Classical Guitar Duet Separation Using GuitarDuets: A Dataset of Real and Synthesized Guitar Recordings

Marios Glytsos[13], Christos Garoufis[123], Athanasia Zlatintsi[123], Petros Maragos[13]

[1]Robotics Institute, Athena Research Center, Athens, Greece
[2]Institute of Language and Speech Processing, Athena Research Center, Athens, Greece
[3]School of ECE, National Technical University of Athens, Athens, Greece
mariosgly@gmail.com, {christos.garoufis,athanasia.zlatintsi}@athenarc.gr, maragos@cs.ntua.gr

Introduction

Music Source Separation (MSS) focuses on separating instruments from composite audio. Research typically targets **multi-timbral** separation, where instruments have **distinct sounds** (e.g., vocals, drums).

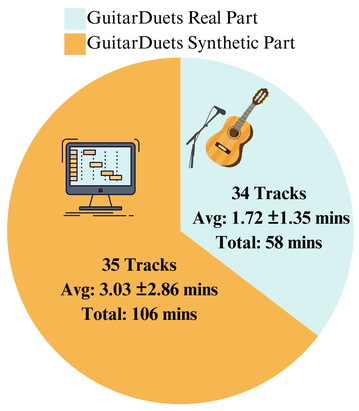
Challenge: In contrast, **monotimbral** MSS, deals with separating instrumental tracks of the **same type** (e.g. bowed strings, guitars), which poses a challenge due to the similar timbres of the instruments.

Contribution:

- Introduction of the **GuitarDuets** dataset, consisting of 3 hours of real & synthesized **classical guitar duet** stem recordings coupled with **note-level annotations** for the synthesized part.
- Cross-dataset evaluation using **DeMucs,** a state-of-the-art separation model.
- Exploration of score integration into the architecture, via a **joint transcription-separation** framework.

Dataset

- **Real Recordings - GuitarDuets (R):**
  - **Simultaneously recorded** performances from 2 guitars.
  - Recording settings: Acoustically treated room, high-quality (Presonus PM-2) microphone.
  - Timbral diversity ensured by using multiple guitars.
  - Independently recorded, **bleeding-free** testing set.
- **Synthesized Recordings - GuitarDuets (S):**
  - Utilized MIDI scores and a variety of **virtual instrument plugins** ⟶ diverse timbral characteristics.



Datasets	Real Data Incl.	Monotimbral	Polyphonic	Note Annotations	Duration
musdb18	✓	✗	✓	✗	ca. 10h
MoisesDB	✓	✗	✓	✗	ca. 14.5h
URMP	✓	✗	✓	✓	1h 6min
SLAKH	✗	✗	✓	✓	ca. 145h
EnsembleSet	✗	✓	✗	✓	6h 9min
GuitarSet	✓	✓	✓	✓	3h 3min
GuitarDuets	✓	✓	✓	Partial	2h 44min

Methodology

Separation Model:

- Hybrid Transformer DeMucs (HT-DeMucs) [1]: Combines U-Nets in temporal and spectral domains.

Joint Transcription-Separation Framework:

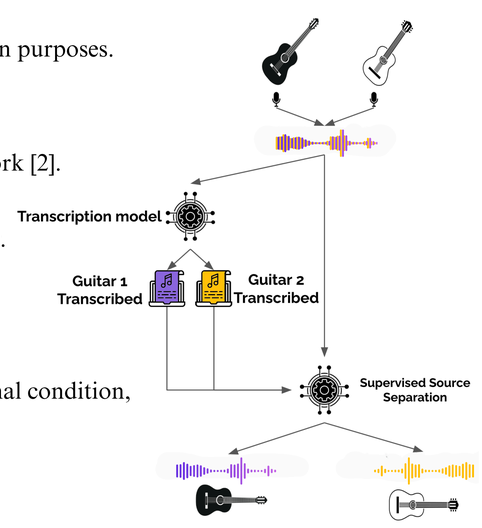
- Independent networks for transcription and separation purposes.

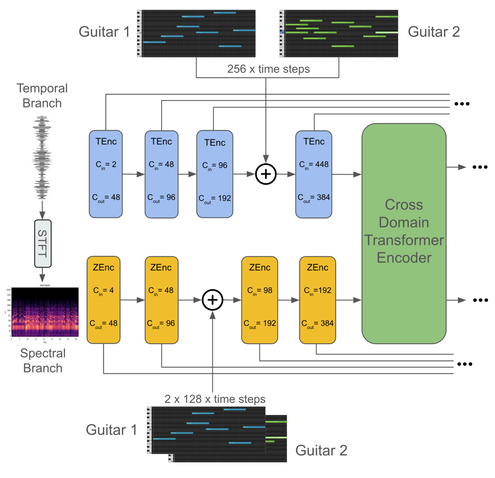
Transcription Network:

- Based on the Residual Shuffle-Exchange (RSE) network [2].
- Intakes **combined audio** from both guitars.
- Generates **pianoroll estimates**, separate for each guitar.

Separation Network:

- Adaptation of the baseline HT-Demucs model.
- Receives the generated **pianoroll estimates** as an external condition, along with the **mixed audio**.
- Creates **separate audio outputs** for each guitar.





Intuition

- Note Interdependencies: The transcription network captures **patterns** and **interdependencies** between notes and guitars.
- The separation model refines the output based on **timbre**, leveraging the transcription information to improve accuracy.

Loss Function: **Permutation-invariant** L1 loss: Addresses the challenge of separating monotimbral sources.
Metrics: SDR, SI-SDR, SIR, SAR between the estimated guitar tracks and the ground truth audio.

Acknowledgments

This research was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the "3rd Call for H.F.R.I. Research Projects to support Post-Doctoral Researchers" (Project Number: 7773). For more information: https://i-mreplay.athenarc.gr/

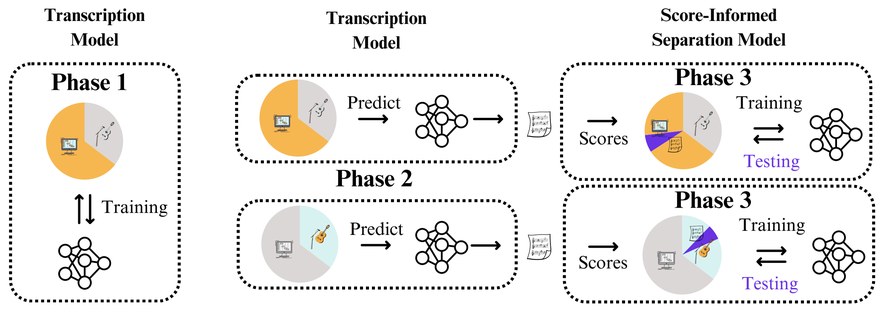
Experimental Setup

Cross-Dataset Experiments: Conducted using various combinations of **GuitarDuets(R),** **GuitarDuets(S),** and **GuitarSet** [3] using HT-DeMucs; test results correspond to **GuitarDuets(R)**

Joint Transcription and Separation Experiments:

First Experiment: Experimented with GuitarDuets(S) to determine the **optimal branch** of Demucs for inputting note-level annotations using ground truth labels.

Second Experiment: 1) Trained a **transcription model** on GuitarDuets(S) using its ground-truth annotations, 2) Used it to generate **transcript estimates** on both GuitarDuets(S) and GuitarDuets(R) and 3) Utilized the estimated transcripts as **auxiliary information** for separation in both subsets.



Results & Discussion

Cross-Dataset Evaluation:

Source Datasets			Metrics			
GuitarDuets(R)	GuitarDuets(S)	GuitarSet	SDR	SI-SDR	SAR	SIR
✓	✓	✓	2.566	0.262	4.866	7.631
✓	✗	✓	2.941	1.021	**10.191**	6.952
✗	✓	✓	2.815	-0.018	8.248	7.560
✗	✗	✓	3.005	**1.337**	7.152	**8.145**
✗	✓	✗	1.836	-1.098	6.357	6.359
✓	✗	✗	2.983	0.019	4.526	7.643
✓	✓	✗	**3.401**	0.591	4.692	7.905

- Combining **real** and **synthesized** data from GuitarDuets leads to the highest SDR scores.
- **Artifact introduction** when mixing different domains/datasets due to structural differences.

Score-Informed Evaluation:

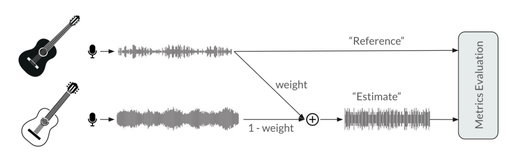
Dataset	Note Labels	Branch Conditioning		Metrics			
		Time	Frequency	SDR	SI-SDR	SAR	SIR
GuitarDuets(S)	Ground Truth	✗	✗	2.572	-0.039	3.973	9.256
		✓	✗	4.404	1.595	4.085	10.352
		✗	✓	3.924	1.442	4.068	9.703
		✓	✓	**4.790**	**1.766**	**4.339**	**11.309**

- Ground-truth note labels: Significant metric improvement, especially regarding **SIR**.
- Better results when conditioning on **both DeMucs branches**.

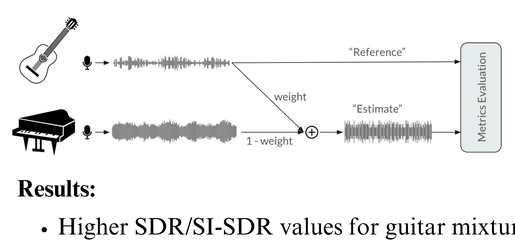
Dataset	Note Labels	Branch Conditioning		Metrics			
		Time	Frequency	SDR	SI-SDR	SAR	SIR
GuitarDuets(S)	Estimated	✓	✓	**2.696**	-0.057	3.271	8.845
	None	✗	✗	2.572	**-0.039**	**3.973**	**9.256**
GuitarDuets(R)	Estimated	✓	✓	**3.174**	**0.531**	**4.818**	**8.100**
	None	✗	✗	2.983	0.019	4.526	7.643

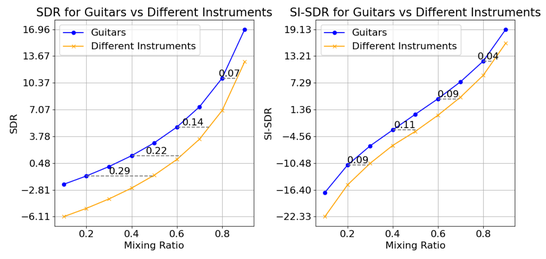
- Evident **improvement** in all metrics when using estimated transcripts on GuitarDuets(R).
- Slightly **deteriorated performance** in GuitarDuets(S)⟶reduced generalization of the transcription model?

Comparison to SOTA?: Most studies focus on **distinct instruments**, making direct comparison of our results difficult due to the **timbral similarity** of the guitars.



Thus, (SI-)SDR values were calculated for normalized artificial guitar-guitar (**monotimbral**) and guitar-piano (**multitimbral**) mixtures, mixed at varying ratios *a*.





Results:

- Higher SDR/SI-SDR values for guitar mixtures than for different instruments.
- SDR value of 5 required *a* approx. 0.8 for multitimbral but 0.6 for monotimbral, showing timbral similarity affects metric accuracy.

Conclusions & Future Work

- Presented the **GuitarDuets** dataset for the task of mono-timbral music source separation.
- Note-level information can aid separation through a **joint transcription-separation** framework, when integrated within a **permutation-invariant training** scheme..

Future Work:

- Expand the size of the dataset, by both recording and synthesizing additional data.
- Explore different architectures and conditioning methods for the joint framework.

References

[1] Rouard, Simon, Francisco Massa, and Alexandre Défossez. "Hybrid transformers for music source separation.", in Proc. ICASSP 2023.
[2] Andis Draguns, Emils Ozolins, Agris Sostaks, Matiss Apinis, and Karlis Freivalds. "Residual shuffle-exchange networks for fast processing of long sequences". CoRR, abs/2004.04662, 2020.
[3] Xi, Qingyang, et al. "GuitarSet: A Dataset for Guitar Transcription.", in Proc. ISMIR 2018.