Music Segmentation Using Deep Learning

The Dark Knights

Ayan Banerjee & Soham Bodhak

RKMVERI, MSc BDA

May 3, 2025

Contents

Introduction

- **Music segmentation** is the task of dividing a music track into meaningful structural sections (e.g., verse, chorus, bridge).
- These segments are essential for tasks such as music analysis, retrieval,and remixing.
- Traditional methods depend on hand-crafted features and rules, which often fail with complex or varied music.
- Deep learning enables automatic feature learning directly from raw or low-level inputs.
- In this project, we implement a U-Net model inspired by the paper *Splitter: Learning to Segment Music with Noisy Labels*, using spectrograms as input and predicting segment boundaries.

Contents

Problems

- **Objective:** Given an audio file, detect the boundaries between different musical segments (e.g., intro, verse, chorus).
- **Challenges:**
  - Segment boundaries are often subjective and imprecise.
  - Training data may contain noisy or weak labels.
  - Audio is high-dimensional and varies over time.
- **Goal:** Train a U-Net based deep learning model to predict a boundary probability map from spectrogram features, robust to label noise.

Contents

Dataset

- We use the MUSDB18 dataset, a widely used benchmark for music source separation and structure-related tasks.
- It consists of 150 full-length stereo music tracks (approximately 10 hours total) of which **100** are used for training and **50** songs for test set.
- Each song is a multi-stream .stem.mp4 file containing 5 stems (mixture, drums, bass, other, vocals).
- Log-magnitude spectrograms are extracted from audio segments as model input.

Contents

Literature Review

- Romain Hennequin, Anis Khlif, Felix Voituret, and Manuel Moussallam.A fast and efficient music source separation tool with pre-trained models.Spleeter: https://github.com/deezer/spleeter, 2019.
- Minseok Kim, Woosung Choi, Jaehwa Chung, Daewon Lee, and Soonyoung Jung.Kuielab-mdx-net: A two-stream neural network for music demixing. Technical report,Korea University, 2021
- Zafar Rafi, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, Derry FitzGerald, and Bryan Pardo. The musdb18 corpus for music separation. https://zenodo.org/records/1117372, 2017.
- Daniel Stoller, Sebastian Ewert, and Simon Dixon. Wave-u-net: A multi-scale neural network for end-to-end audio source separation. In Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR), pages 334-340. ISMIR, 2018.

Contents

Preprocessing

- Stems are extracted using FFmpeg into mono WAV files at 22.05 kHz.
- Audio clips are fixed to 4 seconds (DURATION $= 4$).
- Each track is separated into 5 stems, of which 4 (drums, bass, other, vocals) are reconstructed.

Contents

Enhanced U-Net

- The model follows a U-Net structure with encoder-decoder design with 8,967,941 Trainable Parameters.
- **ResidualBlock:** Each block contains: Two Conv2d layers (3x3, padding=1) with InstanceNorm and LeakyReLU.
- Each encoder block(total 3) is a residual block + MaxPooling(2)
- Each decoder block uses Upsample followed by Conv2d(3x3) and Attention Mechanism is applied at the third decoder level
- Output: 4-channel spectrogram mask reconstructed using Sigmoid activation.

Contents

Loss Function and Evaluation Metrics

- **Loss:** Computed using CombinedLoss(Weighted combination of MSE(0.7) and Spectral Convergence loss(0.3)) on the test set.
- **Metrics:**Signal-to-Distortion Ratio (SDR): Calculated for reconstructed stems.

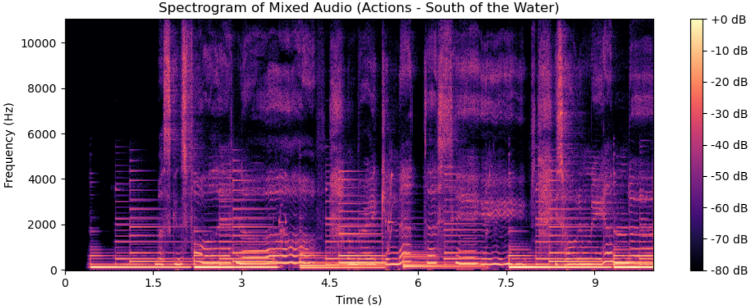
Contents

Training and Evaluation

- Optimizer: AdamW (lr=0.001, weight-decay=1e-4).
- Spectrograms generated from mono audio with 4-second chunks.
- Mixed input is mapped to clean stem outputs.
- Audio is processed in 4-second chunks with 25% overlap to minimize boundary artifacts.

Contents

Plotting



Spectrogram of Mixed Audio (Actions - South of the Water)
Frequency (Hz)
10000
8000
6000
4000
2000
0
0
1.5
3
4.5
6
7.5
9
Time (s)
+0 dB
-10 dB
-20 dB
-30 dB
-40 dB
-50 dB
-60 dB
-70 dB
-80 dB

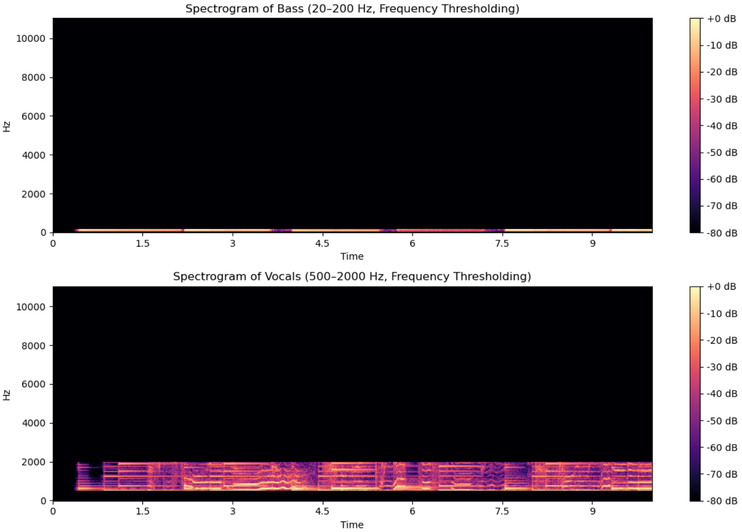
Evaluation Results for Stage 1: Frequency Thresholding

The SDR (Signal-to-Distortion Ratio) values obtained using the simple frequency thresholding method are shown below:

Source	**SDR (dB)**
Bass	Nan
Vocals	-16.45
Drums	-6.61

Table: SDR values for different sources using frequency thresholding. The SDR for bass is undefined due to a silent or invalid estimation.

Plotting


Spectrogram of Bass (20–200 Hz, Frequency Thresholding)
Hz
10000
8000
6000
4000
2000
0
0
1.5
3
4.5
6
7.5
9
Time
+0 dB
-10 dB
-20 dB
-30 dB
-40 dB
-50 dB
-60 dB
-70 dB
-80 dB
Spectrogram of Vocals (500–2000 Hz, Frequency Thresholding)
Hz
10000
8000
6000
4000
2000
0
0
1.5
3
4.5
6
7.5
9
Time
+0 dB
-10 dB
-20 dB
-30 dB
-40 dB
-50 dB
-60 dB
-70 dB
-80 dB

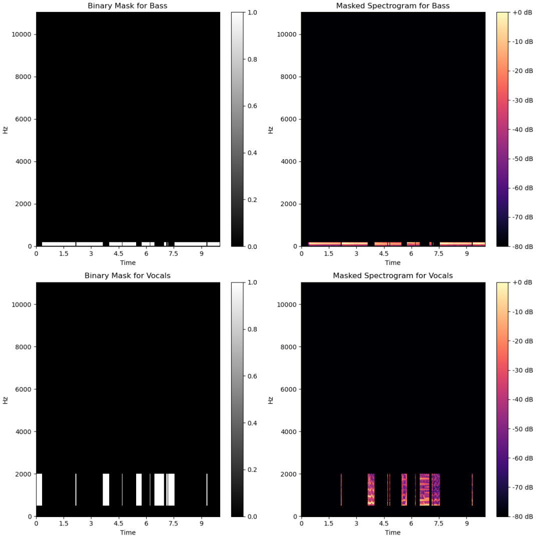
Evaluation Results for Stage 2: Spectrogram Masking

The SDR (Signal-to-Distortion Ratio) values for the separated sources using the spectrogram masking method are summarized below:

Source	**SDR (dB)**
Bass	-41.80
Vocals	-13.88
Drums	-7.03

Table: SDR values for different sources using spectrogram masking.

Plotting


Binary Mask for Bass
Masked Spectrogram for Bass
Binary Mask for Vocals
Masked Spectrogram for Vocals
Hz
Time
10000
8000
6000
4000
2000
0
0 1.5 3 4.5 6 7.5 9
1.0
0.8
0.6
0.4
0.2
0.0
+0 dB
-10 dB
-20 dB
-30 dB
-40 dB
-50 dB
-60 dB
-70 dB
-80 dB

Evaluation Results for Stage 3: Classical ML and DL methods

The SDR (Signal-to-Distortion Ratio) values for the separated sources using the ML method are summarized below:

Source	**SDR (dB)**
Bass	-7.00
Vocals	-7.1
Drums	-4.16

Table: SDR values for different sources using ML.

Evaluation Results fopr Enhanced U-Net

After training, the model was evaluated on the MUSDB18 test set. The evaluation loss and Signal-to-Distortion Ratio (SDR) were computed as follows:

Metric	Value
Evaluation Loss	20.295310
Average SDR	0.338778

Table: Evaluation performance of the Enhanced U-Net on the test set.

The model was used to perform inference on a new audio track. The following separated stem files were generated:

- `inferred_outputs/drums_reconstructed.wav`
- `inferred_outputs/bass_reconstructed.wav`
- `inferred_outputs/other_reconstructed.wav`
- `inferred_outputs/vocals_reconstructed.wav`

Evaluation Results for Spleeter

The SDR (Signal-to-Distortion Ratio) values for the separated sources using the ML method are summarized below:

Source	**SDR (dB)**
Bass	5.51
Vocals	6.86
Drums	6.71

Table: SDR values for different sources.

Contents

Results

- Enhanced U-Net achieved stable reconstruction across all 4 stems, taking 15 minutes for a single epoch.
- Attention mechanisms improved SDR on dense overlapping sources.
- Segment transitions align with peaks in boundary mask outputs.
- Output audio reconstructions saved and compared to ground truth.

Contents

Future Work

- If more proper resources are available:
- We can Use a more deeper architecture and Transformer based model.
- We can use a more high quality version of musdb18 dataset.

Contents

Conclusion

- Deep learning models like U-Net can effectively learn music structure from weak labels.
- Enhanced U-Net with residual blocks and attention mechanisms improves robustness.
- MUSDB18 provides a rich dataset to train and evaluate such models.
- Future Work: Train on larger, annotated datasets and explore explicit segmentation prediction.