches, and developing evaluation metrics that better capture the perceptual quality and timbral accuracy of the separated sources.

It is important to acknowledge some limitations of this literature review. Firstly, the review focused primarily on studies published in peer-reviewed journals and conference proceedings, potentially overlooking relevant work in unpublished reports or non-English publications. Secondly, the search keywords and inclusion criteria might have inadvertently excluded some studies that could have been valuable for this review.

In the next section, the attention mechanism is explained. This mechanism is at the core of transformer architecture and is therefore a prerequisite for understand how transformers work.

Inclusion criteria	Exclusion criteria
Papers which are well cited	Papers in which only the abstract is available
Papers which are open access	Duplicate records
Articles published in English	Papers not written in the English language
Papers providing clear information about the datasets and sample size	Papers not reporting sample size
Papers published after 2014	Papers published before 2014
Published in conferences/journals	Papers without appropriate evaluation metrics
Peer reviewed	Poor performance

Table 1. Exclusion and inclusion criteria.

4 Attention mechanism

Before the introduction of the attention mechanism, traditional seq-to-seq models relied on fixed-length context vectors to encode the entire input sequence. This approach had limitations when dealing with long sequences because the fixed-length context vector struggled to retain all the necessary information. The first paper which brought the idea of attention mechanism to the world was Bahdanau et al. 2015 [34]. It proposes the encoder-decoder model with an additive

attention mechanism. The encoder processes the input sequence and generates a set of contextual representations or hidden states. These representations capture important information about the input sequence and serve as the basis for subsequent decoding. The attention mechanism revolutionized this paradigm by introducing a dynamic and adaptive approach to incorporating context information. Instead of using a fixed-length context vector, the attention mechanism allows the model to selectively attend to different parts of the input sequence, giving higher weights to more relevant elements. This selective focus enables the model to effectively handle long sequences and capture the dependencies between the input and output.

There are different types of attention mechanisms of which the most interesting one for this paper is the self-attention. Self-attention is a mechanism that allows a sequence to attend to other elements within itself, enabling the model to capture relationships and dependencies between different positions in the same sequence. This mechanism has been introduced in the revolutionary paper „Attention is all you need“ by Vaswani, Ashish, et al. [35] as a part of a groundbreaking transformer architecture which is explored in the next section.

5 The transformer architecture

The transformer architecture was proposed as an alternative to recurrent neural networks (RNNs) which were the dominant models in sequence modelling at the time. Unlike RNNs and their variants, such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), the transformer relies solely on the attention mechanism to capture dependencies between different elements of a sequence. (You can find more about these architectures in the book “Deep Learning” by Goodfellow, Bengio and Courville [36])

This architecture follows an encoder-decoder structure which solves the issue of input and output sequences having different lengths and structures which makes it difficult to directly map the input to the output. The encoder accepts an input sequence and generates a compact, fixed-length vector representation known as a hidden or "latent representation." This condensed form aims to have the crucial information from the input sequence. Subsequently, the decoder utilizes this latent representation to create an output sequence. The foundational elements for constructing the encoder-decoder architecture predominantly revolve around neural networks. The following figure is a simple representation of the encoder-decoder structure followed by a transformer architecture which is built upon it.

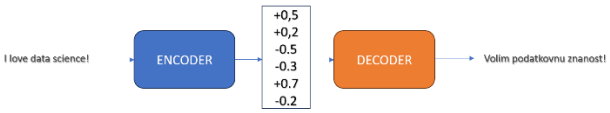

Figure 2. Encoder-Decoder structure

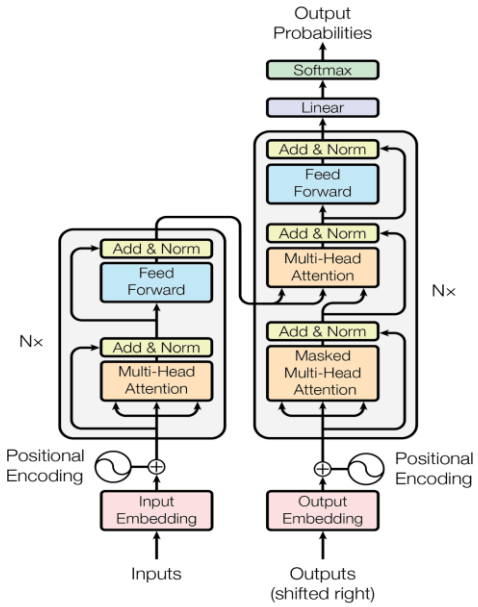

Figure 3. The encoder-decoder structure of the Transformer architecture [35]

The encoder, on the left half of the Transformer architecture, maps an input sequence to a sequence of continuous representations, which is then fed into a decoder. The decoder, on the right half of the architecture, receives the output of the encoder together with the decoder output at the previous time step to generate an output sequence. Before the input gets to the encoder it passes through an embedding block which converts the human-readable array of strings (a word) into its’ machine-readable representation, a vector of numbers.

The transformer does not make use of recurrence therefor a positional encoding is required for it to be able to differentiate between elements based on their relative positions in the sequence. The encoder and decoder consist of modules stacked on each other several times (represented as Nx in the image). To mitigate the vanishing gradient problem, the transformer architecture employs skip connections (residual connections), which allow the direct flow of information from one layer to another. Specifically, skip connections are added around each sub-

module (e.g., self-attention and feed-forward networks) in both the encoder and decoder.

5.1 Encoder

The encoder is composed of a stack of N = 6 identical layers. Each layer has two sub-layers. The first is a multi-head self-attention mechanism, and the second is a simple, position wise fully connected feed-forward network. A residual connection around each of the two sub-layers is employed, followed by layer normalization.

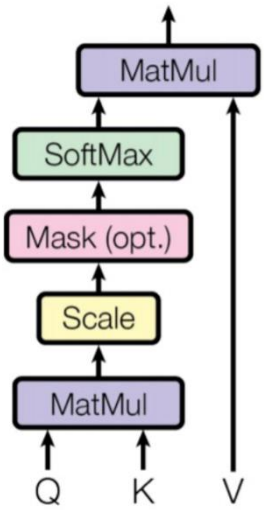

Figure 4. Scaled dot-product attention [35]

The following equation describes the attention mechanism in the above image:

$$Attention(Q, K, V) = softmax\left(\frac{(QK^T)}{\sqrt{d_k}}\right)V$$

The dot products of Q and K are scaled down by $\sqrt{d_k}$ for better normalization.
Here is the explanation for Q, K and V vectors:

- The Query vector represents the element that is currently being considered or queried. For each position in the input sequence, a Q vector is derived, capturing information about that specific position. The Q vectors are obtained by multiplying the input sequence with a learned weight matrix during the encoding process.
- The Key vector represents the elements in the input sequence that are compared against the Query vector. It provides context and helps determine the relevance or importance of each element in relation to the Query vector. Like the Q vector, the K vectors are obtained by multiplying the input sequence with another learned weight matrix during the encoding process.
- The Value vector carries the actual information associated with each element in the input sequence. It represents the content or meaning of the elements.

The self-attention mechanism calculates the attention scores between the Query (Q) and Key (K) vectors to determine the relevance or importance of each element. The attention scores are transformed into attention weights by applying the softmax function to ensure they sum up to 1. Finally, the attention weights are used to compute a weighted sum of the Value (V) vectors. The Value vectors are multiplied by the corresponding attention weights, and the resulting weighted sum represents the aggregated information or context from the entire input sequence.

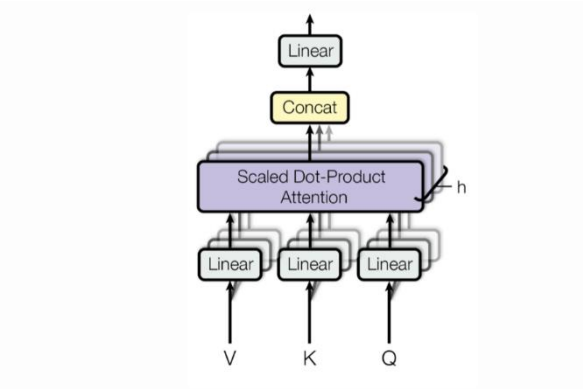

Figure 5. Multi-Head Attention [35]

The image above describes how this attention mechanism is parallelized. The multi-head attention mechanism enables the model to pay attention to multiple parts of the key simultaneously.

5.2 Decoder

The decoder also consists of a stack of N = 6 identical layers that are each composed of three sublayers:

1. The first sublayer receives the previous output of the decoder stack, augments it with positional information, and implements multi-head self-attention over it. The decoder is modified to attend *only* to the preceding words, and this is achieved by introducing a mask over the values produced by the scaled multiplication of matrices Q and K.
2. The second layer implements a multi-head self-attention mechanism like the one implemented in the first sublayer of

the encoder. On the decoder side, this multi-head mechanism receives the queries from the previous decoder sublayer and the keys and values from the output of the encoder. This allows the decoder to attend to all the words in the input sequence.

3. The third layer implements a fully connected feed-forward network, like the one implemented in the second sublayer of the encoder.

6 Transformer in HT Demucs

Hybrid Transformer Demucs (HT Demucs) is a hybrid temporal/spectral bi-U-Net based on Hybrid Demucs [37], where the innermost layers are replaced by a cross-domain transformer encoder, using self-attention within one domain, and cross-attention across domains. The cross-domain Transformer Encoder simultaneously processes the 2D signal from the spectral branch and the 1D signal from the waveform branch. Unlike the original Hybrid Demucs, which required meticulous parameter tuning (such as STFT window, hop length, stride, padding, etc.) to align the time and spectral representation, this new architecture can handle heterogeneous data shapes, making it more adaptable and flexible.

In the following figure, a single self-attention encoder layer of the transformer is shown. It has normalizations before the self-attention and feed-forward operations which stabilizes training. The first two normalizations are performed independently on each token, known as layer normalizations. The third normalization, however, is applied collectively to all tokens, referred to as time layer normalization.

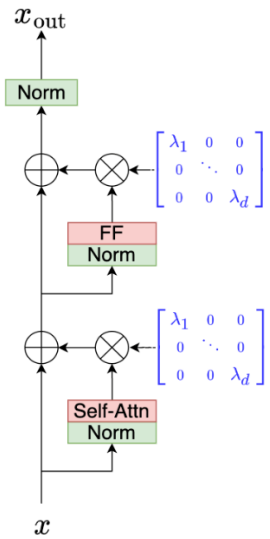

Figure 6. Self-attention Encoder layer of the Transformer [33]

The cross-attention encoder layer is unchanged, but it incorporates cross-attention with the representation from the other domain. In the following figure, there is a depiction of a cross-domain transformer encoder consisting of five layers. It alternates between self-attention encoder layers and cross-attention encoder layers in both the spectral and waveform domains. Additionally, 1D and 2D sinusoidal encodings are added to the scaled inputs, and the spectral representation is reshaped to be treated as a sequence.

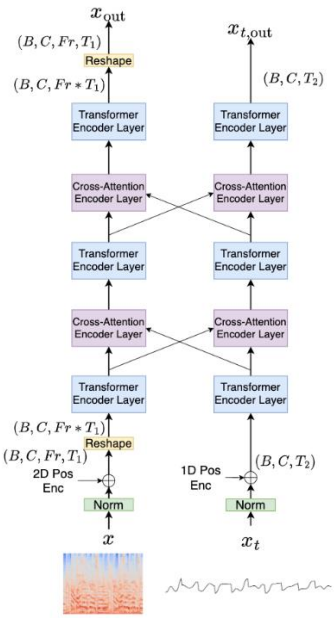

Figure 7. The Cross-domain Transformer Encoder [33]

Finally, the hybrid transformer demucs architecture is shown on figure 8.

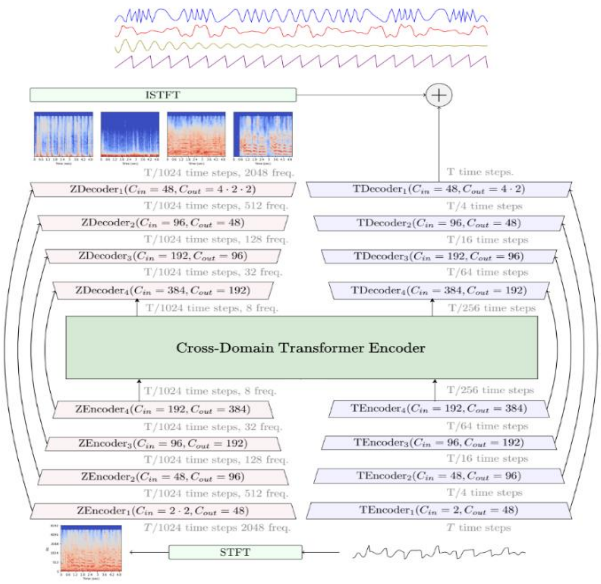

Figure 8. Hybrid Transformer Demucs architecture [33]

The input waveform is processed both through a temporal encoder and a spectral encoder (the latter is prefixed by STFT). The decoder is built symmetrically. The output spectrogram goes through the ISTFT and is summed with the waveform outputs, giving the final model output. The Z prefix is used for spectral layers, and T prefix for the temporal ones. The advantage of the hybrid approach is that the model can learn strong temporal structure from sources, such as drums, while using the spectrogram domain for sources that have a more harmonic structure, such as vocals. This way the model is taking advantage of strengths of both domains rather than being limited to one of them.

7 Experiment

In this experiment, the goal was to evaluate the performance of a hybrid transformer Demucs model in separating songs into four stems: voice, drums, bass, and other. The model was trained on MUSDB18 dataset + additional 800 songs and has shown promising results [33]. To test the model, 4 short song examples (~1 minute duration) were taken from the DSD100 dataset [38]. To achieve better results, a fine-tuned version of the algorithm was used.

7.1 Evaluation

Measuring the results of a source separation approach is a challenging problem. Generally, there are two main categories for evaluating the outputs of a source separation approach: objective and subjective. Objective measures rate separation quality by performing a set of calculations that compare the output signals of a separation system to the ground truth isolated sources. Subjective measures involve having human raters give scores for the source separation system's output.

Both objective and subjective measures have their advantages and disadvantages. Objective measures face limitations in capturing various aspects of human perception solely through computational methods. Nevertheless, they offer the advantages of speed and cost-effectiveness compared to subjective measures. Conversely, subjective measures are characterized by their high cost, time-consuming nature, and susceptibility to the variability of human raters. However, they can provide greater reliability than objective measures due to the involvement of actual human listeners in the evaluation process.

When it comes to objective measures, Source-to-Distortion Ratio (SDR), Source-to-Interference Ratio (SIR), and Source-to-Artifact Ratio (SAR) are, to date, the most widely used methods for evaluating a source separation system's output. Signal-to-Noise Ratio (SNR) is not used as widely but does appear sometimes in source separation. In this paper, only SDR is used for the sake of simplicity and conciseness.

8 Results

To get results, a google colab version of the code was used from:
https://github.com/facebookresearch/demucs.
In addition to this code, a museval library was used to evaluate the model and results are the following:

	Drums (SDR)	Bass (SDR)	Other (SDR)	Vocals (SDR)
Song1	-1.423	-2.372	-3.598	-4.430
Song2	-2.948	-1.425	7.992	-4.810
Song3	-0.735	-27.114	-3.138	-0.433
Song4	-2.183	-1.696	-2.548	-4.805
Avg	-1.822	-8.152	-0.323	-3.620

Table 2. SDR measures for 4 snippet songs

The model's performance in terms of Source-to-Distortion Ratio (SDR) varies across the different stems and songs. The negative SDR values suggest that the estimated sources have higher distortion or interference compared to the reference sources. Each song in the evaluation has different separation results across the stems. This indicates that the complexity of the songs, the composition of the mixtures, and the characteristics of the individual sources can influence the model's separation performance.

For comparison, the HT Demucs [33] is evaluated at a 9.20 dB of SDR with extra training data and using sparse attention kernels and per source fine-tuning. The evaluation done in this paper deviates tremendously from the already mentioned state-of-the-art performance.

9 Discussion

One of the fundamental reasons for disparities in results could be the disparity in dataset size and diversity. The original paper had access to 950 songs which include a broader range of musical genres, recording conditions, and audio quality. In contrast,

when evaluating a model on a smaller dataset with only a few songs, the results may not generalize well. In addition, the original paper used the library *BSSEvalv4* (https://github.com/sigsep/bsseval) whereas this paper used the library *museval* (https://sigsep.github.io/sigsep-mus-eval/).
This is due to simplicity of use and limitations provided by the google colaboratory environment which is itself also a contributing factor to the huge deviation from the original evaluation results.

10 Conclusion

In the pursuit of advancing the field of Music Source Separation (MSS), this paper has delved into the Transformer architecture and its application to the challenging task of isolating individual audio sources from music mixtures. While this architecture has shown promise, it is imperative to acknowledge some of the limitations and considerations that warrant the attention of the MSS community. One notable limitation in the current landscape of MSS research is the community's heavy reliance on the MUSDB18 dataset. Creation and use of diverse datasets that encompass various musical genres, audio quality levels, and recording conditions should be encouraged. This would enable the development and evaluation of models that better generalise.

Moreover, reproducing the results reported in MSS papers can often be challenging due to computational demands and information accessibility. Collaboration of any kind should be maximally promoted, which means open-sourcing projects: sharing the code, pretrained models and benchmarks. Also, models can be made more accessible to a broader range of researchers by exploring methods for optimizing efficiency to mitigate hardware limitations. Lastly, rigorous documentation of experiments, including detailed descriptions of model architectures, hyperparameters, and evaluation protocols can aid others in reproducing research findings.

In conclusion, addressing dataset limitations, hardware constraints, and promoting community collaboration are pivotal steps towards furthering research in this field.

References

[1] Zafar Rafii, Antoine Liutkus, Fabian-Robert Stoter,Stylianos Ioannis Mimilakis, and Rachel Bittner, “The musdb18 corpus for music separation,” 2017.

[2] Zafar Rafii, Antoine Liutkus, Fabian-Robert Stoter, Stylianos Ioannis Mimilakis, and Rachel Bittner, “Musdb18-hq - an uncompressed version of musdb18,” Aug. 2019.

[3] Mark D Plumbley, Samer A Abdallah, Juan Pablo Bello, Mike E Davies, Giuliano Monti, and Mark B Sandler. Automatic music transcription and audio source separation. Cybernetics &Systems, 33(6):603–627, 2002.

[4] Hiromasa Fujihara, Masataka Goto, Jun Ogata, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G Okuno. Automatic synchronization between lyrics and music cd recordings based on viterbi alignment of segregated vocal signals. In Eighth IEEE International Symposium on Multimedia (ISM'06), 257–264. IEEE, 2006.

[5] Toni Heittola, Anssi Klapuri, and Tuomas Virtanen. Musical instrument recognition in polyphonic audio using source-filter model for sound separation. In ISMIR, 327–332. 2009.

[6] Annamaria Mesaros and Tuomas Virtanen. Automatic recognition of lyrics in singing. EURASIP Journal on Audio, Speech, and Music Processing, 2010(1):546047, 2010.

[7] Bidisha Sharma, Rohan Kumar Das, and Haizhou Li. On the importance of audio-source separation for singer identification in polyphonic music. In INTERSPEECH, 2020–2024. 2019.

[8] Daniel Stoller, Sebastian Ewert, and Simon Dixon. Jointly detecting and separating singing voice: a multi-task approach. In International Conference on Latent Variable Analysis and Signal Separation, 329–339. Springer, Cham, 2018.

[9] C. Macartney and T. Weyde, “Improved speech enhancement with the wave-u-net,” arXiv preprint arXiv:1811.11307, 2018.

[10] R. Giri, U. Isik, and A. Krishnaswamy, “Attention wave-u-net for speech enhancement,” in 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). IEEE, 2019, pp. 249–253.

[11] A. Defossez, G. Synnaeve, and Y. Adi, “Real time speech enhancement in the waveform domain,” Proc. Interspeech, pp. 3291–3295, 2020.

[12] T. Jenrungrot, V. Jayaram, S. Seitz, and I. Kemelmacher-Shlizerman, “The cone of silence: Speech separation by localization,” Advances in Neural Information Processing Systems, vol. 33, pp. 20 925–20 938, 2020.

[13] H. Liu, Q. Kong, Q. Tian, Y. Zhao, D. Wang, C. Huang, and Y. Wang, “VoiceFixer: Toward general

speech restoration with neural vocoder," arXiv preprint arXiv:2109.13731, 2021.

[14] Fabian-Robert Stöter, Antoine Liutkus, and Nobutaka Ito. The 2018 signal separation evaluation campaign. In 14th International Conference on Latent Variable Analysis and Signal Separation, 2018.

[15] Yuki Mitsufuji, Giorgio Fabbro, Stefan Uhlich, FabianRobert Stoter, Alexandre D ¨ efossez, Minseok Kim, ´ Woosung Choi, Chin-Yun Yu, and Kin-Wai Cheuk, "Music demixing challenge 2021," Frontiers in Signal Processing, vol. 1, jan 2022.

[16] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. Independent component analysis. John Wiley & Sons, 2004.

[17] P. Smaragdis, C. Fevotte, G. J. Mysore, N. Mohammadiha, and M. Hoffman. Static and dynamic source separation using nonnegative factorizations: A unified view. IEEE Signal Processing Magazine, 31 (3), 2014.

[18] Sam T. Roweis. One microphone source separation. In Advances in Neural Information Processing Systems, 2001.

[19] Francis Bach and Michael I. Jordan. Blind one-microphone speech separation: A spectral learning approach. In Advances in neural information processing systems, 2005.

[20] Emad M. Grais, Mehmet Umut Sen, and Hakan Erdogan. Deep neural networks for single channel source separation. In International Conference on Acoustic, Speech and Signal Processing (ICASSP), 2014.

[21] Stefan Uhlich, Franck Giron, and Yuki Mitsufuji. Deep neural network based instrument extraction from music. In International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015.

[22] Stefan Uhlich, Franck Giron, and Yuki Mitsufuji. Deep neural network based instrument extraction from music. In International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015.

[23] Jen-Yu Liu and Yi-Hsuan Yang. Denoising auto encoder with recurrent skip connections and residual regression for music source separation. In 2018 17th IEEE International Conference on Machine

[24] Naoya Takahashi and Yuki Mitsufuji. Multi-scale multi-band densenets for audio source separation. In Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). IEEE, 2017. Learning and Applications (ICMLA), 2018.

[25] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in International Conference on Medical image computing and computer-assisted intervention. Springer, 2015, pp. 234–241.

[26] N. Takahashi, N. Goswami, and Y. Mitsufuji, "MMDenseLSTM: An efficient combination of convolutional and recurrent neural networks for audio source separation," in 2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC). IEEE, 2018, pp. 106–110.

[27] Naoya Takahashi and Yuki Mitsufuji, "D3net: Densely connected multidilated densenet for music source separation," 2020.

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

[29] Yi Luo and Jianwei Yu, "Music source separation with band-split rnn," 2022.

[30] Daniel Stoller, Sebastian Ewert, and Simon Dixon, "Wave-u-net: A multi-scale neural network for endto-end audio source separation," arXiv preprint arXiv:1806.03185, 2018.

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

[32] Minseok Kim, Woosung Choi, Jaehwa Chung, Daewon Lee, and Soonyoung Jung, "Kuielab-mdx-net: A twostream neural network for music demixing," 2021.

[33] 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.

[34] Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." *arXiv preprint arXiv:1409.0473* (2014).

[35] Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

[36] Goodfellow, I., Bengio, Y., and Courville, A. Deep Learning. MIT Press, 2016.

[37] Défossez, Alexandre. "Hybrid spectrogram and waveform source separation." *arXiv preprint arXiv:2111.03600* (2021).

[38] Liutkus, Antoine, et al. "The 2016 Signal Separation Evaluation Campaign." Proceedings of the 12th International Conference on Latent Variable Analysis and Signal Separation ({LVA/ICA} 2015), Liberec, Czech Republic, August 25-28, 2015. Edited by Petr Tichavský et al., Springer International Publishing, 2017, pp. 323-332.