# 1st Latin American Music Information Retrieval Workshop

December 9 to 11, 2024

Rio de Janeiro, Brazil

# Proceedings

# MUSIC SOURCE SEPARATION IN NOISY BRAZILIAN CHORO RECORDINGS

**Pedro Donadio**[1,2] **Martín Rocamora**[3,4] **Luiz Wagner Biscainho**[2]
[1] Department of Electronics and Computation, Universidade Federal do Amazonas, Brazil
[2] Universidade Federal do Rio de Janeiro, Brazil
[3] Music Technology Group, Universitat Pompeu Fabra, Spain
[4] Facultad de Ingeniería, Universidad de la República, Uruguay
pedro.donadio@smt.ufrj.br, rocamora@fing.edu.uy, wagner@smt.ufrj.br

## ABSTRACT

Choro music is considered the first musical style to originate in Brazil, dating back to the 1870s. Some historical recordings from the early 20th century include noise inherent to the process of recording and playing shellac records. In this work, we investigate the instrument separation task applied to historical recordings of this Brazilian music genre, using models originally trained on clean tracks. We used a choro dataset composed of modern recordings of songs from the most important composers of this style, and a 78 RPM (rotations per minute) noise dataset to emulate old choro records. Using an available neural network architecture — Hybrid Demucs — trained to separate the characteristic choro musical instruments into the string, wind, and percussion families without background noise, we evaluate the separation result in the presence of different types of 78 RPM noise. Furthermore, we study the impact of the additive noise on separation when the signal-to-noise ratio (SNR) ranges from 10 to 40 dB. The experiments showcase that the model is robust, although the performance depends on the type and level of noise.

## 1. INTRODUCTION

The task of music source separation consists in isolating the sound of each instrument, or a family of instruments, from an audio mixture (i.e., a track containing various instruments playing together) [1–7]. In recent years, deep learning approaches have achieved state-of-the-art performance in music source separation. Various neural network architectures have been explored for this task, as demonstrated in previous works [8–12]. We chose to base the investigation of this work on the Hybrid Demucs model [10], which uses an encoder-decoder architecture, due to its excellent performance in instrument separation and the availability of a model we trained on a dataset of choro music within our current research.

In many contexts, isolating the singer's voice or even erasing the sound of a specific instrument from a musical excerpt (preserving the rest of the mixture) are practical tasks that may help professional and amateur musicians, music students, audio engineers, musicologists, and researchers. However, most of the works in this area explore separation on modern recordings, obtained in controlled studio environments, such as those found in MUSDB [13] database. The open question of our interest is the effectiveness of these models on historical recordings, i.e., data extracted from discs where noise is an inherent part of the recording/reproduction process.

In this work, we explore the traditional musical genre called *choro*, which, along with samba, is one of the most representative elements of Brazilian and Latin American culture. Historical recordings of choro [14, 15] date back to the beginning of the 20th century and typically involve groups where a rhythmic and harmonic base of 6-string guitars, 7-string guitars, *cavaquinho*, and *pandeiro* accompanies soloists playing flute, mandolin, clarinet, or *cavaquinho*. This music is instrumental in its conception (although some choros have received lyrics later) and is formed through the fusion of various rhythms performed in a characteristic choro way, such as polka, baião (a rhythm from northeast of Brazil), waltz, and maxixe (also known as the Brazilian Tango). It results from a mixture of the African rhythm lundu with European genres. All these aspects introduce different types of challenges. Regarding rhythm, there are various notions of meter, which may vary between 2/4, 3/4, or 4/4. In melodic terms, the improvisational nature of choro allows the soloist freedom to create melodies in certain musical passages. From a technical perspective, the timbral overlap of various instruments increases the difficulty of separation. The main composers of the genre are Ernesto Nazareth (1863-1934), Heitor Villa-Lobos (1887-1959), Pixinguinha (1897-1973), Jacob do Bandolim (1918-1969) and Waldir Azevedo (1923-1980). It is interesting to note that, in most cases, the recordings have the composers themselves performing as soloists. For example, Jacob do Bandolim, who was an excellent mandolinist; Waldir Azevedo, credited with introducing the *cavaquinho* as a soloist instrument in choro groups; and Pixinguinha, who played both the flute and saxophone, and is also considered one of greatest musicians in Brazil.

**Attribution:** P. Donadio, M. Rocamora, and L.W. Biscainho, "Music source separation in noisy Brazilian choro recordings", in *Proc. of the 1st Latin American Music Information Retrieval Workshop,* Rio de Janeiro, Brazil, 2024.

This study investigates the impact of background noise present in the original mix on the performance of a musical instrument family separation system. We utilize a Hybrid Demucs model specifically trained on a choro music dataset composed of 10 albums of modern recordings from the most prominent composers of the genre. In its original form, the model is capable of isolating instruments into three families: strings, wind, and percussion. We conduct a series of experiments by adding different levels and types of 78 RPM noise to the mixtures, simulating the characteristics of old shellac recordings. The noisy mixtures are then separated by the model and compared to the corresponding results obtained from clean (non-noisy) mixtures. In both cases, the separation performance is evaluated according to the most usual objective metrics adopted for source separation [16] — signal-to-distortion ratio (SDR), signal-to-interference ratio (SIR), and signal-to-artifact ratio (SAR) — supplemented by informal but careful subjective listening evaluation. The results demonstrate that the model is robust and can effectively separate instrument families under most conditions, although the performance depends on the type and level of noise.

## 2. METHODS

### 2.1 The Model

Hybrid Demucs[1] (v3) is a neural network that utilizes convolutional layers as encoder/decoder schema. Its main highlight is the fact that it receives the input signal simultaneously in the time and frequency domains, subjecting it to 6 encoder layers and 6 decoder layers. The model was trained for 360 epochs with a batch size equal to 4, using Adam optimizer and a learning rate $\mathrm{Lr} = 10^{-4}$.

The available model has been trained (from scratch) on a choro music dataset that consists of 10 albums for training and 2 albums for validation[2]. Table 1 presents the number of songs per album utilized for training, validation and testing the model.

The data structure for each song includes a mixture (with all instruments playing together) and 4 sources: the string family (6-string guitar, 7-string guitar, *cavaquinho*, mandolin, etc.), the wind family (flute, saxophone, clarinet, etc.), the percussion family (*pandeiro*, *reco-reco*, *tamborim*, snare drums, etc.), and the "others" family (consisting of instruments that do not belong to the other three classes). All tracks have a sampling rate of 48 kHz.

In a future publication, we will provide a detailed presentation of the choro dataset, outlining all stages of track production for each instrument family and their specific characteristics, as well as details of the training process.

### 2.2 The noise dataset

The noise dataset described in [17] comprises various noise segments extracted from recordings of 78 RPM shellac records. These include electrical circuit noise, ambient noise, noise caused by the turntable, and clicks. From this set, we choose 5 noise samples with different characteristics to simulate historical choro recordings; Figure 1 shows their respective spectrograms. To make the tracks from the noise dataset compatible with the clean mixtures from the choro dataset, the noise tracks have been pre-processed: we took only one of the two original stereo channels, resampled it from 44.1 to 48 kHz (to match the sampling rate of the choro dataset), periodically looped the noise in cycles consistent with a 78 RPM recording, and ensured that both the noise track and the musical track were the same length. The original noise track titles (with references to various metadata) in the original dataset are extremely long and were shortened to *cristree*, *vucchella*, *majourney*, *springfield*, and *alacarte* for reference. The samples used in this study are available upon request.

**Table 1**: List of songbooks in *Choro* dataset.

| Number | Songbook | # Songs |
|---|---|---|
| 1 | Altamiro Carrilho | 13 |
| 2 | Benedicto Lacerda | 12 |
| 3 | Chiquinha Gonzaga | 12 |
| 4 | Choro Meets Bach | 14 |
| 5 | Ernesto Nazaré 1 | 11 |
| 6 | Ernesto Nazaré 2 | 11 |
| 7 | Ernesto Nazaré 3 | 11 |
| 8 | Inéditos | 14 |
| 9 | Jacob do Bandolim 1 | 12 |
| 10 | Jacob do Bandolim 2 | 12 |
| 11 | Pixinguinha | 12 |
| 12 | Roda de Choro | 12 |
| 13 | Severino Araújo | 12 |
| 14 | Waldir Azevedo 1 | 12 |
| 15 | Waldir Azevedo 2 | 12 |
| 16 | Zequinha da Abreu | 12 |

## 3. EXPERIMENTS

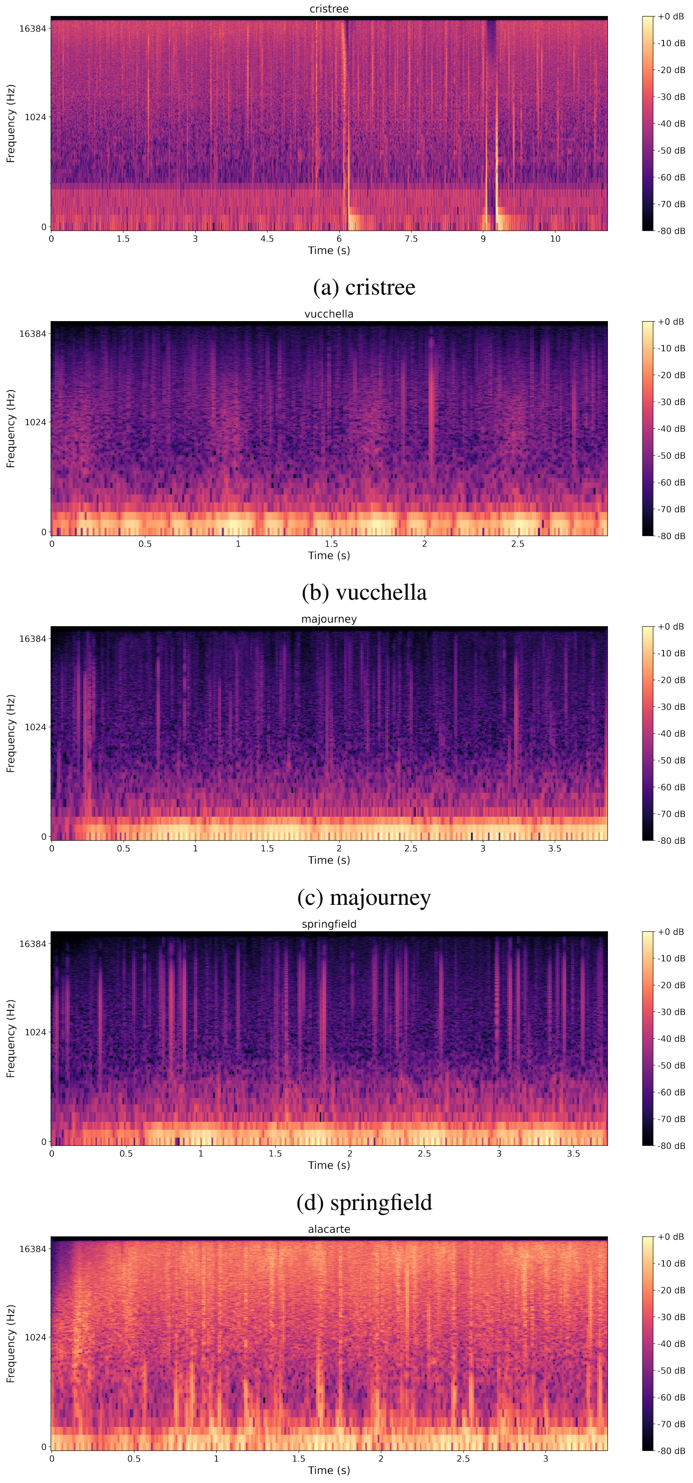
The experiments were conducted to test the separation models on simulated historical choro recordings. To do that, 20 tracks were carefully selected from the test set of the choro dataset to ensure they were free of leakage. One of the typical characteristics of choro is the rich contrapuntal interaction between performers, which is facilitated and spontaneously captured in studio if the musicians play in the same space. This usually results in some sound leakage from one instrument (or family of instruments) into the microphone of another. Of course, this effect would prevent the proper use of the original segregated tracks as reference signals for evaluating system performance.

To simulate historical 78 RPM noisy recordings, we combine the clean tracks of choro test set and the 5 preconditioned noise tracks according to

$$C_{\text{noisy}} = \beta.(G.N_{pp} + C_{\text{clean}}), \qquad (1)$$

[1] Available at https://github.com/facebookresearch/demucs/tree/v3

[2] The albums mentioned are available for purchase at https://www.choromusic.com/



(a) cristree

(b) vucchella

(c) majourney

(d) springfield

(e) alacarte

**Figure 1**: Spectrogram of the five noise samples selected.

where $\beta$ represents the global track gain, $G$ is the gain applied to the pre-processed noise track $N_{pp}$ to produce the desired SNR, and $C_{\text{clean}}$ is the track extracted from choro test set. Noise gains were adjusted to induce SNR values of 10, 20, 30, and 40 dB for each track. In total, 400 noisy mixtures were generated according to this scheme, outlined in Figure 2. It should be noticed that the choro dataset albums used for training were not subjected to any pre-selection regarding instrument leakage.

Another key aspect to highlight is the choice to separate simulated historical tracks instead of original recordings from the early 20th century, which was based on two main considerations. First, we aim to test the robustness of the model, i.e., its ability to accurately separate instrument families in the presence of background noise. To achieve this, it is essential to vary the types of noise both in terms of their spectral components and their power levels—an approach that would be impractical with real recordings, given the limited availability of such material. Secondly, to calculate the objective measures commonly used in the source separation community, it is essential to compare the separated track with the ground truth, i.e., the isolated recording of the instrument without noise. Clearly, acquiring such material from original recordings of that era is practically unfeasible, as technology for multichannel recording per instrument was not yet available.

To assess the quality of separation, three objective metrics are widely adopted in the literature: signal-to-**distortion** ratio (SDR), signal-to-**interference** ratio (SIR) and signal-to-**artifact** ratio (SAR), defined in [16]. Table 2 presents these metrics computed to compare the performances for different levels of SNR. The average value is calculated considering all the test tracks and all selected noises. It can be observed that for SNRs of 20, 30, and 40 dB, the separation results are closely aligned with those obtained from the clean tracks in the last row. This indicates a degree of robustness in the system when processing noisy recordings. Table 3 addresses the results obtained exclusively for the SNR of 10 dB. For this case of worst performance, it is possible to observe the behavior of the separation for each noise individually. As the values of the metrics SDR, SIR, and SAR metrics decrease, respectively, the signal tends to become almost imperceptible, presents a higher amount of artifacts that mask the original signal, and the separation is affected by interference from other families, indicating that the system struggles to differentiate them.

Overall, the separation works well for recordings with noise. From an auditory perspective, the instrument families are separated in a similar way as in the cases of noise-free tracks for almost all noise types and SNR values, with the percussion family being the most critical case for low SNR values.

Lower SNRs tend to impair the separation quality, particularly for the percussion family. We verified auditorily that in certain cases, the nature of the noise is very similar to that of the *pandeiro* (the primary percussive instrument in the database), especially due to the high-frequency sound produced by the *platinelas* (small iron plates). This is predominantly observed in the cristree and alacarte noise signals, which contain a considerable amount of information above 512 Hz. In contrast, the other noise signals do not exhibit this characteristic with the same intensity. Reflecting the low values in SDR and SAR, a similar behavior was found through perceptual evaluation.

As for the string family, separation is quite reasonable

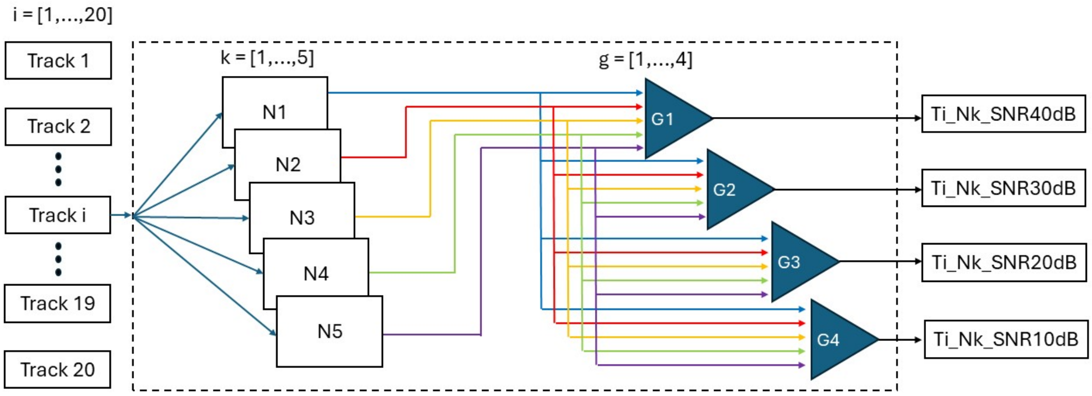


**Figure 2**: Scheme for preparing the noisy tracks of the test set.

**Table 2**: Average value of SDR, SIR and SAR for each SNR (considering the 20 noisy tracks in the test set).

| Noise SNR | STRINGS | | | WIND | | | PERCUSSION | | |
|---|---|---|---|---|---|---|---|---|---|
| | SDR | SIR | SAR | SDR | SIR | SAR | SDR | SIR | SAR |
| 20 dB | 14.210 | 28.248 | 16.104 | 14.134 | 24.596 | 15.590 | 5.657 | 16.918 | 5.234 |
| 30 dB | 15.777 | 28.298 | 17.878 | 15.365 | 25.130 | 17.112 | 7.176 | 19.231 | 7.380 |
| 40 dB | 16.128 | 28.230 | 18.503 | 15.582 | 25.280 | 17.414 | 7.453 | 19.452 | 8.233 |
| Clean | 16.130 | 28.373 | 18.399 | 15.700 | 25.407 | 17.576 | 7.6172 | 19.340 | 8.334 |

**Table 3**: Average value of SDR, SIR and SAR computed for the 20 noisy tracks in test set with an SNR of 10 dB.

| Noise category | STRINGS | | | WIND | | | PERCUSSION | | |
|---|---|---|---|---|---|---|---|---|---|
| | SDR | SIR | SAR | SDR | SIR | SAR | SDR | SIR | SAR |
| alacarte | 12.749 | 27.398 | 14.309 | 14.169 | 24.699 | 15.617 | -0.016 | 6.042 | -6.685 |
| cristree | 11.880 | 27.291 | 14.128 | 14.035 | 25.083 | 16.023 | 0.597 | 7.475 | -6.396 |
| majourney | 6.264 | 27.656 | 9.166 | 9.545 | 21.626 | 10.154 | 6.373 | 17.835 | 6.850 |
| springfield | 5.663 | 27.465 | 8.090 | 9.935 | 21.005 | 10.774 | 6.077 | 17.126 | 6.395 |
| vucchella | 5.102 | 26.898 | 7.091 | 10.484 | 20.101 | 11.447 | 5.631 | 16.241 | 6.122 |

with an SNR of 10 dB with the alacarte and cristree noises. This occurs mainly because, for these two cases, the model extracts some of the noise from the string family and classifies it as percussion, thus favoring the string metric, at the cost of impairing the percussion metric.

In the case of the wind family, results for SNR of 30 and 40 dB show little variation for SDR, SAR e SIR values, indicating that separation is easier for this family. In many tracks, the system acts as a form of denoiser, partially eliminating noise, particularly during pauses in the melody played by the wind instrument.

## 4. CONCLUSIONS AND FUTURE WORK

In this work, we investigate the effects of separating musical instruments into families on historical recordings of the traditional Brazilian genre choro. To achieve this, we employ a deep learning approach using an architecture known as Hybrid Demucs, trained on a choro database. We use 20 leakage-free test tracks combined with a database of noise from 78 RPM records to simulate historical recordings, creating different levels of SNR (10, 20, 30, and 40 dB). Finally, we evaluate the results using traditional objective metrics for music separation (SDR, SIR, and SAR), in addition to a careful consideration of our own subjective evaluation.

The results obtained are promising, demonstrating that the system is robust when dealing with tracks containing additive noise, even though it has been pretrained on clean recordings. Some families, such as percussion, face greater challenges in separation at lower SNRs, while others, such as wind instruments, show good results across all SNR levels. Some separation results, as well as the noises used, are available for listening at `https://www02.smt.ufrj.br/~pedro.donadio/index_Lamir.html`.

Several ideas for future work arise from this approach, with the main one being fine-tuning using tracks with noise. The expectation is that, in addition to separating the families, a form of denoising will occur if the noise is included as training data for the model.

## 5. REFERENCES

[1] Y. Özer and M. Müller, “Source separation of piano concertos with test-time adaptation,” *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pp. 493–500, June 2020.

[2] S. Sarkar, E. Benetos, and M. Sandler, “Ensembleset: A new high quality synthesised dataset for chamber ensemble separation,” *23rd International Society for Music Information Retrieval Conference (ISMIR)*, 2022.

[3] F. Stöter, A. Liutkus, and N. Ito, “The 2018 signal separation evaluation campaign,” *Proc. Int. Conf. Latent Variable Anal. Signal Separation*, p. 293–305, 2018.

[4] G. Fabbro, “The sound demixing challenge 2023 – music demixing track,” *Transactions of the International Society for Music Information Retrieval*, vol. 7, no. 1, p. 63–84, 2024.

[5] E. Cano, D. FitzGerald, A. Liutkus, M. D. Plumbley, , and F. Stöter, “Musical source separation: An introduction,” *IEEE Signal Process. Mag.*, vol. 36, no. 1, p. 31–40, Jan. 2019.

[6] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, D. FitzGerald, and B. Pardo, “An overview of lead and accompaniment separation in music,” *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 26, no. 8, p. 1307–1335, Aug. 2018.

[7] P. Patel, S. Shah, S. Prasad, A. Gada, K. Bhowmick, and M. Narvekar, “Audio separation and classification of indian classical instruments,” *Eng. Appl. Artif. Intell*, vol. 133, 2024.

[8] F. Stoter, S. Uhlich, A. Liutkus, and Y. Mitsufuji, “Open-unmix - a reference implementation for music source separation,” *Journal of Open Source Software*, 2019. [Online]. Available: https://doi.org/10.21105/joss.01667

[9] R. Hennequin, A. Khlif, F. Voituret, and M. Moussallam, “Spleeter: a fast and efficient music source separation tool with pre-trained models,” *Journal of Open Source Software*, vol. 5, no. 50, p. 2154, 6 2020.

[10] A. Défossez, “Hybrid spectrogram and waveform source separation,” in *ISMIR 2021, Music Demixing Workshop (DMX)*, 11 2021.

[11] Y. Luo and N. Mesgarani, “Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation,” in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 8 2019, pp. 1256–1266.

[12] ——, “Tasnet: Time-domain audio separation network for real-time, single-channel speech separation,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, 4 2018, pp. 696–700.

[13] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, “The MUSDB18 corpus for music separation,” december 2017. [Online]. Available: https://doi.org/10.5281/zenodo.1117372

[14] J. B. Siqueira, *Três vultos históricos da música brasileira: Mesquita - Callado - Anacleto.* Rio de Janeiro: FUNARTE, 1970.

[15] A. Diniz, *Almanaque do Choro*, 3rd ed. Rio de Janeiro: Zahar, 2003.

[16] E. Vincent, R. Gribonval, and C. Févotte, “Performance measurement in blind audio source separation,” *IEEE Transactions on Audio, Speech and Language Processing, Institute of Electrical and Electronics Engineers*, vol. 14, p. 1462–1469, 6 2006.

[17] E. Moliner and V. Välimäki, “A two-stage u-net for high-fidelity denoising of historical recordings,” in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 841–845.