Comparing Different Spectral Losses of a U-Net Based Audio Source Separation System

Yilong Tang

Music Informatics Group, Center for Music Technology, Georgia Institute of Technology

Georgia Tech · College of Design
Center for Music Technology

Abstract

Audio source separation is the process of extracting individual sound sources from a mixed audio signal, which can be useful in various music-related tasks such as lyrics transcription, chord transcription, and drum transcription. The separated instrumental tracks can be used for karaoke software, while audio denoising and speech enhancement tools use source separation to remove unwanted noise and enhance speech quality.

The goal for this semester project is re-implement the Spleeter source separation system and experiment with different loss functions to compare their performance.

Method

- **Dataset**: MUSDB18, The dataset contains files encoded with Native Instruments stem format and the multi-track is represented in 5 sources. Each song was randomly segmented into six-second clips, and PyTorch data loader function was used to load data in batches.
- **Network Architecture**: U-Net is a fully convolutional neural network with three main components: the input encoder, the separation module, and the output decoder. Each encoder layer's size is halved, and the channel size is doubled, and after separating the encoded spectrogram, it goes through the decoder layers, which upsample the spectrogram to its original size. The skip connections connect the encoder path to the decoder path at the same dimension, which allows the network to preserve the details of the input spectrogram that might be lost during downsampling and upsampling.

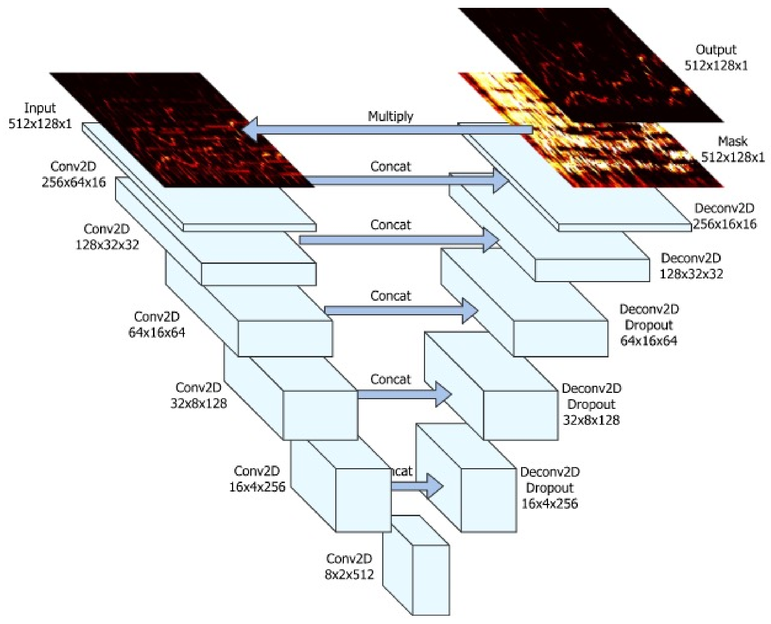

Figure 1. Network Architecture

- **Network Input and Output**: The input spectrogram to the network is calculate using STFT with the window size of 4096 and hop length of 1024. The output of the final layer is a soft mask or a ratio mask.

$$X_i = (\hat{M}_i \odot |Y|) \odot e^{j \cdot \angle Y} \quad (1)$$

- **Loss Functions**:
  Spectrogram MSE loss:

$$L_{spec} = \frac{1}{N} \sum_{n=1}^{N} \left| Y_n - \hat{Y}_n \right|^2 \quad (2)$$

  Magnitude Spectrogram MSE loss:

$$L_{mag} = \frac{1}{N} \sum_{n=1}^{N} \left| |Y_n| - \left| \hat{Y}_n \right| \right|^2 \quad (3)$$

  Mel Spectrogram MSE loss:

$$L_{mel} = \frac{1}{N} \sum_{n=1}^{N} \left| M_n - \hat{M}_n \right|^2 \quad (4)$$

  Spectrogram MAE loss:

$$L_{spec_{l1}} = \frac{1}{N} \sum_{n=1}^{N} \left| Y_n - \hat{Y}_n \right| \quad (5)$$

Evaluation

Many advanced source separation systems use objective evaluation metrics because organizing a listening test is difficult and evaluating audio quality with many recordings takes time. The BASS evaluation metric includes SDR, SIR, and SAR, and the "museval" python package is used to calculate SDR in this experiment. The evaluation was conducted on a testing file provided by the MUSDB18 dataset, which contains 50 songs truncated to 2 minutes each to save time.

Result

Among the four loss functions tested, the mean square error (MSE) spectrogram loss performed the best, with an SDR value of 4.9744. The L1-norm spectrogram loss used by Spleeter performed worse, with an SDR of 4.6194. The mel spectrogram loss and additional log scale magnitude spectrogram loss functions were also tested, but they performed poorly with SDR values of 0.766 and 1.1534, respectively.

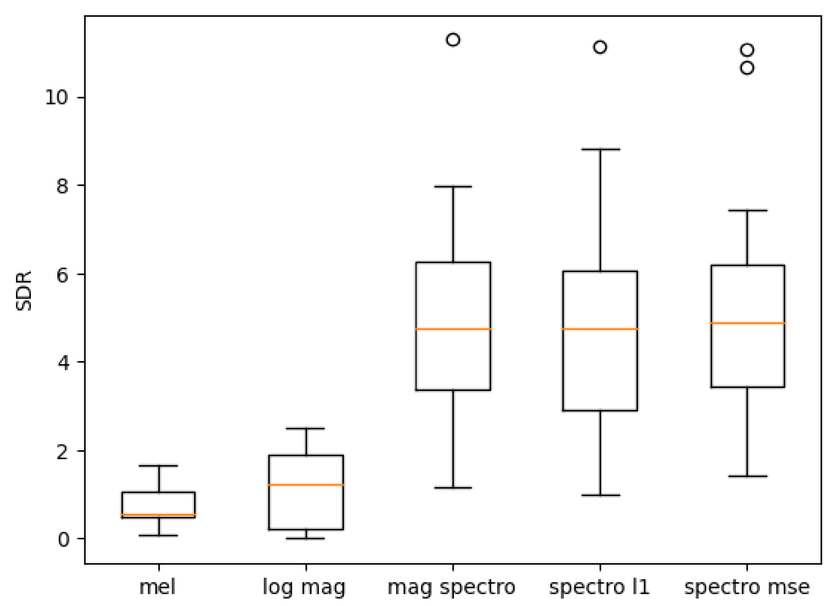

Figure 2. Systems Train with Different Loss For Vocals Source Separation

The best-performing model was tested on four stems, and its SDR performance was worse than the Spleeter model's performance on all instruments.

Table 1. SDR Comparison With Spleeter

	Spleeter Trained on MUSDB18	MSE Spectrogram Loss Model
Vocals	5.10	4.5213
Drums	5.15	5.0379
Bass	4.27	4.0649
Other	3.21	3.0807

Future research can explore other loss functions, such as the perceptual loss function, to retain more meaningful audio features by mimicking human hearing perception.

References

[1] R. Hennequin, A. Khlif, F. Voituret, and M. Moussallam, "SPLEETER: A FAST AND STATE-OF-THE ART MUSIC SOURCE SEPARATION TOOL WITH PRE-TRAINED MODELS," en, 2019.

[2] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, "SINGING VOICE SEPARATION WITH DEEP U-NET CONVOLUTIONAL NETWORKS," en, 2017.

[3] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, *MUSDB18 - a corpus for music separation*, Dec. 2017. DOI: 10.5281/zenodo.1117372. [Online]. Available: https://doi.org/10.5281/zenodo.1117372 (visited on 04/26/2023).

[4] I. Ananthabhotla, S. Ewert, and J. A. Paradiso, "Using a Neural Network Codec Approximation Loss to Improve Source Separation Performance in Limited Capacity Networks," in *2020 International Joint Conference on Neural Networks (IJCNN)*, ISSN: 2161-4407, Jul. 2020, pp. 1–7. DOI: 10.1109/IJCNN48605.2020.9207053.