Data-Driven Approach in Isolating Vocals and Instruments from Music

Department of Electronics and Computer Engineering

Institute of Engineering, Thapathali Campus

06/03/2022

Prepared by:

- **Anish Dahal (THA074BEX004)**
- **Prajwol Pakka (THA074BEX022)**
- **Sujal Subedi (THA074BEX043)**

Supervised by:
Er. Dinesh Baniya Kshatri
Lecturer

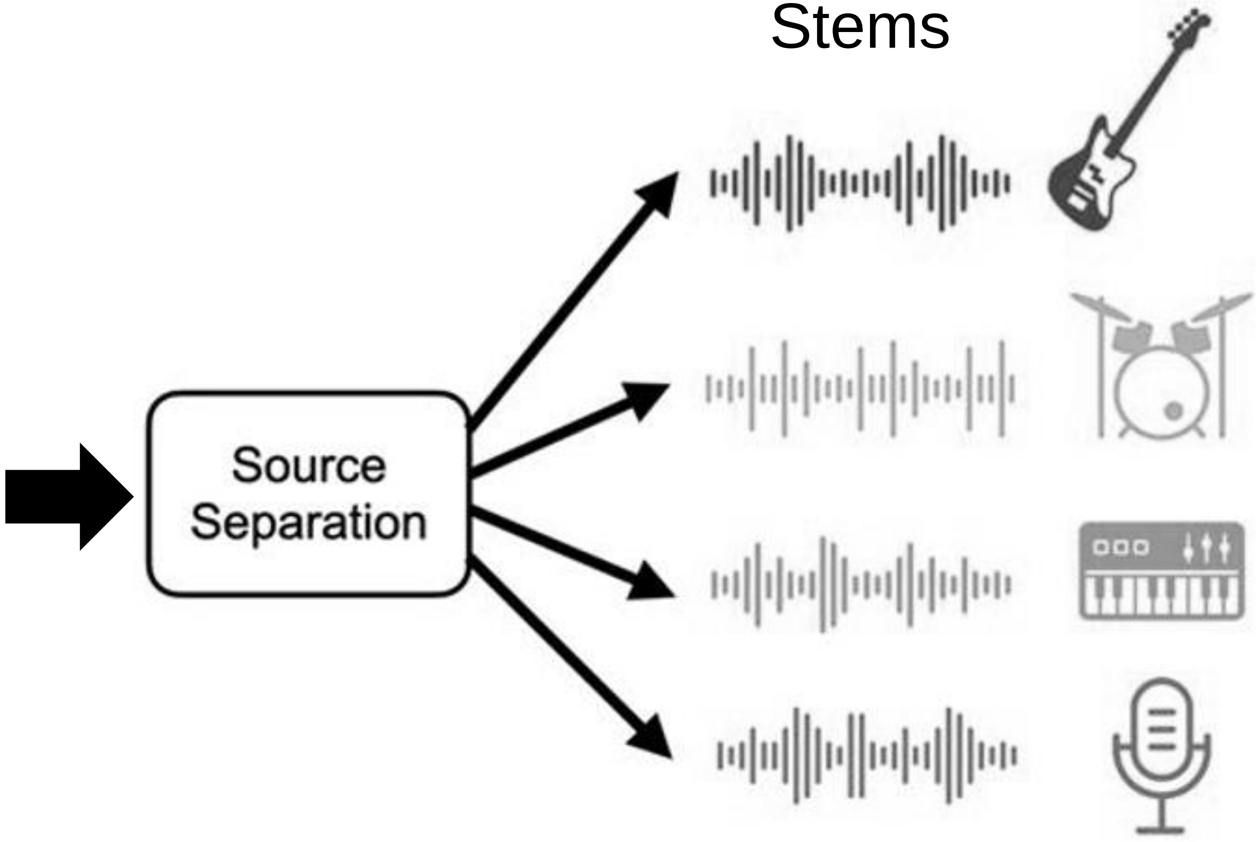
Presentation Outline

- Motivation
- Introduction
- Project Objectives
- Scope of Project
- Project Applications
- Methodology
- Results
- Analysis and Discussion
- Future Enhancement
- Conclusion
- References

Motivation

Mixture

Stems



Introduction

- Two approaches are used: 2DFT approach and Machine learning approach

- Only vocal and instrumental are separated by 2DFT approach

- Vocal, Instrumental, Drum and Bass stems can be separated with Machine Learning Approach

Project Objectives

- To separate vocals and instrumentals from a song
- To isolate drum and bass stems from the instrumentals

Scope of Project

- Project Capabilities:
  - Vocals and instrumentals can be extracted from a song
  - Instrumentals can be broken down into drums and bass stems
- Project Limitations:
  - Not all instruments can be isolated due to limited datasets
  - Might not provide satisfactory results for every songs

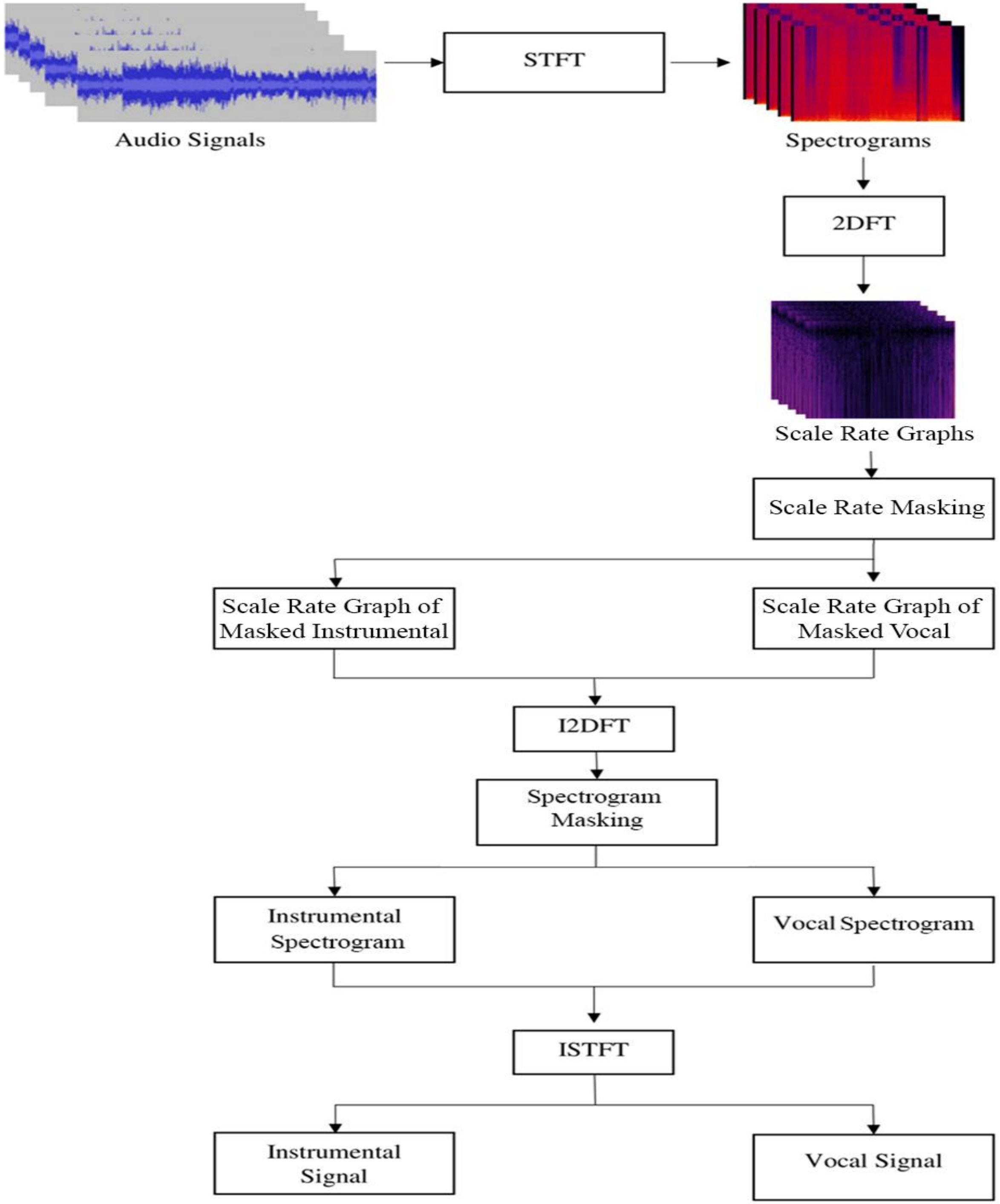
Project Applications

- Karaoke
  - Singing along prerecorded background instrumentals
- Audio Mixing
  - Extracted stems can be used to produce new audio mixtures
- Lyrics Extraction
  - Unraveling the vocals to obtain the lyrics of a song
- Singer Identification
  - Recognizing the vocalists of a song for copywrite issues

Methodology
(2DFT Processing Approach)

- Spectrograms of songs are obtained using STFT
- 2DFT of the spectrograms are taken to create scale rate graphs
- Masks generated in scale rate domain to separate the stems
- Scale rate graphs of stems are changed to spectrograms via I2DFT
- Another mask for extracted spectrograms are generated
- Masks are then applied to get spectrograms of instrumentals and vocal
- ISTFT is applied to masked spectrograms to get separated waveforms

Methodology
(2DFT Block Diagram)



Audio Signals

STFT

Spectrograms

2DFT

Scale Rate Graphs

Scale Rate Masking

Scale Rate Graph of Masked Instrumental

Scale Rate Graph of Masked Vocal

I2DFT

Spectrogram Masking

Instrumental Spectrogram

Vocal Spectrogram

ISTFT

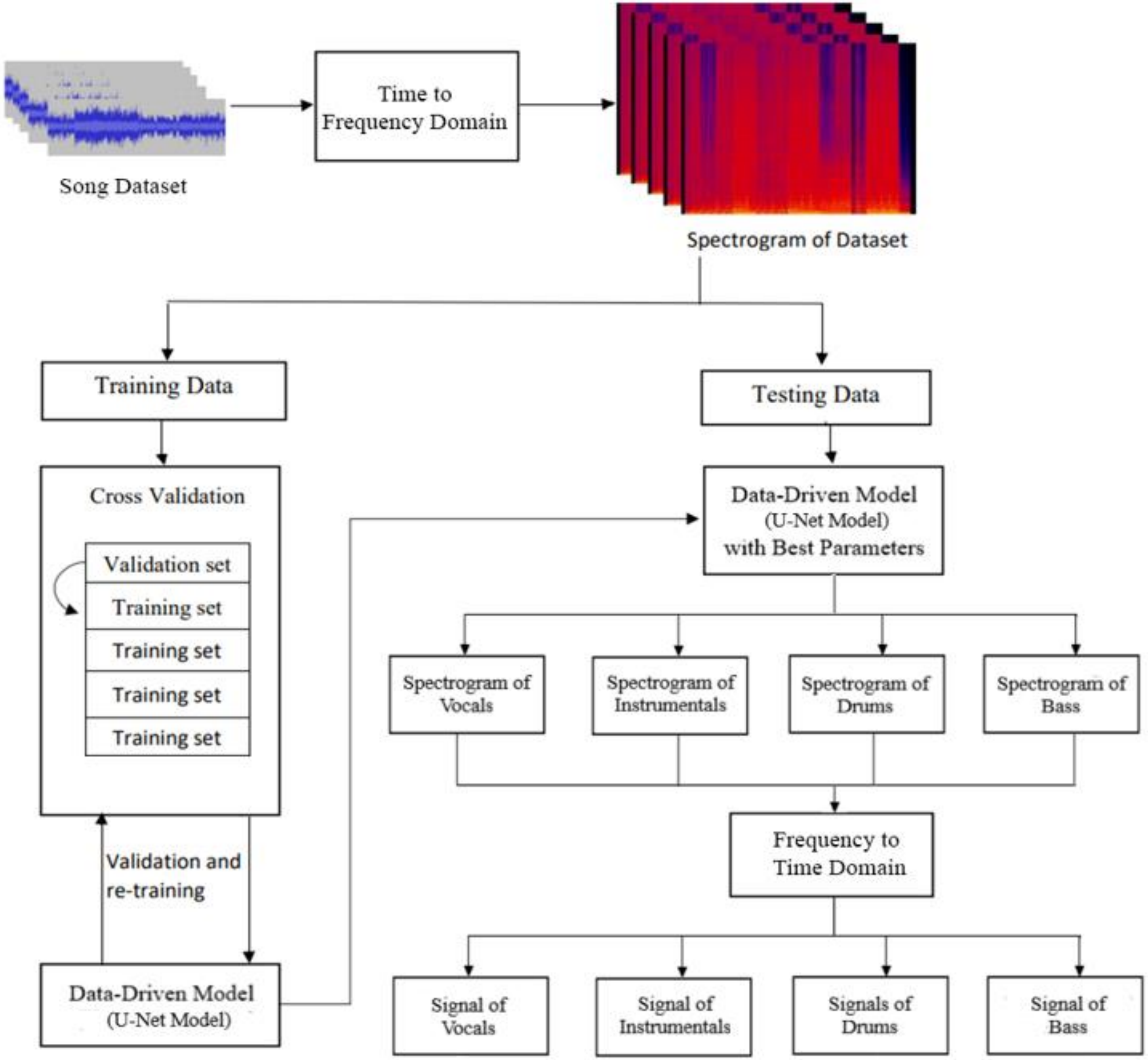
Instrumental Signal

Vocal Signal

Methodology
(Machine Learning Approach)

- Datasets of songs are available in .wav format
- Songs are changed to frequency domain with STFT
- U-Net CNN model is trained using spectrograms
- Spectrograms of Vocal, Instrumental, Bass and Drum are isolated
- Isolated stems are converted back to .wav format

Methodology
(ML Block Diagram)



Song Dataset

Time to Frequency Domain

Spectrogram of Dataset

Training Data

Testing Data

Cross Validation

Validation set

Training set

Training set

Training set

Training set

Validation and re-training

Data-Driven Model (U-Net Model)

Data-Driven Model (U-Net Model) with Best Parameters

Spectrogram of Vocals

Spectrogram of Instrumentals

Spectrogram of Drums

Spectrogram of Bass

Frequency to Time Domain

Signal of Vocals

Signal of Instrumentals

Signals of Drums

Signal of Bass

Methodology
(Dataset Splitting)

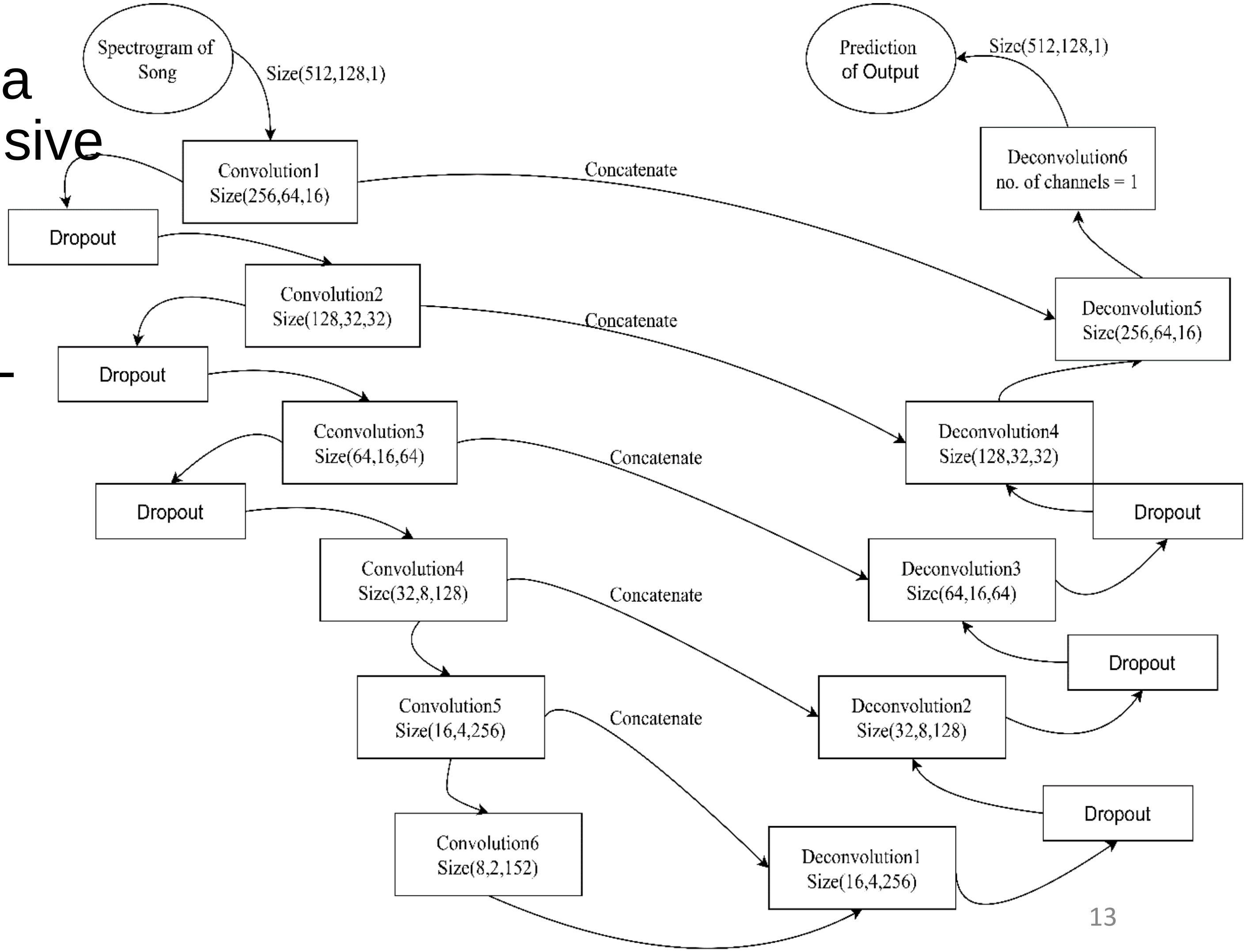
- MUSDB dataset is used which consists of 150 songs
- Training set contains 100 songs and test set contains 50 songs
- Training dataset undergo 5 fold cross validation (4:1)

Methodology
(U-Net Architecture)

- U-Net architecture consists of a contracting path and an expansive path

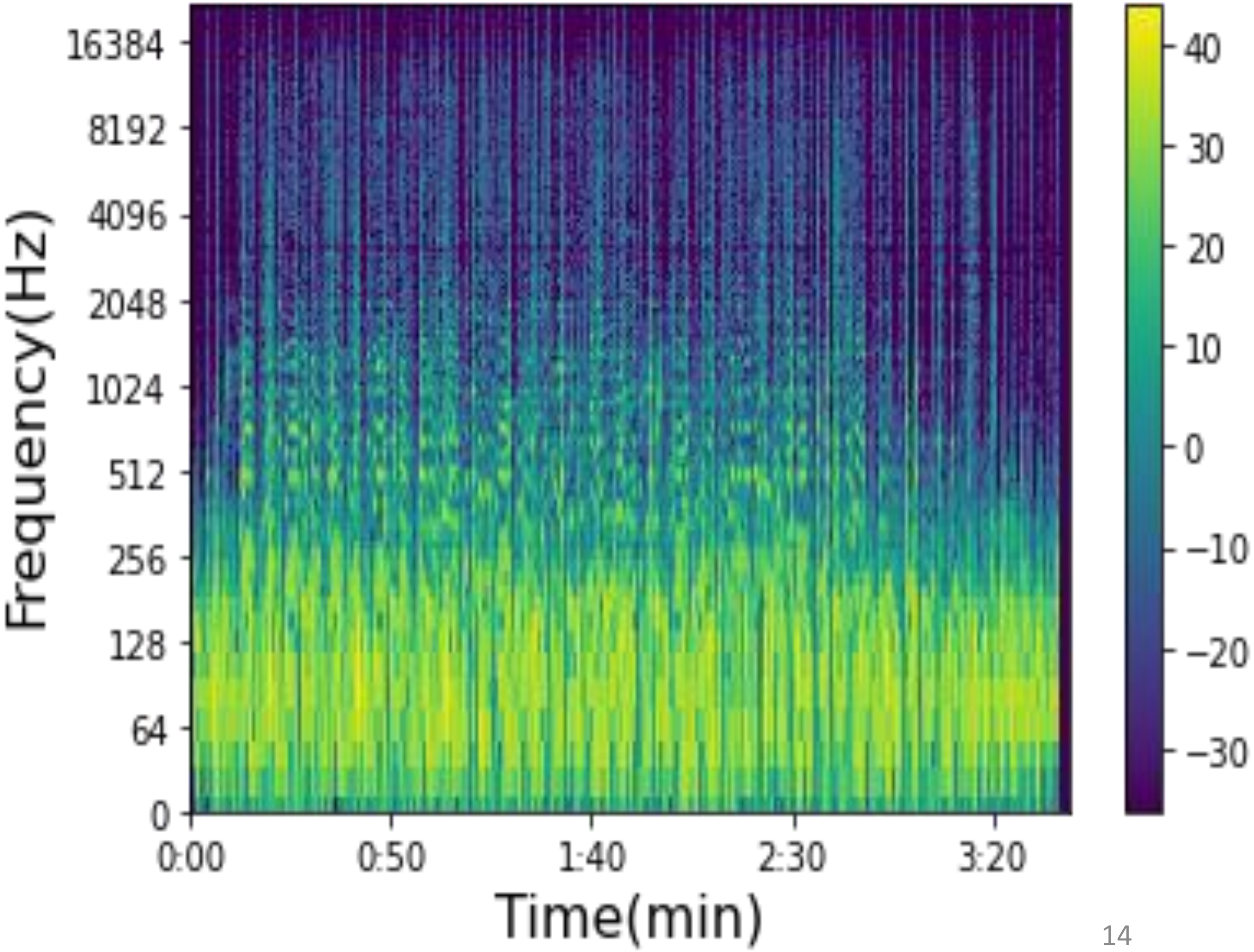
- On the contracting path, down-convolution takes place

- Up-convolution occurs on the expansive path

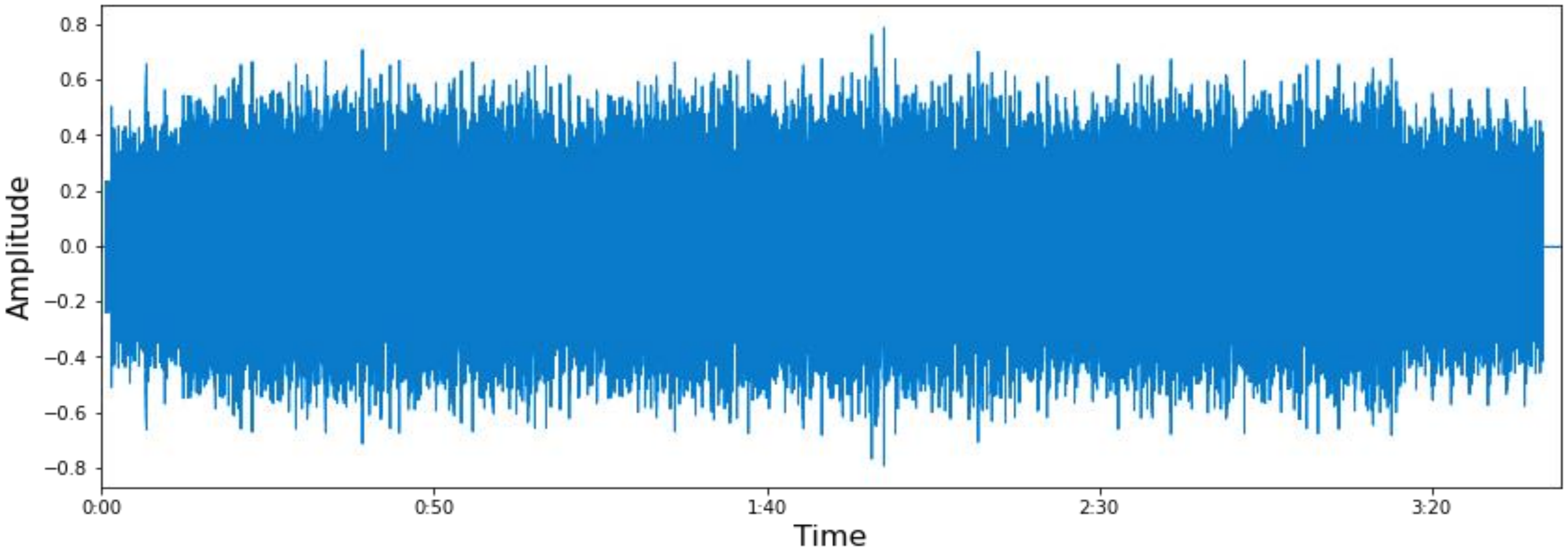

Spectrogram of Song
Size(512,128,1)
Convolution1 Size(256,64,16)
Dropout
Convolution2 Size(128,32,32)
Dropout
Cconvolution3 Size(64,16,64)
Dropout
Convolution4 Size(32,8,128)
Convolution5 Size(16,4,256)
Convolution6 Size(8,2,152)
Concatenate
Concatenate
Concatenate
Concatenate
Concatenate
Deconvolution1 Size(16,4,256)
Dropout
Deconvolution2 Size(32,8,128)
Dropout
Deconvolution3 Size(64,16,64)
Dropout
Deconvolution4 Size(128,32,32)
Deconvolution5 Size(256,64,16)
Deconvolution6 no. of channels = 1
Size(512,128,1)
Prediction of Output

Results
(Spectrogram of Selected Song)

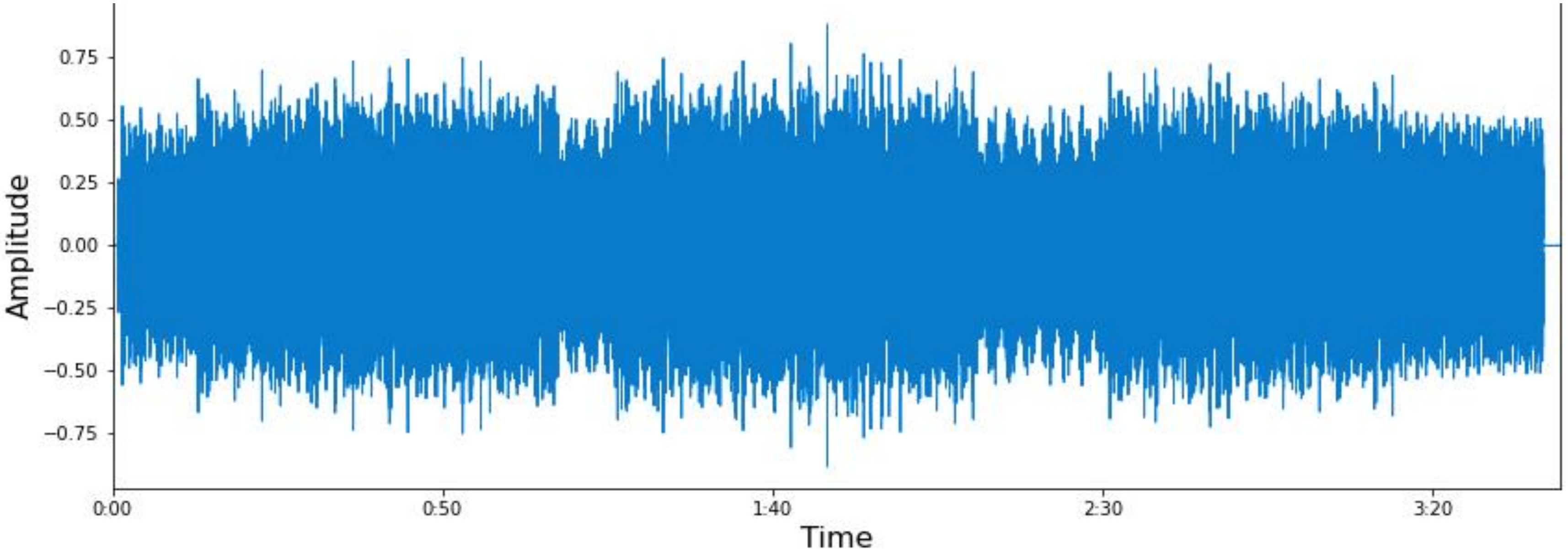
- The song “Run Run Run by Arise” is selected
- The spectrogram of the selected song is generated using STFT
- The window function used is the Hanning Window



Results (Waveform Comparison)-[1]

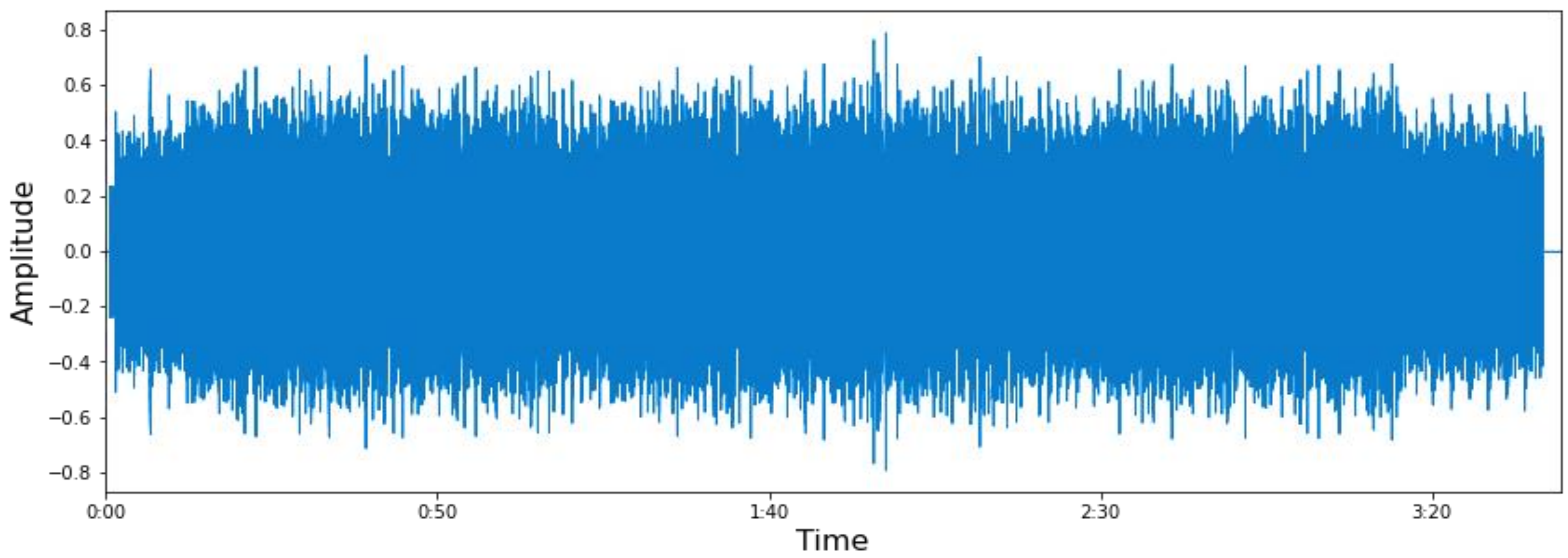


Ground-truth Instrumental Waveform of Run Run Run Song

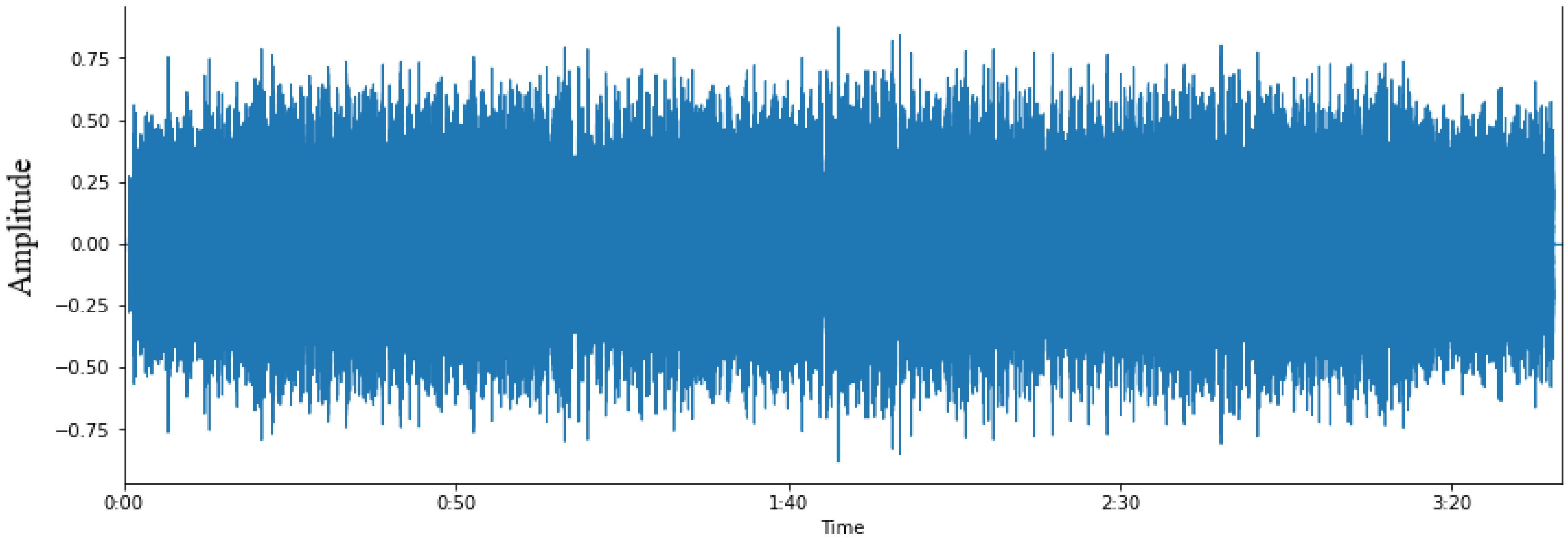


2DFT Instrumental Waveform of Run Run Run Song

Results (Waveform Comparison)-[2]

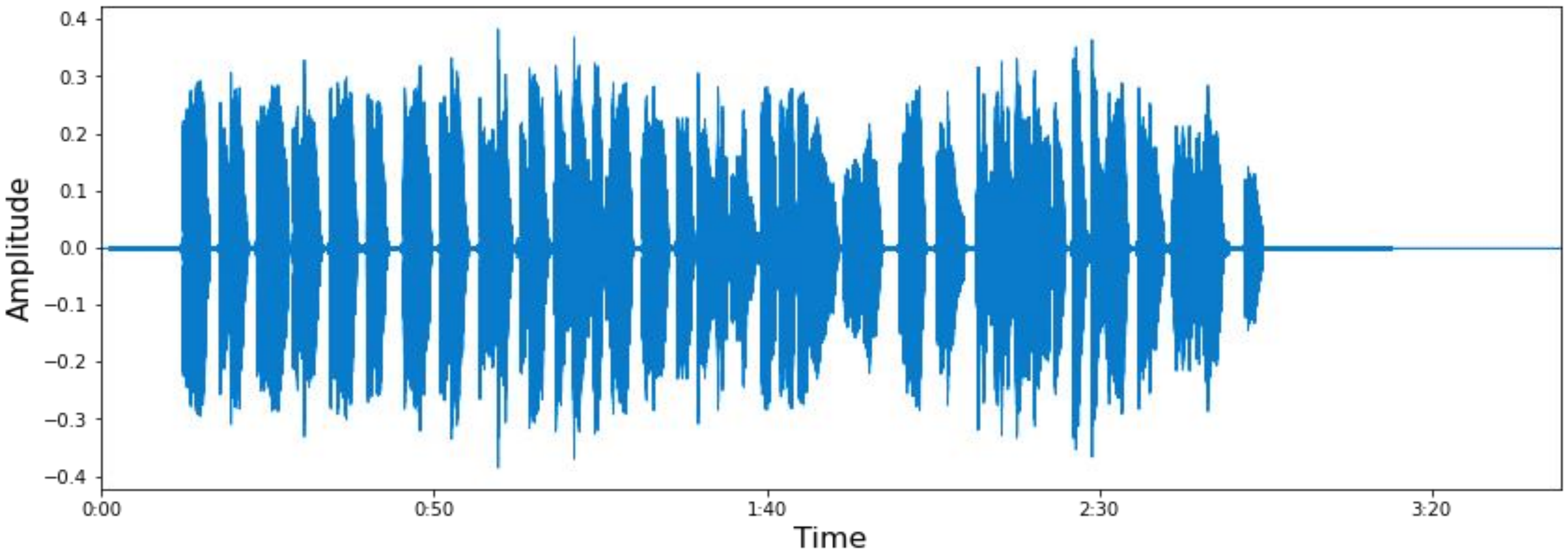


Ground-truth Instrumental Waveform of Run Run Run Song

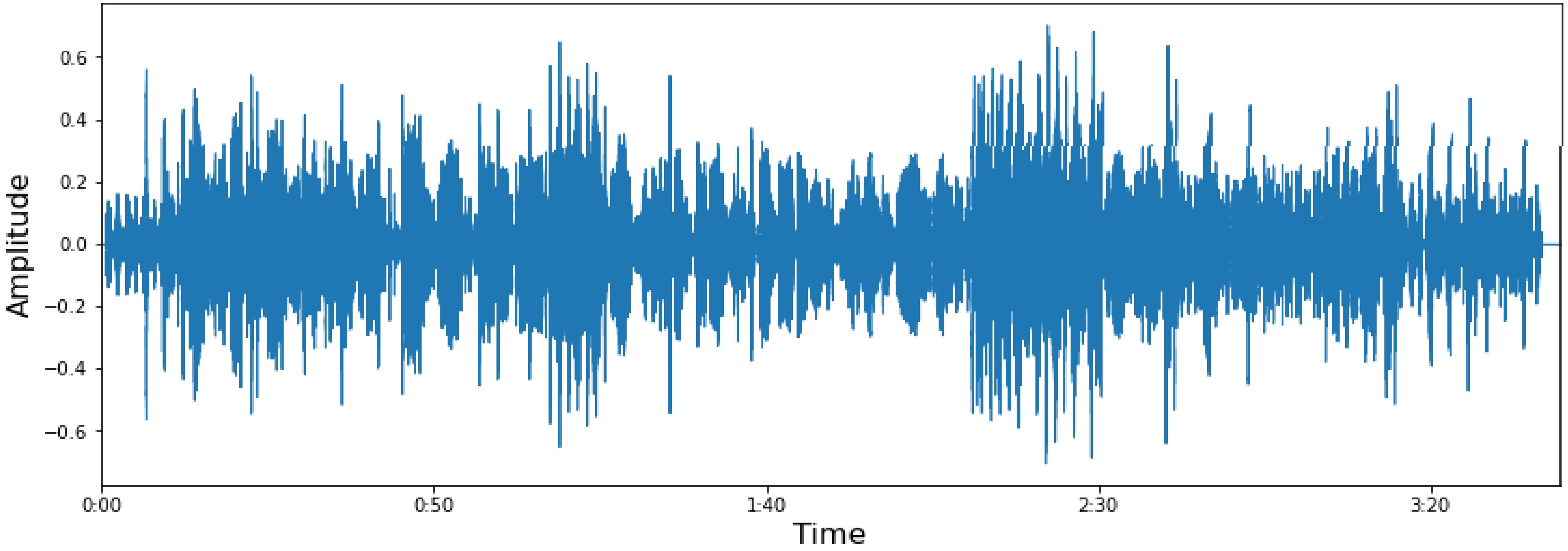


ML Instrumental Waveform of Run Run Run Song

Results (Waveform Comparison)-[3]

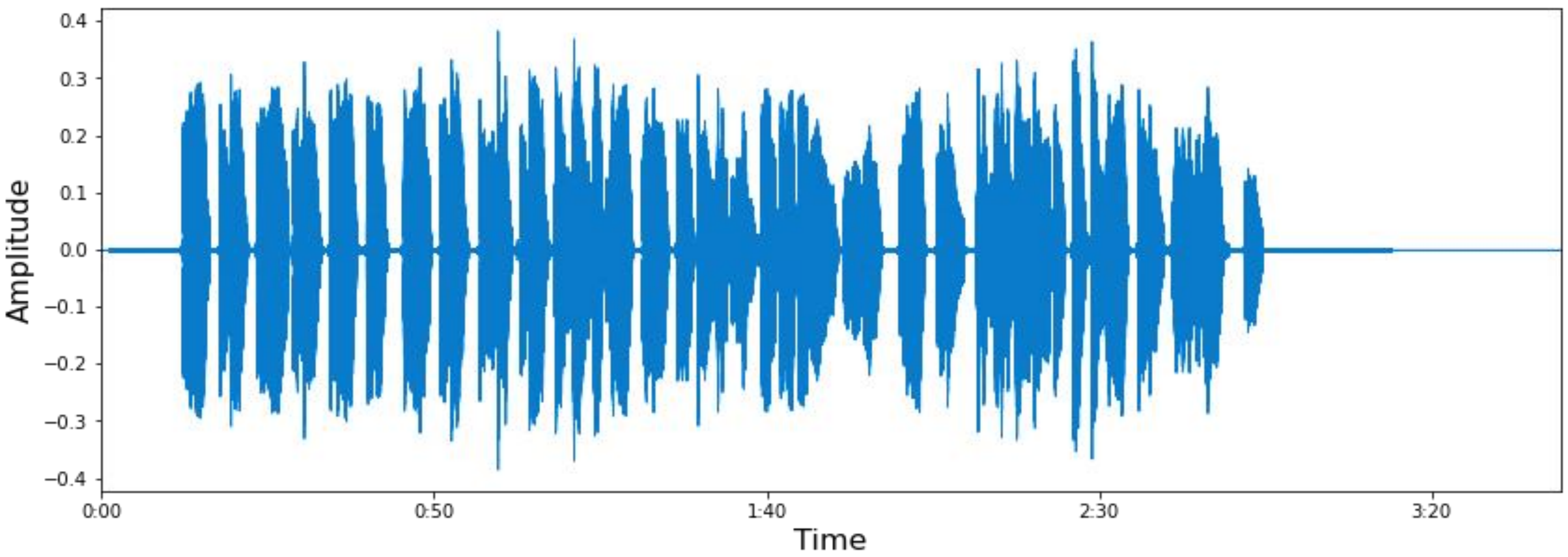


Ground-truth Vocal Waveform of Run Run Run Song

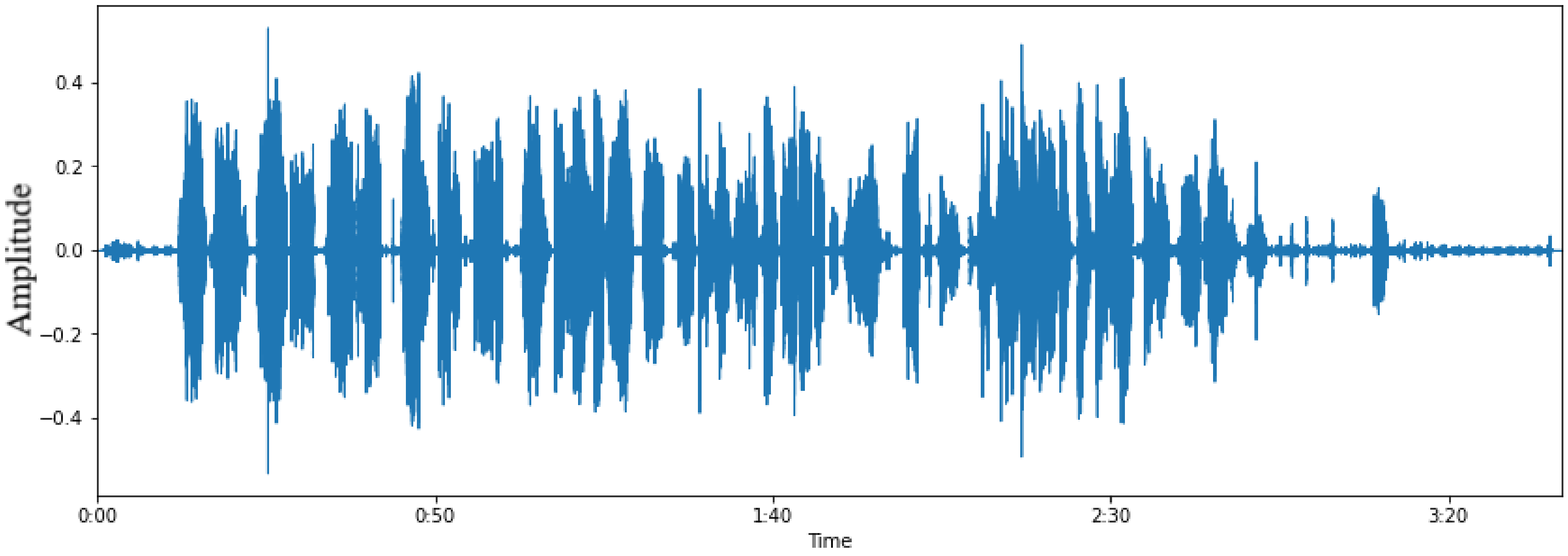


2DFT Vocal Waveform of Run Run Run Song

Results (Waveform Comparison)-[4]

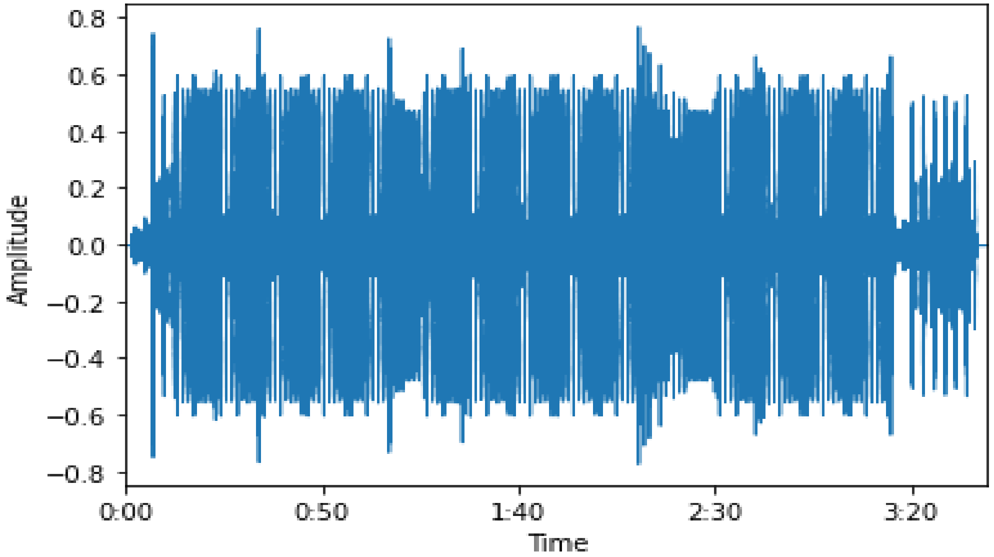


Ground-truth Vocal Waveform of Run Run Run Song

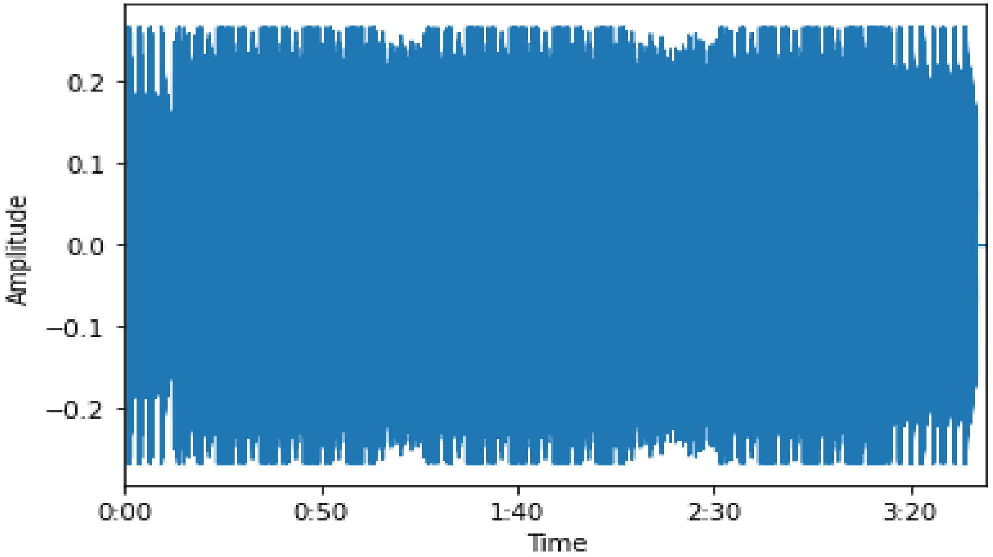


ML Vocal Waveform of Run Run Run Song

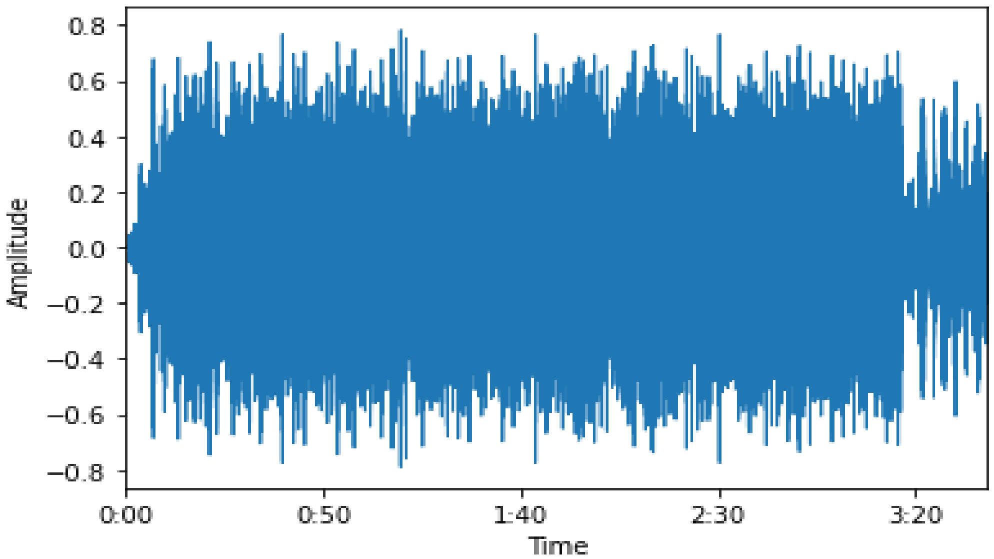
Results (Waveform Comparison)-[5]



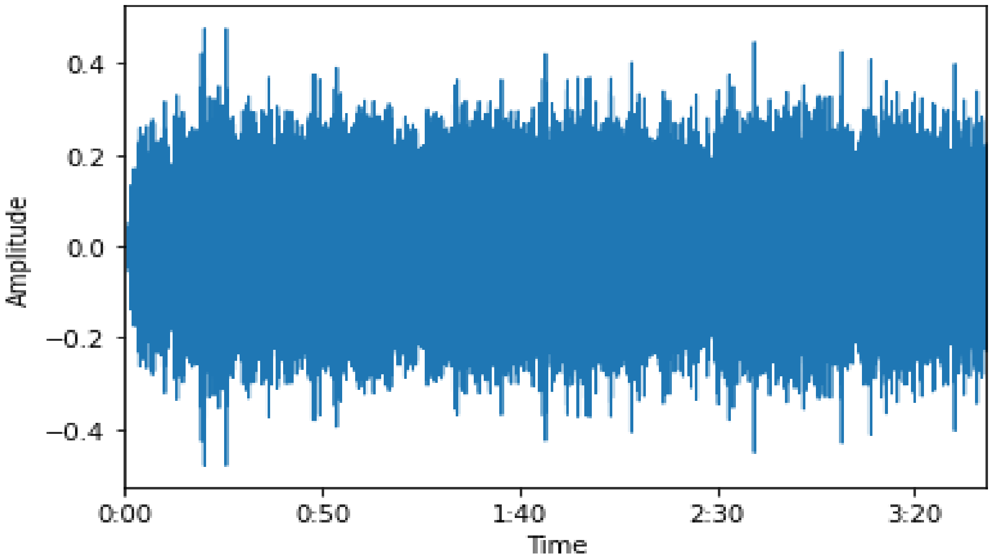
Ground-truth Drum of Run Run Run Song



Ground-truth Bass of Run Run Run Song



ML Drum of Run Run Run Song



ML Bass of Run Run Run Song

Results
(Evaluation Metrics of Vocal)

Genre	2DFT Approach			Machine learning Approach		
	SNR (dB)	SAR (dB)	SDR (dB)	SNR (dB)	SAR (dB)	SDR (dB)
Pop/Rock	2.1018	-0.6405	0.8527	6.3181	7.7059	6.3181
Country	0.9763	-0.9569	0.2435	5.1654	4.8321	5.1654
Rap	-2.4698	-4.7934	-3.211	3.3399	2.4673	3.3399
Jazz	-0.1313	-2.2790	-0.6493	4.1843	3.0958	4.1843
Reggae	2.4286	-0.8136	1.1246	5.3954	4.8879	5.3954
Electronic	-9.3916	-11.6256	-10.0340	4.4968	3.5516	4.4968
Heavy Metal	-0.7201	-4.0395	-1.4521	7.6181	7.8850	7.6181

Results
(Evaluation Metrics of Instrumental)

Genre	2DFT Approach			Machine learning Approach		
	SNR (dB)	SAR (dB)	SDR (dB)	SNR (dB)	SAR (dB)	SDR (dB)
Pop/Rock	6.8457	5.2404	6.1578	9.1852	9.1473	9.1852
Country	5.5380	3.5657	4.5305	8.6397	9.3898	8.6397
Rap	4.6192	2.1449	3.8704	9.7534	9.9783	9.7534
Jazz	4.4232	1.4462	3.4528	8.6592	8.5258	8.6592
Reggae	10.3305	9.1015	9.4276	8.6700	9.7357	8.6700
Electronic	3.6487	0.8474	3.0422	9.2185	9.2330	9.2185
Heavy Metal	7.4202	6.1805	6.6965	7.3812	7.1152	7.3812

Results
(Cosine Similarity of Instrumental and Vocal)

Genre	2DFT Approach		Machine learning Approach	
	Instrumental	Vocal	Instrumental	Vocal
Pop/Rock	0.8728	0.6786	0.9420	0.8843
Country	0.8062	0.6629	0.9299	0.8572
Rap	0.7743	0.4968	0.9459	0.7581
Jazz	0.7416	0.6023	0.9358	0.7885
Reggae	0.9440	0.6544	0.9303	0.8651
Electronic	0.7237	0.2512	0.9432	0.8032
Heavy Metal	0.8900	0.5293	0.9046	0.9096

Results
(Evaluation Metrics of Drum and Bass)

Genre	SNR (dB)		SAR (dB)		SDR (dB)	
	Drum	Bass	Drum	Bass	Drum	Bass
Pop/Rock	4.2249	-3.2061	2.7190	-5.8124	4.2249	-3.2061
Country	3.0561	2.0732	2.2697	0.0581	3.0561	2.0731
Rap	2.3053	3.9658	0.8206	2.7393	2.3053	3.9658
Reggae	-2.5902	1.2126	-3.5596	-1.6525	-2.5902	1.2126
Electronic	2.1252	4.0147	1.7054	2.3588	2.1252	4.0147
Heavy Metal	0.4338	2.7370	-5.5563	1.8676	0.4338	2.7370
Jazz	6.0837	3.5682	6.8360	2.3651	6.0837	3.5682

Results
(Cosine Similarity of Drum and Bass)

Genre	Drum	Bass
Pop/Rock	0.7929	0.4160
Country	0.7801	0.6976
Rap	0.7198	0.7878
Reggae	0.5439	0.6268
Electronic	0.7150	0.7914
Heavy Metal	0.4495	0.7338
Jazz	0.8926	0.7667

Analysis and Discussion
(Comparison with Other Related Projects)-[1]

Model	SDR of Vocal (dB)	SDR of Instrumental (dB)	Overall SDR (dB)
This Project	5.28	11.43	8.355
Dedicated U-Nets (x2)	5.09	12.95	9.02
C-U-Net	4.42	12.21	8.31
UW	5.06	12.98	9.02
DWA	5.20	12.96	9.08

Analysis and Discussion
(Comparison with Other Related Projects)-[2]

Model	SDR of Drum (dB)	SDR of Bass (dB)	Overall SDR (dB)
This Project	4.03	3.21	3.56
Wave-U-Net	4.22	3.21	3.56
Open-Unmix	5.73	5.23	5.43
Meta-Tasnet	5.91	5.58	5.96
D3Net	7.01	5.25	6.50

Future Enhancement

- For more stem separation, a dataset with more instruments can be used

- Constant Q-Transform can be used for domain transformation

- Other machine learning model can be explored

Conclusion

- From the 2DFT approach, only vocal and instrumental were separated

- A U-Net model was trained by using the spectrograms of the songs

- U-net model was able to separate vocal, instrumental, drum, and bass stems with better performance

- All of the project objectives were successfully completed

References – [1]

- A. J. Simpson, G. Roma and M. D. Plumbley, "Deep Karaoke: Extracting Vocals from Musical Mixtures Using a Convolutional Deep Neural Network," in *12th International Conference on Latent Variable Analysis and Signal Separation*, Liberec, Czech Republic, 2015.

- "The DSD100 Dataset," SiSEC MUS, [Online]. Available: https://www.sisec17.audiolabs-erlangen.de/#/dataset. [Accessed 06 06 2021].

- P. Seetharaman, F. Pishdadian and B. Pardo, "Music/Voice separation using the 2D fourier transform," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2017.

References – [2]

- K. M. M. Prabhu, "Review of Window Functions," in *Window Functions and Their Applications in Signal Processing*, New York, USA, CRC Press, Taylor & Francis Group, 2014, pp. 87-126.

- A. Koretzky, "Towards Data Science," Audio AI: isolating vocals from stereo music using Convolutional Neural Networks , 04 02 2019. [Online]. Available: https://towardsdatascience.com/audio-ai-isolating-vocals-from-stereo-music-using-convolutional-neural-networks-210532383785. [Accessed 06 06 2021].

- Z. Rafii, A. Liutkus, F.-R. Stoter, S. I. Mimilakis, D. FitzGerald and B. Pardo, "An Overview of Lead and Accompaniment Separation in Music," *IEEE/ACM Transactions on Audio, Speech and Language Processing,* vol. 26, no. 8, pp. 1307-1335, 2018.