# Musical Source Separation of Brazilian Percussion

Richa Namballa [1] Giovana Morais [1] Magdalena Fuentes [1, 2]

[1]Music and Audio Research Laboratory, New York University

[2]Integrated Design & Media, New York University

## Objectives

- Improve representation of non-Western instruments in the task of musical source separation (MSS).
- Use an existing dataset of Brazilian *samba* percussion to create artificial mixtures to train a source separation model for the *surdo*, a low-pitched drum with a distinctive timbre.

## Background

- **Musical source separation (MSS)** is a core task in music information retrieval (MIR) that aims to "de-mix" audio into instrument stems.
- Most systems are trained to process Western instruments only.
- Limited inclusivity due to lack of diverse training data.
- Creation of new datasets is time-consuming and expensive.
- Investigate the feasibility of building an MSS system with **artificially-created** mixtures from an **existing** dataset: the **Brazilian Rhythmic Instruments Dataset (BRID)** [1, 2].

## Data

- BRID contains 274 solo tracks across 10 different instruments and 5 rhythmic styles.
- **Target source** is the **surdo**, a large tom-like drum which plays a repeated pattern throughout the piece (Figure 1).

Figure 1. A surdo, commonly used in Brazilian samba performances (Source: Adobe Stock).

- Generated mixtures by randomly combining solo surdo tracks (26) with other solo instruments from the same musical style.
- Number of stems in each mixture varied.
- Ensured each instrument-style combination was represented in each train/validation/test split.
- Solos within a style had the same tempo and could be mixed, without time-alignment.
- No repetition of an instrument type within a single mixture and no duplicate mixtures.

## Methods

- 2D convolutional U-Net encoder-decoder adapted for the frequency domain [3].
- Mimicked single-task implementations (architecture, hyperparameters, processing) of [4] and [5], as seen in Figure 2.
- Trained surdo-separation model for 1000 epochs with a single Nvidia RTX8000 GPU (approximately 30 minutes).
- Separated audio is reconstructed from the output spectrogram using the inverse Short-Time Fourier-Transform (iSTFT).
- Evaluated using the Source-to-Distortion Ratio (SDR) [6] and by listening to the separated audio of the test set.

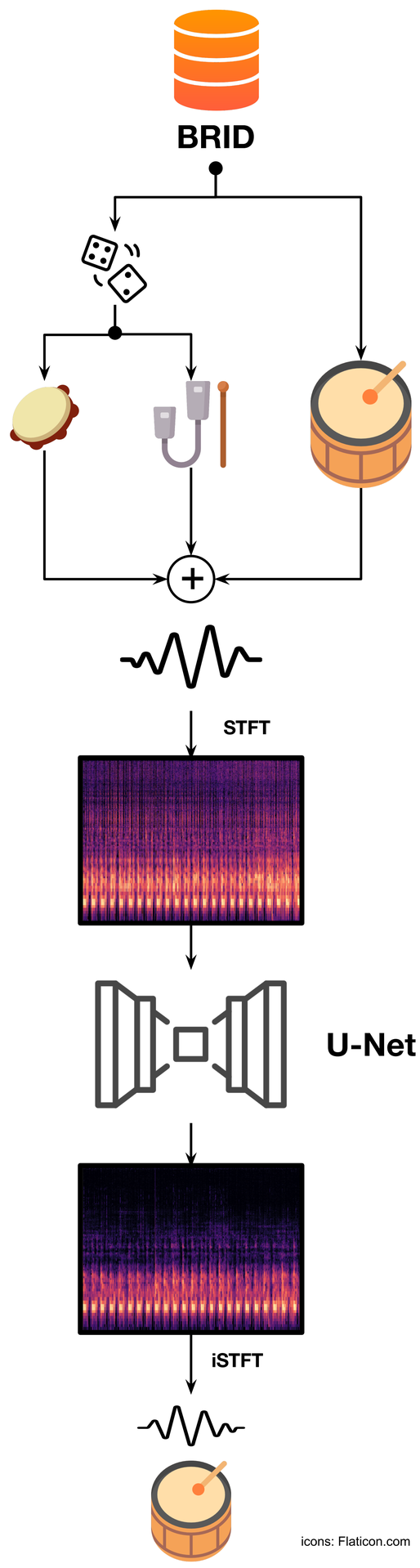


Figure 2. Overview of the training pipeline for the surdo-separation model. Solo tracks are randomly selected from BRID and combined with a surdo stem to generate mixtures for training.

## Results

| Dataset | Size | Mean ± SD | Median |
|---|---|---|---|
| Training | 100 | 16.83 ± 7.13 | 16.97 |
| Validation | 10 | 13.27 ± 9.61 | 12.92 |
| Testing | 30 | 17.57 ± 8.80 | 16.00 |

Table 1. SDR performance of the surdo separation model.

- Impressive SDR metrics (Table 1) demonstrating clear separation.
- Qualitatively, **minimal distortion**; only bleed from other instruments in the mixture.
- Checked for over-fitting by applying the model to the BRID group performances and a YouTube video of a percussion ensemble[a].
- Attribute clean separation to homogeneity of BRID, the repetitive nature of the surdo rhythmic pattern, and the unique timbre and frequency range of the instrument.
- Shows that a simple MSS model can perform a decent separation of an instrument **without large amounts of data, as long as the style features a certain amount of homogeneity**.
- Future Work
  - Resample data splits to check for possible biases from small data size.
  - Explore the performance of this pipeline on other percussion instruments from Brazil (BRID) and beyond.

## Demo

## References

[1] P. D. Tomaz Jr., W. S. d. Silva Jr., and L. W. P. Biscainho, "Separação automática de instrumentos de percussão brasileira a partir de mistura pré-gravada," Master's thesis, Universidade Federal do Rio de Janeiro, June 2016.

[2] L. S. Maia, P. D. Tomaz Jr., M. Fuentes, M. Rocamora, L. W. P. Biscainho, M. V. M. da Costa, and S. Cohen, "A novel dataset of Brazilian rhythmic instruments and some experiments in computational rhythm analysis," in *AES Latin American Conference 2018*, (Montevideo, Uruguay), Audio Engineering Society, 2018.

[3] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, 18th International Conference, Proceedings, Part III*, (Munich, Germany), pp. 234–241, Springer, 2015.

[4] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep U-net convolutional networks," in *Proceedings of the 18th International Society for Music Information Retrieval Conference*, (Suzhou, China), pp. 745–751, 2017.

[5] G. Meseguer-Brocal and G. Peeters, "Conditioned-U-net: Introducing a control mechanism in the U-net for multiple source separations," in *Proceedings of the 20th International Society for Music Information Retrieval Conference*, (Delft, Netherlands), pp. 159–165, 2019.

[6] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[a] https://youtu.be/mmlK94QvwiA?si=mGwgwiioUWHmsyFD