Volume: 9 Issue: IV Month of publication: April 2021

DOI: https://doi.org/10.22214/ijraset.2021.33953

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue IV Apr 2021- Available at www.ijraset.com

A Instrument - A System for Generating Instrumental Covers

Srishti Prasad[1], Bhavesh Singh[2], Ali Abbas Rizvi[3], Purva Raut[4]
[1, 2, 3, 4]Dwarkadas J. Sanghvi college of Engineering, Mumbai, India

Abstract: In today's digital world and the fast-growing musical instrument technology, the ability to adapt the songs to any particular instrumental sound, becomes a priceless possession for musicians. With numerous existing tools which cater to remixing original music with instruments, a tool to convert human lyrics into instrumental rhythms is a need of the hour, or music composers of background song tracks. Often, music students start with listening to popular songs, and then attempt to recreate them through instruments. This paper aims at filtering out the human vocals of songs using a library Spleeter [14], and turning them into instrumental versions, creating an ideal platform for young talented music enthusiasts to explore and learn to play various instruments. The vocals are converted to covers using the Differentiable Digital Signal Processing (DDSP) library [13] and pre-trained Convolutional Neural Networks (CNNs) - Resnet 101 [15], which enables direct integration of signal processing elements with deep learning techniques. Currently in the music industry, after recording original songs with lyrics, they are re-recorded for covers. The proposed model can also help music professionals and artists in turning existing songs to a particular instrumental cover to suit their need and also allows them to release various versions of their songs, without the burden of extra effort and money.
Keyword: Differentiable Digital Signal Processing, Spleeter, Libroso, Vocals, Convolutional Neural Networks, Resnet101, Lyrics, Instrumental Rhythms.

I. INTRODUCTION

As an art form, music plays an intrinsic part in people's lives. Not only for entertainment, but there are numerous instances where learning music can affect the intelligence in children. And there are many more benefits apart from this. Music as a discipline is available in schools to help young students develop psychological strengths, cognitive attributes like expression, and soft skills simultaneously. This is highly noticeable during the instrumental music sessions in classrooms.

Recently, the Modern Era has seen a tempestuous period with change in taste and style of songs. Majority of the modern "art-music" composers have traversed the path of unconventional sounds, and have located their music in terms of tone and texture of the instrumental beats, skipping the more regular and established features of harmony and melody in music. The 20th Century music is an aesthetic stance underlying the period of change and development, demanding introduction of some advanced characteristics, which were not always present, which includes fewer lyrical melodies than other periods, and fun instrumental rhythms.

Unfortunately, the majority of the music students who think of taking up an instrument and learning, often give-up midway, because of the scarce availability of sources which allow them to hear and understand the beats and thus, to play instruments. There are lots of rules to learn, requiring an ample amount of time and no matter how young you start, there's always someone more prodigious than you. Thus there is a need to teach beginners to play a particular instrument in order to master it independently, irrespective of whether one excels in understanding notes or not. Due to the lack of instrument learning platforms and the parochial availability of song covers for certain types of instruments, it becomes necessary to develop an inexpensive system which would allow the young students to acquire diverse instrumental covers with ease.

The current system of creating covers involves professional musicians to manually understand the beats and play them on the various instruments. This process is more of a trial and error process where musicians continue to improve their covers until they sound perfect. This is a very time consuming process and also requires the musician to possess the instruments to play them and thus, to avoid this hard work there is a need for a low-cost, easily accessible system, that can automatically generate the covers not just for a single instrument, but for all possible instruments.

II. LITERATURE SURVEY

Separating vocals from songs with the help of a Convolutional Neural Network [8] is a process where Andrew J. R. Simpson et al. trained a convolutional Deep Neural Network (DNN) to output probabilistic approximations of the absolute binary mask for extraction of vocals. The audio signals from every song were categorized as vocal or non-vocal . All these signals are sampled at 44 kHz which they transform to spectrograms using Short-Time Fourier Transform (STFT) using certain parameters. A binary mask is calculated with the help of spectrograms of the sound signal, where each component of the mask is found by contrasting the magnitudes of the corresponding component.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue IV Apr 2021- Available at www.ijraset.com

This is done by allocating the mask a '0' if the vocal spectrogram has lesser magnitude and '1' otherwise. Now these spectrograms are divided into various windows and hence they have a huge dataset which they send to feed-forward deep neural networks. The biased-sigmoid activation function is used throughout the DNN with no bias for the output layer. The DNN is prepared by applying some parameter k of iterations of stochastic gradient descent [10]. The model is then utilized as a feed- forward probabilistic system after the training process.

Another system is the music/voice separation using a similarity matrix [9] which is an improvisation over the traditional approach of extracting vocals from music. In the conventional approach the principle was to distinguish the musical component from the vocal component in a musical mix, by simply separating the primary recurrent structure which was based on an assumption that the background had certain repeated patterns over specific patterns. Hence, Zafar Rafii et.al. have proposed a generalised approach for doing so considering redundancies occur irregularly or without a fixed period, consequently permitting the handling of music pieces with quick differing recurring structures and remote repeating components. In this approach they have used a similarity matrix which is a 2-D representation in which each point is used to find the dissimilarity amongst any component pairs of a sequence. As the recurrent patterns are responsible for creating the music structure, a similarity matrix derived from a sound signal might assist in revealing the musical structure that is present within it, which is later used to produce spectrograms to identify repeating elements.

R. Hennequin et.al. have managed to produce a remarkable library Spleeter [1][14] which is an efficient tool for music source separation from a given soundtrack. It contains pretrained models which are based on TensorFlow and can separate audio files into 2, 4 or 5 stems. They have also claimed that Spleeter is extremely fast and have backed their statements with statistical data where they state that it can separate a mix audio file into 4 stems about 100 times faster than real-time on a single GPU using the pre-trained 4-stems model. With the help of this open source library published by R. Hennequin et.al., vocal extraction can be easily carried out in this use case.

Human understanding is affected by general patterns as well as in-depth waveform coherence. Due to this, efficient audio synthesis becomes a fundamentally complex machine learning task. Autoregressive models, like WaveNet [4][5][6], replicate local structure although they present slow repetitive sampling and show a paucity of global hidden structure [2]. Global hidden conditioning and systematic parallel sampling, on the other hand, are present in Generative Adversarial Networks (GAN), however, GANs scuffle in producing locally-coherent sound waves. Through replication of log magnitudes and expeditious frequencies with adequate frequency resolution in the spectral domain, Jesse Engel et.al. in their paper demonstrate that GANs can produce audio which has a high degree of accuracy. After thorough tests carried out on the NSynth dataset, they portray that GANs can perform far better than strong WaveNet [4] [5] [6] standards on automatic as well as human assessment criterions, and produce sound multiple times quicker.

Jesse Engel et.al. have also proposed a novel system named Differentiable Digital Signal Processing (DDSP) which helps to combine traditional signal processing elements with deep learning methods [3][12][13]. In simpler terms, they have put forward an easily understandable and modular approach to generative modelling without losing out on the advantages of deep learning. They have proposed a complete system which links Digital Signal Processing to deep learning, retaining the benefits of strong inductive biases without sacrificing the expressive power of neural networks. Their library can be used to accommodate deep learning techniques for producing the instrumental covers from the vocals extracted using Spleeter.

III. PROPOSED METHODOLOGY

The Proposed System can be broadly divided into two modules, the first being the Vocal Extraction and the second Vocal Conversion. In this section, these modules are described in detail.

A. Vocal Extraction

In this step, the aim is to extract the human vocals form the input song with a minimal amount of noise and background notes. The system uses Spleeter [1][14], which is a deep learning based library for extracting various components of songs separately. These components include the vocals, different musical notes and noises. Once the human vocals are extracted using Spleeter, the next task is to remove the zero frequency amplitudes from the extracted vocals. This is important because the songs may or may not consist of vocals throughout its duration. Thus for efficient computation, the vocals at zero frequency amplitudes need to be trimmed. This can be carried out by the use of Notch Filter (band-stop filter). In signal processing, a band-stop filter or band-rejection filter is a filter that passes most frequencies unaltered, but attenuates those in a specific range to very low levels. Thus these filters are used to overcome the problems which provide a better computational efficiency. Once the vocal is extracted it is divided into various sub parts, and spectrograms are generated which are passed to the next module.

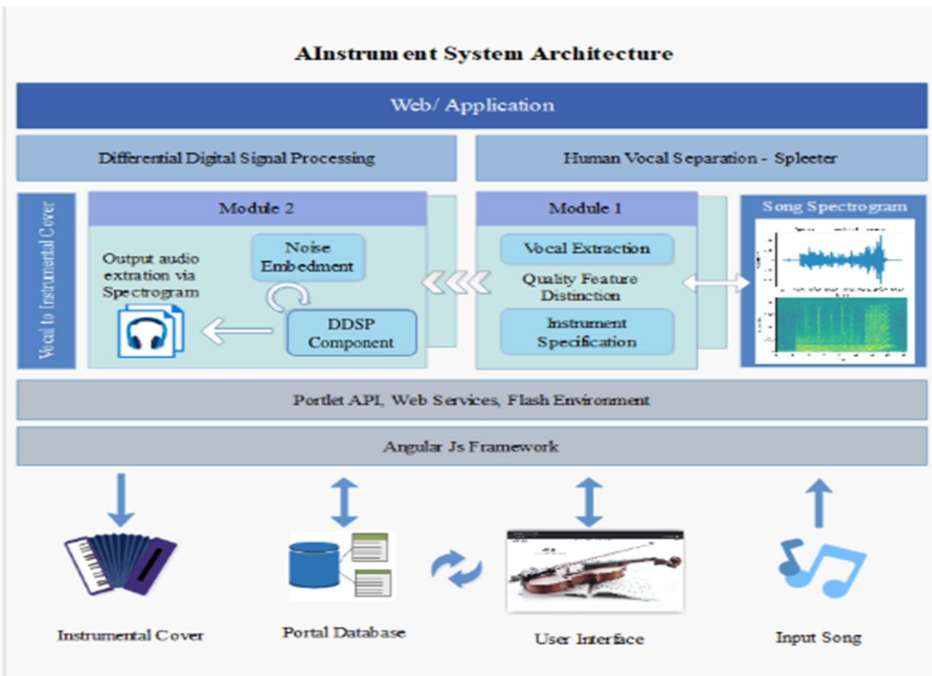
B. Vocal Conversion

The main idea is to use standard interpretable DSP elements to create complex and realistic signals by accurately controlling their parameters. For example, a combination of linear filters with sinusoidal oscillators (DSP elements) can generate the tone and rhythm of a realistic instrument if the responses and frequencies are tuned in just the appropriate way. However, it is slightly complex to dynamically control all of these parameters manually, that is why synthesizers with easy controls often sound unnatural or "synthetic". With DDSP[3], a neural network is utilised to transform a user's input into complex controls of DSP, which can produce more realistic sounding signals. This input can be any form of controlled signal, like features extracted from the audio itself. Since the DDSP components are differentiable, one can easily train the neural network to adapt to any dataset using standard backpropagation.

IV. SYSTEM WORKFLOW

Figure 1 represents the entire workflow of the system. The song is passed as input to Module 1 which uses Spleeter to extract vocals and features from songs. After extracting the vocals and features they are passed through to module 2. In module 2 there is a DDSP Component that has been trained to learn the timbre characteristics of instruments violin and flute. The timber characteristics of the learned instrument is transferred to the vocals that are generated from module 1 based on the instrument selected by the user. After this, a spectrogram for the generated cover is produced for testing and accuracy purposes.

Figure 1: System Workflow



V. RESULTS

Since this is a novel system, not many resources were previously available for usage. In order to check the accuracy of the system, actual instrumental covers played by musicians were obtained and spectrograms for the same were generated. Spectrograms are frequency vs time graphs that act as a visual representation of audio signals. After these were generated, the spectrograms for the output of the system were generated for the same instrument, and both were compared to check for the accuracy of the proposed system. The following figures show the spectrogram comparisons.

Figure 2.1 shows the spectrogram of the musicians violin cover for a song, while figure 2.2 shows the spectrogram generated by the system for the violin cover of the same song.

Similarly Figure 3.1 shows the spectrogram for the flute cover of a song played by the musician and Figure 3.2 shows the spectrogram output of the system for the same.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue IV Apr 2021- Available at www.ijraset.com

Figure 2.1: Violin Spectrogram of musician

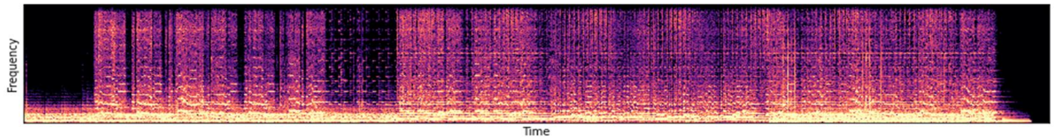


Figure 2.2: Violin spectrogram of system

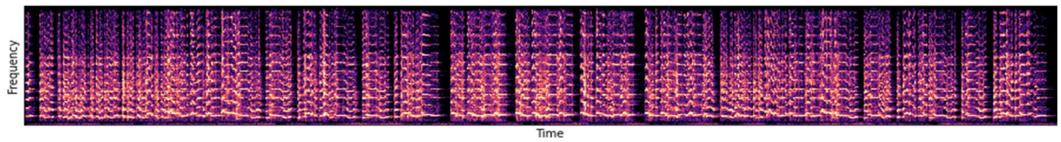


Figure 3.1: Flute spectrogram of musician

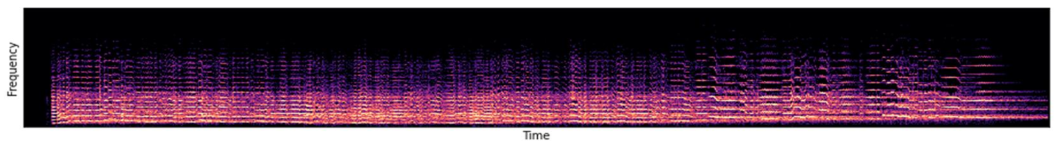
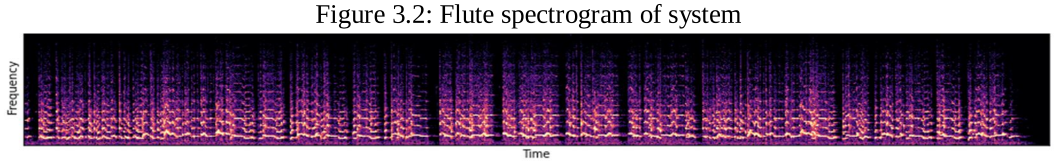


Figure 3.2: Flute spectrogram of system



VI. CONCLUSION

The voice signals of the song are easily distinguishable from the music track using Spleeter [1][14] and hence generating the cover from these vocals is possible through DDSP [3][12][13] and pretrained models. Thus, this system acts as an instrumental cover generator which will readily produce covers for a profusion of instruments, making covers easily accessible to those who rely on it for various purposes. It will have the following effects:

A. This system will remove the dependency of the instrument learning process of callow students on musicians.
B. This will enable game developers, YouTubers and yogis to more readily use instrumental covers in their games, videos and yoga sessions respectively.
C. Thus, the system will not only boost the availability of instrument covers but also provide a platform for inexpert learners to continue growing independently.

REFERENCES

[1] L. Prétet, R. Hennequin, J. Royo-Letelier and A. Vaglio, "Singing Voice Separation: A Study on Training Data," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom, 2019, pp. 506-510.
[2] Jesse Engel and Kumar Krishna Agrawal and Shuo Chen and Ishaan Gulrajani and Chris Donahue and Adam Roberts, “GANSynth: Adversarial Neural Audio Synthesis”.
[3] Jesse Engel and Lamtharn Hantrakul and Chenjie Gu and Adam Roberts,” DDSP: Differentiable Digital Signal Processing”.
[4] Aaron van den Oord and Sander Dieleman and Heiga Zen and Karen Simonyan and Oriol Vinyals and Alex Graves and Nal Kalchbrenner and Andrew Senior and Koray Kavukcuoglu, “WaveNet: A Generative Model for Raw Audio”.
[5] Comparison metrics taken from DeepMind’s Website.{last updated : Sep 8, 2016 , url : ”https://deepmind.com/blog/article/wavenet-generative-model-raw-audio”.}
[6] Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Douglas Eck, Karen Simonyan, and Mohammad Norouzi. "Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders." 2017.
[7] Book by Ian Goodfellow and Yoshua Bengio and Aaron Courville, “Deep Learning”, MIT Press 2016. @book{Goodfellow-et-al-2016, title=Deep Learning, author=Ian Goodfellow and Yoshua Bengio and Aaron Courville, publisher=MIT Press, note=url : http://www.deeplearningbook.org, year=2016}
[8] Andrew J. R. Simpson and Gerard Roma and Mark D. Plumbley, 2015 “Deep Karaoke: Extracting Vocals from Musical Mixtures Using a Convolutional Deep Neural Network”

[9] Zafar Rafii and Bryan Pardo, "Music/Voice separation using the Similarity Matrix". In 13th International Society for Music Information Retrieval Conference (ISMR 2012).
[10] Referred from {last updated : Dec 21, 2017 , url : "https://towardsdatascience.com/gradient-descent-algorithm-and-its-variants-10f652806a3".}
[11] R. Kasantikul and W. Kusakunniran, "Improving Supervised Microaneurysm Segmentation using Autoencoder-Regularized Neural Network," 2018 Digital Image Computing: Techniques and Applications (DICTA), Canberra, Australia, 2018, pp. 1-7.
[12] Jay K. Patel and E. S. Gopi, "Musical Notes Identification using Digital Signal Processing". In 3rd International Conference on Recent Trends in Computing 2015 (ICRTC-2015) Published by Elsevier B.V.
[13] Architecture Referred from the official magenta/ddsp website {last updated : Jan 15, 2020 , url : "https://magenta.tensorflow.org/ddsp".} .
[14] AI tool Spleeter {last updated : Nov 5, 2019 , url : "https://www.theverge.com/2019/11/5/20949338/vocal-isolation-ai-machine-learning-deezer-spleeter-automated-open-source-tensorflow".}
[15] He, K., Zhang, X., Ren, S., and Sun, J., "Deep Residual Learning for Image Recognition", <i>arXiv e-prints</i>, 2015.
[16] Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Douglas Eck, Karen Simonyan, and Mohammad Norouzi. "Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders." 2017.