INTERNATIONAL RESEARCH JOURNAL ON ADVANCED SCIENCE HUB
e-ISSN : 2582 - 4376
Open Access
RSP SCIENCE HUB
(The Hub of Research Ideas)
Available online at www.rspsciencehub.com

Special Issue of First International Conference on Management, Science and Technology (ICMST 2021)

Research on DNN Methods in Music Source Separation Tools with emphasis to Spleeter

Louis Ansal C A[1], Ancy C A[2]
[1] Department of Computer Science, St Paul's College, Kalamassery, Kerala, India.
[2]Assistant Professor, Department of Computer Science, St Paul's College, Kalamassery, Kerala, India.
louisansalca@gmail.com[1], ancyroseann@gmail.com[2]

Abstract

This paper tries to attempt a review on deep neural network (DNN) method in music source separation (MSS) tools with emphasis to Spleeter by Deezer, an enhanced deep learning model for music sourceseparation. It is a set of pre-trainedmodel written in python using the Tensorflow machine learning library used for musicsource separation. It was developed by Deezer, on the need to separate a given mixed music track to its constituentinstrumental or vocal tracks usually known as stems. Spleeter offers 3 pre-trainedmodels namely 2, 4, and 5 stemseparation models that are capable of separating a given mix into 2, 4, and 5 stems respectively, which can be used forvarious needs like remixing, up-mixing,music transcription, etc. This paper is the first of its kind to review on DNN methods in MSS.In this paper, we will learn about the purpose and useof Spleeter developed by Deezer as well as about the technical aspect behind this software product that includes areas like ArtificialIntelligence (AI), Machine Learning and Deep Learning, and further about Time-Frequency (TF) masking and U-NetConvolution Neural Network (CNN) which are the methodology and architecture employed in it respectively. From thereview, we learned that Spleeter by Deezer is one of the latest advancement in MSS problem that comparatively has one of the best signal to distortion ratio (SDR), signal to artifacts ratio (SAR), signal to interference ratio (SIR), and sourceimage to spatial distortion ratio (ISR) and produce a state of the art solution, and it has also paved a way togreater development in MSS problem in the future.

Keywords:Music Source Separation,Artificial Intelligence, Machine Learning, Deep Neural Network, Convolution Neural Network, U-Net andTime-Frequency Masking.

1. Introduction

Spleeter by Deezer is a set of pre-trained models written in python using the Tensor flow machine learning library used for music source separation (MSS). These models are already trained and show state-of-the-art performance in MSS. The MSS problem has been a large research area for music signal researchers for the past few decades. It is based on the concept that songs or music recordings are a mix of separate instrumental tracks like vocals, piano, guitar, bass, drums, etc usually known as stems. MSS aims to get back the separate instrumental tracks from the given mix that is to recover the stems from the given mixed track. By doing so it opens up numerous possibilities in fields like remixing, up-mixing, music transcription, music recommendation, music classification, etc. Human brains can hear distinct parts of the mix distinctively from the rest of the mix, just by concentrate on a particular instrument; humans can isolate it in their brains. But it is not MSS; the rest of the parts of the mix will be still

audible. In MSS the separate tracks are approximated as close as possible and they are separated from the given mix, since the stem tracks in the final mixed track are processed using various effects, it increases the difficulty and challenges to separate them perfectly without any bleeds in the separated stems. For years, a lot of researches have been going on to find the ideal solution for MSS by exploring and implementing a lot of strategies, these researches have recently made significant progress, mainly due to the advancement in the fields like machine learning and deep learning methods. Spleeter by Deezer is state of the art advancement in the MSS problem, it is very fast and efficient; the GPU version separates given mixed audio files 100 times faster than real-time. Hence, it is ideal to process large datasets as well.

Spleeter provides 3 pre-trained models for separation[5]:

- 2 Stems model - Separation into 2 stems – Vocals/ accompaniment separation.
- 4 Stems model - Separation into 4 stems – Vocals/ drums/ bass/ other separation.
- 5 Stems model - Separation into 5 stems - Vocals / drums / bass / piano / other separation.

Fig.1. Source separation representation

The 2 and 4 stem models are the ones that show the best performance. Spleeter is designed in a way that it can be used straight from the command line as well as a Python library directly in any development pipeline. It can be installed with pip or be used with Docker.

1.1 Structure of the Paper

To better understand DNN methods in MSS tools with emphasisto the Spleeter by Deezer an enhanced deep learning model for music source separation, the structure of the paper is organized as follows: Chapter 2 introduces the literature reviews we have gone through, Chapter 3 discusses the background of Spleeter, Chapter 4 gives out the conclusion and describes the future scope in this field and finally, we present the references made.

2. Literature Reviews

State of art reveals that many different works were being done in MSS with traditional Machine Learning Algorithm and it is now replaced in the current era with Deep Learning Techniques. In 2006, Emmanuel Vincent et al., [1] came up with an evaluation of Blind Audio Separation in Music Systems based on some time-invariant algorithms. They studied noise distortions, wavelength separation, and its different correction mechanisms too. In 2017, Stefan Uhlich et al., [2] done another work in DNN to improve music source separation using Data Augmentation Technique and Network Blending. There they studied the separation of music into individual instrument tracks using their proposed method. In 2018, Joachim Muth et al., [3] put forward a research article on improving DNN-based Music Source Separation using a set of Phase Features. In that paper, they use the theoretical relationship between amplitude and STFT and found out that derivatives of phase are the best feature representation in MSS. In the same year, Daniel Stoller et al., [4] found out a DNN tool for MSS known as Wave-U-Net. It is an adaptation of U-Net in the 1D time-domain that repeatedly resamples feature maps of music sources. In the year 2019, RomainHennequin et al., [5] presented and released Spleeter which is a new tool for music source separation with pre-trained models. This software separated the audio files into 2, 4, or 5 stems with a single command line using pre-trained models. It uses the Tensorflow framework for fine-tuning the pre-trained model. In 2019, Inria et al., [6] studied a reference implementation for music source separation based on Deep Learning methods. The technology was called as Open-Unmix, which provides implementations for the most popular deep learning frameworks. In the same year, Alexandre Défossez et al., [7] published Demucs, which is a Deep Extractor for Music Sources with extra unlabelled data that is remixed. They considered four sources for their works: drums, bass, vocals, and other accompaniments; and came up with a RNN model that outperformed the existing state-of-the-art waveforms. In 2020, Naoya Takahashi et al., [8] published work on a Multi-dilated DenseNet for Music Separation called D3NET. Here the authors claimed the

importance of rapid growth of the receptive field in multi-resolution data and proposed this novel method as a solution to this. In the same year, another work was put forth by RyosukeSawata et.al.,[9] using Bridging Networks, which is an All-In-One tool for music separation. They performed modifications in network architecture and introduced a CrossNet structure. Results revealed that their method improved the performances of Open-Unmix, a well-known model.

3. Discussions

3.1 Technology

Spleeter by Deezer is mainly based on the sciences like Artificial Intelligence, Machine Learning, and Deep Learning.[5]

3.1.1 Artificial Intelligence

Artificial Intelligence is the intelligence mimicked by machines to act like humans that is the ability of machines to learn and solve problems mimicking the human cognitive functions. AI can be classified into 3 based on capability as weak or narrow AI, strong or general AI, and super AI, and it can also be classified into 4 based on functionality as reactive machines, limited memory, theory of mind, and self-awareness. In the context of AI, Spleeter can be considered as a weak AI in terms of its capability as it is capable of doing a specific task only, In terms of functionality, it can be considered as a limited memory AI as it uses some retained information learned from training data that was used to develop or train the pre-trained models.

3.1.2 Machine Learning

Machine learning is a branch of artificial intelligence, that is used to develop applications or models that learn from sample data or training data to make decisions and predictions and it improves its performance and accuracy with time without being programmed to do so. Machine learning can be classified into 3 as supervised, unsupervised and semi-supervised or reinforcement learning. In the context of machine learning, Spleeter can be considered as unsupervised as the models were trained with the unlabelled dataset by clustering.

3.1.3 Deep Learning

Deep Learning is a part of machine learning that imitates the human brain in processing data and creating a pattern for decision making. It uses deep networks with multiple layers to progressively learn or extract information from raw data input, these networks are also known as deep neural networks (DNN). Spleeter uses a convolutional neural network known as U-Net which is an alternative type of DNN.

3.2 Methodology

Spleeter by Deezer is open source and uses a technique called Time-Frequency (TF) masking.[5] The various musical tracks or stems in a mix are spread across the audible frequency spectrum and each one of these stems corresponds to a specific frequency range. That is, the lead vocals, drums, bass, etc would occupy different frequency bands. Hence by using TF masking, the frequencies that correspond to a particular track can be filtered out from the mix. So by filtering out each track or stems, we end up in the separated stems of the given mixed track.

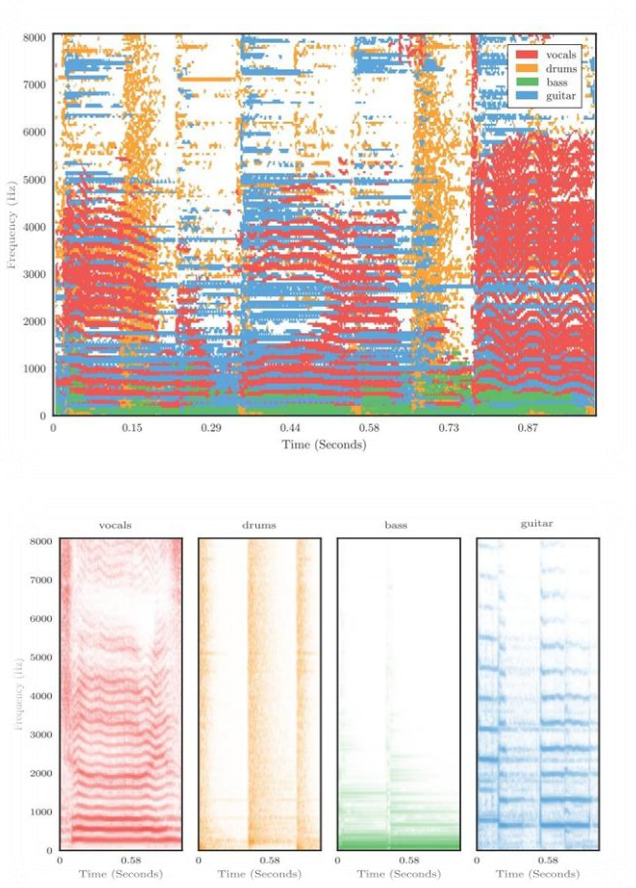

Fig.2. Time-frequency (TF) domain masking representation

In this technique, the process of approximating the frequency bands that correspond to each stem is the hardest part. With the audible frequency range of the human ear being 20Hz-20000Hz, a lot of processing is needed to accurately classify the

frequency response of each separate stem from this broad frequency range. Traditionally, this process was done manually usually on the vocals to filter out the frequencies that correspond to the lead vocals, thereby making a minus track of the original mix that is commonly used for karaoke.

Now, the Spleeter pre-trained models are capable of doing this hectic task its own. The neural networks in the pre-trained models do all the heavy lifting. It's as easy as installing a package and executing the separator function on the command line interface, which then gives out separated stems as .wav files. In addition, Spleeter also allows the users to train custom models with user datasets for source separation.

3.3 Architecture

From the research on this case study, it was understood that, Spleeter is designed upon the architecture namely U-Net. U-Net is a convolutional neural network that was initially developed for image segmentation. U-Net is a pair of convolutional encoders and decoders. In Spleeter this U-Net architecture is slightly modified with extra skip-connections to bring back the detailed information lost during the encoding stage to the decoding stage.[5] U-Net in Spleeter has 5 strided 2D convolution layers in the encoder and 5 strided 2D de-convolution layers in the decoder.

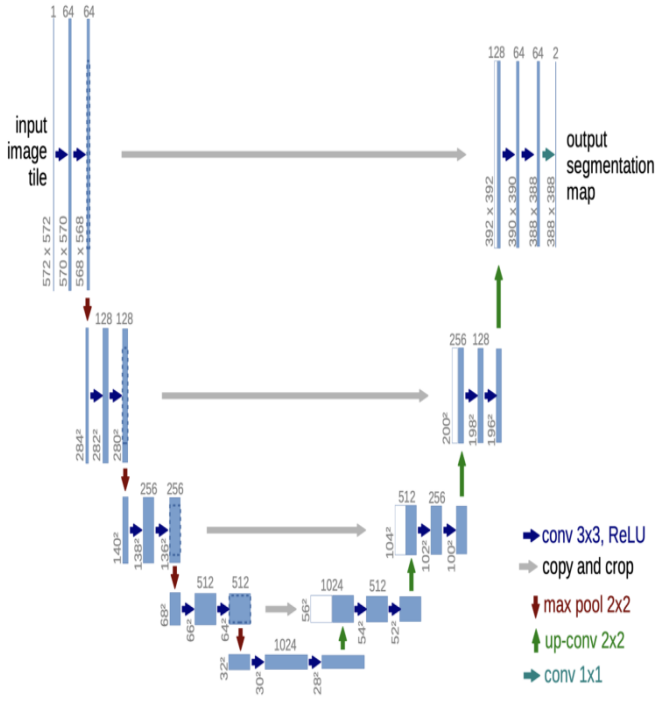

Fig.3. U-Net architecture diagram

3.4 Advantages

- Open-source

The source code is freely available and may bemodified according to the requirement of the user and redistributed.

- Free

It is freely available for all users.

- Easy to use

A set of codes will do the whole music separation.

- Fast Processing

It is very fast, the GPU version separates given mixed audio files 100 times faster than real-time which makes it a good option to process large datasets.

- State of the art audio separation

It gives out state-of-the-art audio stems after source separation.

Table.1. Performance of Spleeter MWF

STEMS	SDR	SIR	SAR	ISR
Vocal	6.86	15.86	6.99	11.95
Bass	5.51	10.30	5.96	9.61
Drums	6.71	13.67	6.54	10.69
Others	4.55	8.16	4.88	9.87

3.5. Challenges

Copyright issues

Copyright of the songs that are separated may not belong to the users; hence it will result in copyright infringement.

- Lack of perfection in separated stems

The separated audio stems are of high quality but they are not perfect they may have bleeding frequencies from other audio stem parts.

- Requires a bit of knowledge in coding

It is used or executed with a set of python codes, so a basic knowledge of coding is required.

3.6 Applications

Spleeter has numerous applications in the field of music such as software for track separation, making music recommendations, classification of music by genre, music transcription, etc.

Conclusions and Future scope

This paper tries to attempt a review onDNN method in MSS tools with a case study on Spleeter which is developed by Deezer, a set of pre-trained models written in python using the Tensorflow machine learning library used for music source separation. Various different methods are analyzed in brief along with case study on Spleeter. It is based on the sciences like AI, Machine Learning, and Deep Learning. From the review, we arrived at a general conclusion that Spleeter is a great tool for MSS, but it hasn't solved the MSS problem. Decades of researches and engineering works have built the tools on which Spleeter is based, and these researches are still going on and advancing daily. It is just a contribution to an ever-growing open ecosystem, a base for others to develop a better solution for the MSS problem. Spleeter by Deezer is just the beginning, it has shown how AI (machine learning – deeplearning - CNN) can be applied to solve the MSS problem which opens a path for great development, that will lead to the development of better models with better Signal to Distortion ratio, Signal to Artifacts ratio, Signal to Interference ratio and Source Image to Spatial Distortion ratio in the future. Spleeter is one of the best solutions at present that will lead to better solutions in the future reaching perfection in music source separation problems. Finally, Music mixing and mastering is a fine art, Spleeter in no way means disrespecting the sound engineers and artists and cause copyright infringement in any manner. Spleeter should be always used responsibly.

References

[1]. Vincent, E., Gribonval, R. and Fevotte, C. (2006). Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech and Language Processing*, 14(4), pp.1462–1469.

[2]. Uhlich, S., Porcu, M., Giron, F., Enenkl, M., Kemp, T., Takahashi, N. and Mitsufuji, Y. (2017). Improving music source separation based on deep neural networks through data augmentation and network blending. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

[3]. Muth, J., Uhlich, S., Perraudin, N., Kemp, T., Cardinaux, F. and Mitsufuji, Y. (2018). Improving DNN-based Music Source Separation using Phase Features. *arXiv:1807.02710 [cs, eess]*. [online] Available at: https://arxiv.org/abs/1807.02710

[4]. Stoller, D., Ewert, S. and Dixon, S. (2018). *Wave-U-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation*. [online] arXiv.org. Available at: https://arxiv.org/abs/1806.03185

[5]. Hennequin, R., Khlif, A., Voituret, F. and Moussallam, M. (2019). *SPLEETER: A FAST AND STATE-OF-THE ART MUSIC SOURCE SEPARATION TOOL WITH PRE-TRAINED MODELS*. [online] www.semanticscholar.org. Available at: https://www.semanticscholar.org/paper/SPLEETER%3A-A-FAST-AND-STATE-OF-THE-ART-MUSIC-SOURCE-Hennequin-Khlif/cf54151c874f9c89ddbe9e3f77e37c59688718ca.

[6]. Stöter, F.-R., Uhlich, S., Liutkus, A. and Mitsufuji, Y. (2019). Open-Unmix - A Reference Implementation for Music Source Separation. *Journal of Open Source Software*, 4(41), p.1667.

[7]. Défossez, A., Usunier, N., Bottou, L. and Bach, F. (2019). Demucs: Deep Extractor for Music Sources with extra unlabeled data remixed. *arXiv:1909.01174 [cs, eess, stat]*. [online] Available at: https://arxiv.org/abs/1909.01174v1

[8]. DeepAI. (2020). *D3Net: Densely connected multidilatedDenseNet for music source separation*. [online] Available at: https://deepai.org/publication/d3net-densely-connected-multidilated-densenet-for-music-source-separation.

[9]. Sawata, R., Uhlich, S., Takahashi, S. and Mitsufuji, Y. (2021). All for One and One for All: Improving Music Separation by Bridging Networks. *arXiv:2010.04228 [cs, eess]*. [online] Available at: https://arxiv.org/abs/2010.04228