Progressive Inference for Music Demixing Report related to task number 1

Giorgio Magalini[1], Alessandro Manattini[2], and Filippo Marri[3]

[1] *la pearà*
[2] *il turtelin*
[3] *la bistecca*

Compiled April 7, 2025

In this work, we present a preliminary study on music demixing using the DEMUCS[1] model. We briefly review the theoretical underpinnings of source separation, outline our experimental setup, and discuss the quantitative and qualitative results obtained. Our findings suggest that DEMUCS can effectively separate drums, bass, vocals, and other musical stems, although certain categories—particularly "other" instruments—remain challenging. Furthermore, we propose a simple strategy to handle silent stems in our evaluation pipeline, ensuring that the system can operate even if certain tracks have no content in the chosen time interval. At the end, we report a brief discussion on the schedule we decided to use for the DEMUCS model.

http://dx.doi.org/10.1364/ao.XX.XXXXXX

1. INTRODUCTION

Music demixing (or source separation) is the task of isolating individual instruments (e.g., drums, bass, vocals) from a single, fully mixed audio track. This has several applications in audio engineering, music production, and research. Traditional approaches include nonnegative matrix factorization (NMF), which factors a magnitude spectrogram into a product of basis and activation matrices. However, recent deep learning techniques have achieved superior performance.

In this work, we employ DEMUCS, a neural network architecture specifically designed for music de-mixing. We evaluate Demucs on a test dataset of songs called MUSDB18-HQ[2] with known ground-truth stems and measure separation quality using Signal-to-Distortion Ratio (SDR). This metric allows us to reach the goal of this first task: to perform a prove of concpet in which it will be demonstrated that the extraction of a single stem gives better result when the track that we want to extract is mixed using an higher value of gain.

Once this is verified, we create an iterative procedure in which the stem is extrcated progressively from the mixed track feeding the NN with the result of the previous extraction. This idea draw inspiration from the diffusion models idea in which the result is reached by progressively removing noise from a white noise input. What is done it is looking for the best schedule parameter that optimises the performance of this iterative algorithm.

2. BACKGROUND

A. Demucs

Demucs is a convolutional neural network that operates (primarily) in the time domain. It combines aspects of waveform modeling with elements of recurrent or convolutional architectures, capturing both local and global context. The key advantage is its ability to generate separated stems that retain temporal details more effectively than purely spectrogram-based methods.

B. Evaluation Metric

The evaluation metric used (Signal to Noise Ratio) is defined as follows:

$$sdr = 10 \log_{10} \frac{||x_{stem}(t)||^2}{||x_{error}(t)||^2} \, [dB] \tag{1}$$

where x_{error} is defined as the difference between the ground-truth extracted track and the norm of the outuput of the neural network. It is implemented using the `bss_eval_sources` method of `mir_eval` library. [3]

C. Diffusion models equation

The equation that is used for the itration is the following.

$$\underline{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\underline{x}_t - \frac{\beta_t}{\sqrt{1-\alpha_t}} \underline{\epsilon}_\theta \left(\underline{x}_t, t \right) \right) + \sigma_t \underline{z} \tag{2}$$

where $\underline{\epsilon}_t \sim \mathcal{N}(0, 1)$

In a diffusion model, the schedule β_t represents the variance of the noise at a given step t. In our application it will represents the gain with which the differents stems will be weighted in the mixture.

3. METHODOLOGY

A. Preliminar operations

The tests were conducted implementing a Python algorithm. The device used to run them has been Metal Performance Shaders (MPS) backend for GPU training acceleration. This MPS backend extends the PyTorch framework, providing scripts and capabilities to set up and run operations on Mac.

Once the dataset was loaded, each stem and song is cropped in 30 seconds long chunks and arranged in a dictionary with its relative label. After that, a preliminar analysis is conducted on a single track to check the main structure of the algorithm. There, after performing normalisation, the mixed track is demixed using the Demucs model. The results are de-normalised by multiplying them by the standard deviation of the mixed track and summing the mean of the same track. Once again, the results are organised in a dictionary. SDR is evaluated as a check.

B. Proof of concept

Now the procedure is iterated on every track selecting by means of a specific function only the chunks in which information is stored. In fact, it could happen that some stems are completely silent, since they have been cropped at a point of the track in which the instrument or group of instruments related to those stems are silent. We want to demonstrate that the Demucs model works better when the stem that we want to extract is mixed with a higher gain with respect to the others. In order to do that, we need a reference that will consists in the mean value of the SDR of every stem of every chunks. This result is achieved implementing the algorithm 1.

Algorithm 1. Main algorithm

```
definition of the weights for the gains
while track in dataset do
    mixing of the stems according to the weights
    computation of the sdr using bss_eval_sources
    evaluation of the mean of the sdr over time
evaluation of the mean of the sdr across all tracks
```

Once the reference is there, we experiment different values of weigths for the mixing of the stems.

C. Schedule choice

Come troviamo e proviamo i diversi valori di β.

4. RESULTS

A. Proof of concept

By setting all the weights equal to 0.25 the SDR mean value we get are the one shown in figure 1.

It can be noticed that the SDR value is almost the same for each stem. However, if the stmes are not uniformly mixed but one of them is mixed with a higher gain with respect to the other one, the SDR of the output related to that significantly increases as it is shown in figure 2.

This demonstrates that the good quality of a stem-extracted signal is proportional to the intensity that the specific stem has in the micture with respect to the other stems.

B. Schedule choice

Qui bisogna mettere tutta la parte in cui spieghiamo quale beta abbiamo scelto e perché

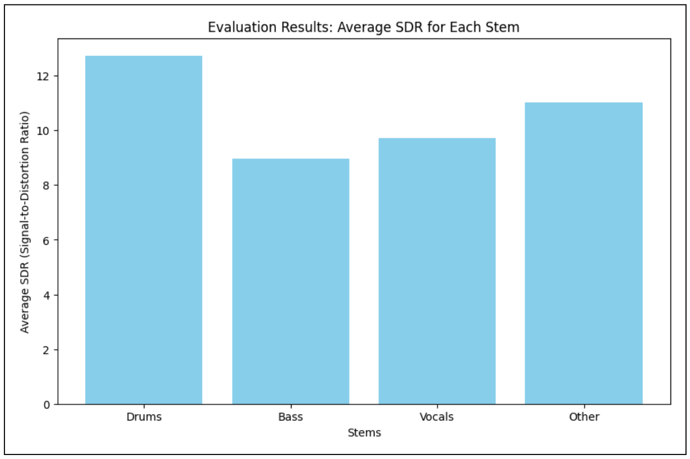

Fig. 1. Average Signal to Noise Distortion for each stem.

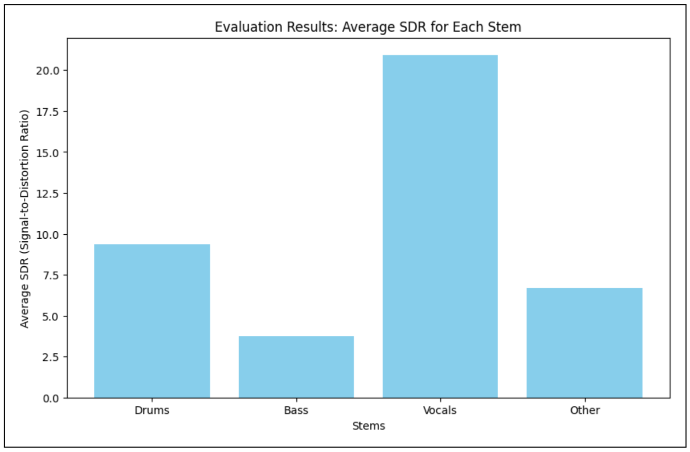

Fig. 2. Average Signal to Noise Distortion for each stem when the voice stem is mixed with an higher value of gain with respect to the others.

REFERENCES

1. A. Défossez, "Hybrid spectrogram and waveform source separation," arXiv preprint arXiv:2111.03600 (2021).
2. Z. Rafii, A. Liutkus, F.-R. Stöter, *et al.*, "Musdb18-hq - an uncompressed version of musdb18 (1.0.0) [data set]," Zenodo (2019).
3. C. Raffel, B. McFee, E. J. Humphrey, *et al.*, "mir_eval: A transparent implementation of common mir metrics," Proc. 15th Int. Conf. on Music. Inf. Retr. (2014).