CLASSICAL GUITAR DUET SEPARATION USING GUITARDUETS - A DATASET OF REAL AND SYNTHESIZED GUITAR RECORDINGS

Marios Glytsos[1,3] **Christos Garoufis**[1,2,3] **Athanasia Zlatintsi**[1,2,3] **Petros Maragos**[1,3]
[1]Robotics Institute, Athena Research Center, Athens, Greece
[2]Institute of Language and Speech Proc., 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

ABSTRACT

Recent advancements in music source separation (MSS) have focused in the multi-timbral case, with existing architectures tailored for the separation of distinct instruments, overlooking thus the challenge of separating instruments with similar timbral characteristics. Addressing this gap, our work focuses on monotimbral MSS, specifically within the context of classical guitar duets. To this end, we introduce the GuitarDuets dataset, featuring a combined total of approximately three hours of real and synthesized classical guitar duet recordings, as well as note-level annotations of the synthesized duets. We perform an extensive cross-dataset evaluation by adapting Demucs, a state-of-the-art MSS architecture, to monotimbral source separation. Furthermore, we develop a joint permutation-invariant transcription and separation framework, to exploit note event predictions as auxiliary information. Our results indicate that utilizing both the real and synthesized subsets of GuitarDuets leads to improved separation performance in an independently recorded test set compared to utilizing solely one subset. We also find that while the availability of ground-truth note labels greatly helps the performance of the separation network, the predicted note estimates result only in marginal improvement. Finally, we discuss the behavior of commonly utilized metrics, such as SDR and SI-SDR, in the context of monotimbral MSS.

1. INTRODUCTION

The task of music source separation (MSS) involves dissecting a musical composition into its constituent sources, typically segregating individual instruments or vocal tracks from a composite audio mixture [1, 2, 3]. Due to the multitude of the co-playing sources, as well as its utility in a variety of applications [1], MSS stands as a significant challenge in the field of Music Information Retrieval (MIR) [4]. The majority of research efforts have focused on multi-timbral music source separation [5,6,7]. In this case, the goal is the separation of distinct instrumental sources from a mixture, where the sources belong to different instrument families or types such as vocals, bass, drums and others, and as such can be framed as an extension of the task of speech denoising into the music domain [8].

Through the advancement of digital signal processing [9] and deep learning [5, 10], considerable progress has been made in extracting distinct instrumental tracks from complex musical compositions. Most recent, deep-learning based approaches for MSS are divided into spectrogram-domain approaches [11], waveform-domain methodologies [10,12,13,14,15] and hybrid ones, working simultaneously in both domains [5]. Spectrogram-domain methods typically isolate sources via mask prediction [16], waveform-domain approaches enhance source separation by applying spectrogram techniques in a learned latent space [14, 17], or directly predicting the isolated waveforms [12], with the additional advantage of incorporating phase information, while hybrid architectures [5] leverage the strengths of both. Moreover, recent findings have highlighted the benefits of using static or dynamic activity labels [18, 19, 20, 21, 22], as well as jointly training transcription and source separation modules [23, 24], which enhances task performance, paralleling efforts in simultaneous speech recognition and separation training [25].

However, an area that remains relatively underexplored is monotimbral music source separation. This subfield of MSS focuses on extracting audio components that belong to the same instrument family, or different builds of the same instrument. It can be viewed as the counterpart of the speaker separation problem [31] in the music domain. While this similarity has led to the development of similar methodologies for network training [32], speaker separation, especially within the context of, rarely available in MSS datasets, multi-microphone recordings [33], can also rely on spatial cues. The limited exploration in this area can largely be attributed to the historical focus on isolating the most prominent instruments in popular music, while the demand for separation of instruments with close timbral characteristics is less pronounced. Indeed, there are very few datasets suitable for training algorithms on the task of separating instrumental tracks from the same

Datasets	Real Data Incl.	Monotimbral	Polyphonic	Note Annotations	Duration
musdb18 [7]	✓	✗	✓	✗	ca. 10h
MoisesDB [26]	✓	✗	✓	✗	ca. 14.5h
URMP [27]	✓	✗	✓	✓	1h 6min
SLAKH [28]	✗	✗	✓	✓	ca. 145h
EnsembleSet [29]	✗	✓	✗	✓	6h 9min
GuitarSet [30]	✓	✓	✓	✓	3h 3min
GuitarDuets	✓	✓	✓	Partial	2h 44min

Table 1: Comparison of the GuitarDuets dataset with existing datasets in the literature for music source separation; we note that GuitarSet is strictly monotimbral, since it was entirely recorded using one guitar.

	GuitarDuets(R)	GuitarDuets(S)
# Tracks	34	35
Dur./Track (mins)	1.72 $\pm$1.35	3.03 $\pm$2.86
Total Dur. (mins)	58	106
Notes/sec.	-	7

Table 2: Detailed statistics of the real and synthesized subsets of the GuitarDuets dataset; note statistics are included for the synthetic subset only.

instrument family in a polyphonic context [30], with the majority of publicly avalaible datasets covering the case of separating mixtures of multiple singing voices [34, 35, 36].

In this paper, we attempt to bridge this gap by introducing GuitarDuets [1], a dataset consisting of a total of ca. 3 hours of real and synthesized guitar duet recordings, along with partial note-level annotations, which can be leveraged as auxiliary score information. We benchmark GuitarDuets in the tasks of i) unconditional guitar duet separation and ii) score-informed duet separation, using the hybrid Demucs [5] as our separation model. We also examine the possibility of integrating note-level predictions into a joint transcription and separation framework. In more detail, the main contributions of this work are:

- Introduction of GuitarDuets, a dataset for monotimbral music source separation, featuring both real classical guitar duet recordings and synthetic recordings generated from online transcriptions and virtual instruments. The synthetic portion includes MIDI representations for each guitar part, enriching the dataset for algorithm training and detailed analysis.
- Extensive cross-dataset evaluation across various conditions, including real and generated synthetic data, as well as the existence or absence of auxiliary score information in specific Demucs branches.
- Development of a joint transcription and separation framework, which incorporates transcription predictions, by adapting existing architectures [5,37] to the task of monotimbral source separation with the introduction of a permutation-invariant [32] loss. We show that incorporation of these note-level predictions can improve the separation of real guitar duets.
- Finally, we analyze the behavior of commonly-utilized source separation metrics in the context of classical guitar duets to understand their effectiveness when applied in sources with similar timbres.

[1] The dataset is available at: https://zenodo.org/records/12802440

2. DATASETS

2.1 Existing Datasets

Datasets available for music source separation or transcription are primarily divided into multitimbral and monotimbral ones, each offering instrument-specific tracks or stems, often accompanied by transcriptions. Multitimbral datasets such as musdb18 [7], URMP [27], MedleyDB [38], MoisesDB [26] and SLAKH [28] are most prominent, featuring both real and synthesized data from a broad spectrum of instruments; some extend to multimodal forms, including for instance audiovisual elements [27].

In contrast, monotimbral instrumental datasets, notably fewer in number, include focused collections such as GuitarSet [30] and EnsembleSet [29]. GuitarSet provides detailed annotations for acoustic guitar recordings, consisting of pairs of comping and soloing performances, while EnsembleSet targets chamber ensembles with high-quality synthetic reproductions of classical music. Despite their utility, these monotimbral datasets face some limitations, namely: GuitarSet's structure, with distinct solo and accompaniment parts, oversimplifies the separation task due to the distinct role of each guitar. Also, the lack of timbral differences between the two guitars prevents the networks from focusing on timbral cues for the separation task. On the other hand, EnsembleSet's reliance on synthetic data introduces a realism gap, underscoring the need for datasets that more accurately capture the dynamics of live musical performances. Moreover, the instruments it contains are largely monophonic, which hinders its use for scenarios with polyphonic co-playing instruments.

2.2 The GuitarDuets Dataset

In this section, we will describe the GuitarDuets dataset, comprising both real recordings of classical guitar duets and synthesized recordings, leveraging virtual instruments and MIDI scores. This approach aimed to provide an original and realistic set of guitar duet recordings for training and evaluating deep learning algorithms on monotimbral MSS, while simultaneously overcoming their limited duration, granting ample training data and enabling analysis between real and synthetic datasets. In total, our dataset comprises 58.6 minutes of real data and 106 minutes of synthesized data, amounting to 164.6 minutes overall. A comparison of the GuitarDuets with the most prominent datasets for MSS in the literature is outlined in Table 1,

whereas detailed statistics about both the real and synthesized subsets of GuitarDuets are given in Table 2.

Real Recordings: For the recordings featuring real classical guitars, we utilized a quiet, acoustically treated room and high-quality condenser microphones (Presonus PM-2), one for each guitar. During the recording process, four different classical guitars were used, with some tracks (16 min.) replayed using different guitars to further enhance timbral diversity. Simultaneous play was crucial for capturing the musical interplay between the two guitarists. The whole recording process resulted in the recording of 27 guitar duets (per-track duration: 123 ± 82 sec.), mostly from the Modern Classical and Nuevo Tango genres and from the Romantic Period. This approach, while essential for the integrity and realism of the dataset, introduced a challenge with cross-microphone sound bleeding. This crossover of sound presented a significant concern, as it compromises the isolation of the individual guitar tracks, impacting the quality of the dataset. In addressing the issue of source bleeding in microphones, we recorded a specialized test set that is free from such leakage. This set, consisting of 7 tracks (per-track duration: 39 ± 13 sec.), was created to ensure the absence of cross-feed between microphones. Each guitar track was exported as a 44,100 Hz, 16-bit WAV file in stereo format, with mixed audio files created by averaging individual guitar performances.

Synthetic Recordings: Despite the inherent realism of the real recordings, their small duration could prove problematic for network training, whereas the single recording setup utilized could import biases. A commonly utilized shortcut to increase the duration of real recordings is to virtually augment them, by synthesizing additional pieces based on note-level transcriptions and virtual music instruments, which has proven effective not only in generating multitrack datasets [28, 29], but also in tasks such as tablature generation [39, 40]. In our case, "Session Guitarist - Picked Nylon" [2], a sample-based virtual instrument, was utilized to generate classical guitar sounds. It offers a wide range of playing styles, capturing the nuances of nylon-stringed guitars. We selected guitar duet MIDI scores from the MuseScore community [3], representing a broad spectrum of pieces. Logic Pro X [4] served as the digital audio workstation (DAW) for transforming MIDI scores into realistic guitar performances. By configuring multiple instances of the PICKED NYLON plugin with distinct timbral settings, we produced different guitar sounds. The final dataset was exported as 44,100 Hz, 16-bit WAV files in both stereo and mono formats for mixed audio file creation.

3. METHODOLOGY

3.1 Separation Architecture

In this work the Hybrid Transformer Demucs [5] was used as the separation backbone, consisting of dual U-Nets [16], operating in both time and spectrogram domains, each

[2] https://www.native-instruments.com/en/products/komplete/guitar/session-guitarist-picked-nylon/

[3] https://musescore.com/

[4] https://www.apple.com/logic-pro/

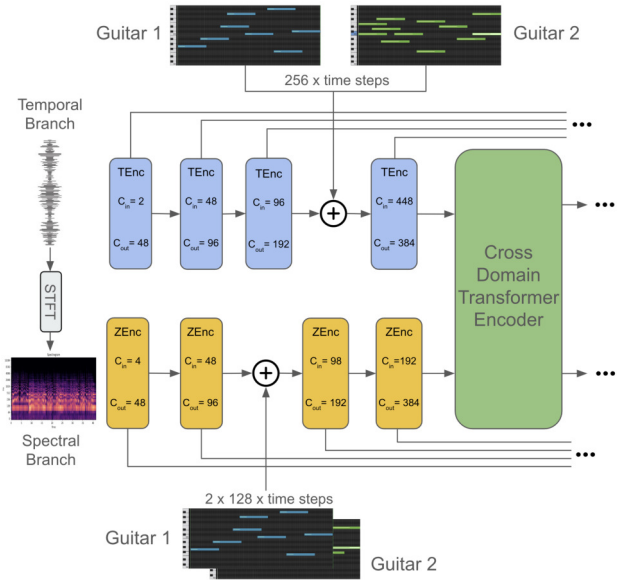

Figure 1: Overview of the incorporation of note-level annotations into the Demucs Architecture.

featuring four encoder and decoder layers. The temporal encoder (TEncoder) downsamples the input waveform through a series of 1D convolutions, whereas the spectral encoder (ZEncoder) gradually compresses the STFT magnitude of the input by applying convolutions across the spectral axis. The traditional convolutional layers, positioned between the encoder and decoder in previous iterations of the Demucs architecture [41] are replaced with a cross-domain Transformer Encoder, composed of interleaved self-attention and cross-attention Encoder layers, each equipped with Layer Scale [42]. The attention mechanism operates with eight heads, and the hidden state size of the feedforward network is four times the dimension of the transformer. The primary decoder layer is shared, branching into both temporal and spectral domains, with the respective decoders built symmetrically to the encoders. The spectral output, post an inverse Short Time Fourier Transform (ISTFT), is merged with the temporal output, producing the model's prediction. We note that in our experiments, the input length is set to 4 seconds.

3.2 Score-Informed Separation

In the context of Score-Informed Separation, the separation network is conditioned on the note-level transcripts of the recordings. To this end, binary vectors indicating the presence or absence of each of the 128 MIDI notes during small temporal frames are concatenated with the intermediate feature maps in each branch, as indicated in Fig. 1. In particular, in the temporal branch, the activity labels of each guitar are inserted after the third TEncoder layer. The binary vector for each guitar has a dimensionality of $128 \times N_s$, where N_s corresponds to the number of samples for each 4-second segment, yielding a combined shape of $256 \times N_s$ for both guitars. Thus, the activity labels have to be downsampled across the temporal axis, to match the resolution of the encoder at this stage. In a parallel manner, within the frequency branch, these binary vectors are introduced following the second ZEncoder

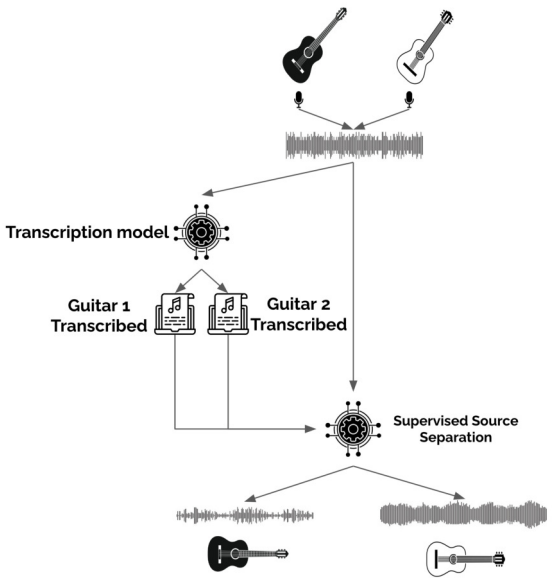

Figure 2: Overview of the proposed methodology for joint transcription and separation of guitar duets.

layer. The shape of the activity labels for concatenation is $2 \times 128 \times N_s$, aligning with the two guitars' MIDI notes. In this case, the activity labels are concatenated with the feature maps across the channel dimension; since the frequency resolution of the ZEncoder, at this stage, matches the number of MIDI notes, resampling occurs only across the temporal axis, by downsampling the note activity labels to the respective temporal resolution of the ZEncoder.

If transcriptions are not available, we use a separate transcription network to generate them creating a joint transcription-separation framework.The first network would intake the combined sounds of the two guitars and generate a binarized piano roll representation for each individual guitar. Afterward, the second model combines the mixed audio and the generated piano rolls to create separate estimates for each guitar as depicted in Figure 2. From a musical endpoint the transcription network could potentially capture note interdependencies and guitar duet patterns through binarized vector features, aiding in note prediction. This transcription informs the separation model, which refines the output by focusing on timbre.

For the transcription architecture, we utilize the Residual Shuffle-Exchange Network (RSE) [37], which has achieved state-of-the-art results in MusicNet [43]. This network enhances the neural Shuffle-Exchange network [44] by employing both Switch and Shuffle layers to capture sequence dependencies effectively, as well as reducing its computational overhead by incorporating strided convolutions. For further details about the architecture we refer to [37, 44]. In our implementation, the RSE's output layer is modified to produce a binarized 2×128-dimensional representation, to assign activity labels for each of the 128 MIDI notes to the corresponding instrument.

4. EXPERIMENTAL EVALUATION

4.1 Experimental Setup

For the separation experiments, we used both the real and synthesized subsets of GuitarDuets, which we will further denote as GuitarDuets(R) and GuitarDuets(S), respectively, as well as the GuitarSet, for which mixtures were generated via addition of the available comping and solo excerpts. We adapted the backbone Demucs model for classical guitar duet separation, modifying it to output two stereo signals, one for each guitar. Data augmentation techniques including channel swapping, time cropping, amplitude scaling and remixing individual guitar parts from different performances were employed during network training to improve generalization. As our loss function we used the quantity:

$$\alpha \cdot \min(|\hat{g_1} - g_1| + |\hat{g_2} - g_2|, |\hat{g_2} - g_1| + |\hat{g_1} - g_2|) + \beta \cdot |(\hat{g_1} + \hat{g_2}) - (g_1 + g_2)|, \quad (1)$$

where the first term corresponds to the traditional permutation-invariant L1 loss between the ground truth signals g_1, g_2 and the output sources $\hat{g_1}$, $\hat{g_2}$, and the second term models the similarity between the sum of the guitar estimates and the input mixtures, encouraging the network to provide separate guitar tracks which neither discard nor duplicate note instances, whereas the weight values were set, after preliminary experiments, to $\alpha = 0.8$, $\beta = 0.2$.

For the transcription architecture experiments, we employed GuitarSet and the GuitarDuets(S) dataset, which contain note-level annotations for individual guitar parts. We transformed labels from GuitarSet (.jams files) and the MIDI files from our dataset to CSV format. All audio files, initially sampled at 44,100 Hz, were resampled to 11,000 Hz and converted to mono, to render them compatible with the RSE backbone [37]. Similar to the separation case, the loss function –in this case, the binary cross entropy– was employed within a permutation invariant framework.

4.2 Cross-Dataset Analysis

For the purposes of the cross-dataset analysis, we consider the GuitarDuets(R) and GuitarDuets(S) subsets as separate datasets, and train the Demucs backbone on various combinations of GuitarSet, GuitarDuets(R) and GuitarDuets(S), using the same experimental protocol and an 80-20 training-validation split; all networks are evaluated on the bleeding-free testing set of GuitarDuets(R). Upon inspection of the results, presented in Table 3, several key insights emerge. Namely, the complete GuitarDuets dataset yielded the highest SDR values for the first guitar. The inclusion of the synthesized subset likely provided additional information that enhanced the model's performance with regards to the SDR. On the other hand, the inclusion of these synthetic parts made the model prone to auditory artifacts, since the highest SAR scores were achieved for the combination of GuitarDuets(R) with the GuitarSet. Finally, we observe that the combination of the complete GuitarDuets dataset with GuitarSet leads in diminished performance, probably due to the structural differences between the training subsets.

Source Datasets			Metrics			
GuitarDuets(R)	GuitarDuets(S)	GuitarSet	SDR	SI-SDR	SAR	SIR
✓	✓	✓	G1: 4.297 G2: 0.835	G1: 3.403 G2: -2.880	G1: 7.670 G2: 2.062	G1: 10.766 G2: 4.495
✓	✗	✓	G1: 4.522 G2: 1.359	G1: 4.280 G2: -2.238	**G1: 9.483** **G2: 10.898**	G1: 6.273 G2: 7.631
✗	✓	✓	G1: 4.493 G2: 1.137	G1: 1.530 G2: -1.566	G1: 8.191 G2: 8.305	G1: 7.038 G2: 8.081
✗	✗	✓	G1: 4.632 **G2: 1.378**	G1: 3.871 **G2: -1.198**	G1: 7.971 G2: 6.332	G1: 7.321 **G2: 8.968**
✗	✓	✗	G1: 3.472 G2: 0.200	G1: 1.857 G2: -4.052	G1: 8.212 G2: 4.502	G1: 8.217 G2: 4.501
✓	✗	✗	G1: 4.952 G2: 1.014	G1: 3.573 G2: -3.536	G1: 7.628 G2: 1.424	G1: 10.413 G2: 4.873
✓	✓	✗	**G1: 5.882** G2: 0.920	**G1: 4.315** G2: -3.133	G1: 8.488 G2: 0.896	**G1: 11.706** G2: 4.104

Table 3: Separation results on the testing set of GuitarDuets(R), according to the datasets utilized during training. G1 corresponds to the 1st guitar, G2 to the 2nd. Higher is better for all metrics.

Dataset	Note Labels	Branch Conditioning		Metrics			
		Time	Frequency	SDR	SI-SDR	SAR	SIR
GuitarDuets(S)	Ground Truth	✓	✗	G1: 4.453 G2: 4.355	G1: 3.117 G2: 0.072	G1: 4.972 G2: 3.197	**G1: 12.411** G2: 8.292
		✗	✓	G1: 4.547 G2: 3.301	G1: 3.293 G2: -0.410	**G1: 4.685** G2: 3.451	G1: 9.523 G2: 9.882
		✓	✓	**G1: 4.717** **G2: 4.863**	**G1: 3.378** **G2: 0.154**	G1: 4.362 **G2: 4.316**	G1: 12.081 **G2: 10.537**
GuitarDuets(S)	Estimated	✓	✓	**G1: 3.414** G2: 1.977	G1: 1.398 **G2: -1.511**	G1: 3.455 G2: 3.087	G1: 10.655 **G2: 7.035**
	None	✗	✗	G1: 2.575 **G2: 2.569**	**G1: 2.436** G2: -2.514	**G1: 4.473** **G2: 3.473**	**G1: 12.795** G2: 5.717
GuitarDuets(R)	Estimated	✓	✓	**G1: 5.313** **G2: 1.035**	**G1: 4.352** **G2: -3.291**	**G1: 7.638** **G2: 1.998**	**G1: 11.110** **G2: 5.089**
	None	✗	✗	G1: 4.952 G2: 1.014	G1: 3.573 G2: -3.536	G1: 7.628 G2: 1.424	G1: 10.413 G2: 4.873

Table 4: Separation results on the testing sets of GuitarDuets(S), GuitarDuets(R), when using solely the respective training sets for training, depending on the availability of note-level annotations and the Demucs branches conditioned on them. G1 corresponds to the 1st guitar, G2 to the 2nd. Higher is better for all metrics.

In our analysis, we observed a consistent discrepancy in the Signal-to-Distortion Ratio (SDR) between the two guitars, where the first guitar exhibited a decent SDR, while the second often fell below a threshold of 1 dB. This pattern suggests that the algorithm may be effectively separating the first guitar by identifying it as the primary source, whereas it perceives the second guitar as background noise, or merely an auditory artifact. It is important to note that the average amplitude of both guitars is on the same scale, so this observation is not attributed to amplitude differences. Notably, we observe that the most consistent SDR values for G2 were achieved when GuitarSet was included in the training set, which we attribute to its relatively noise-free structure.

4.3 Score-Informed Separation Approaches

For the experiments concerning score-informed separation, we investigated the integration of activity labels into our network, considering Demucs' operation across frequency and temporal domains, by using the GuitarDuets(S) as our dataset since it contains note-level annotations. We also investigate, using both GuitarDuets(R) and GuitarDuets(S), whether the joint transcription-separation architecture can aid in effective separation in scenarios where no ground truth data is available. In both cases, a part of the dataset (the bleeding-free subset of GuitarDuets(R), and 10% of GuitarDuets(S)) was used for performance evaluation; the rest were used for training and validation, at an 80:20 ratio.

The analysis, as detailed in Table 4, reveals that while using the temporal branch for note label integration leads to slightly improved results compared to the spectral branch, the hybrid approach achieves the most promising outcomes. This performance can be attributed to the inherent design of the Demucs architecture, which has historically shown improved efficiency when leveraging both domains concurrently [41]. It is also noteworthy that while the integration of ground-truth labels leads to higher SIR values, presumably due to the guidance that these labels

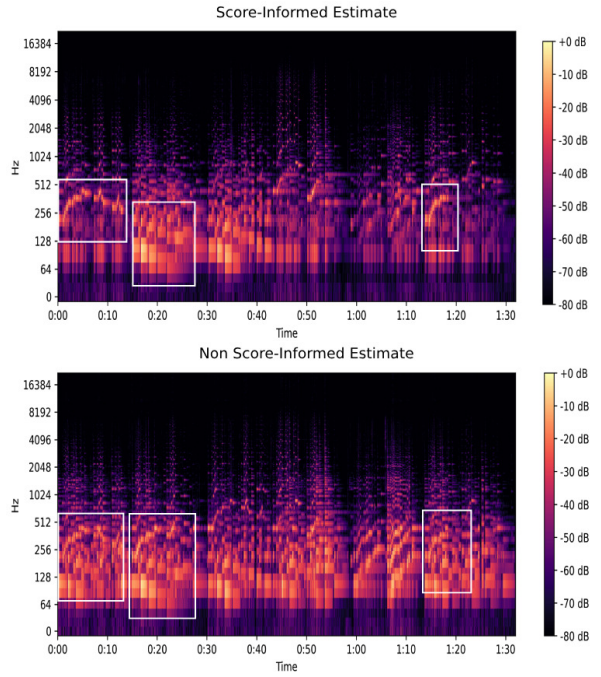

Figure 3: Comparison of spectrogram estimates with estimated (top) note-level annotations, and without those (bottom), for an instance from GuitarDuets (R).

provide to the separation network about the identity of each guitar, the improvement in SAR is marginal.

Regarding the joint transcription and separation framework, the performance does not reach the levels achieved when ground-truth note-level annotations are available, as measured in GuitarDuets(S). On the other hand, while in the case of GuitarDuets(R), the performance is slightly improved when these pseudo-annotations are available, GuitarDuets(S) achieves better results in their absence. A comparison of guitar estimates for the models trained with and without predicted note label information, for an instance of the GuitarDuets(R) test set, can be depicted in Figure 3; we assume that the incorporation of note activity labels enables the separation model to more accurately sustain notes, enhancing the quality of the isolated melodic and accompanying parts. On the other hand, we attribute the performance drop, when using GuitarDuets(S), to the reduced generalization of the transcription network. Since the training set of GuitarDuets(S) was used for its training, the separation network was trained using mostly correct labels, but evaluated with note-level annotations of pieces the transcription network did not use for training.

4.4 Comparative Metric Analysis

In the field of MSS, the evaluation of separation quality is often quantified using metrics such as SDR [45] and SI-SDR [46]. While they have been extensively used in studies focusing on separating different instruments, their behavior on sources with similar timbral characteristics remains less explored. Given that most prior work involves instruments with distinct timbres, direct comparison of our results with SDR values achieved across different datasets may not be appropriate for our study, which focuses on two classical guitars with similar timbral properties.

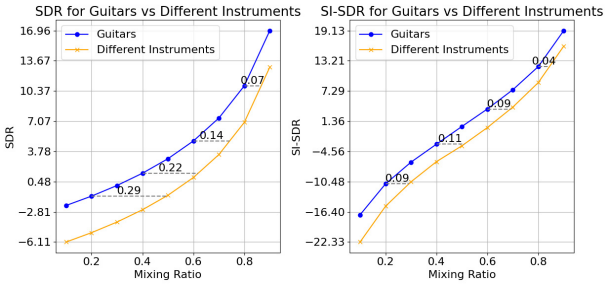

Figure 4: Comparison of the behavior of SDR (left) and SI-SDR (right) when assessing the separation of monotimbral (blue line) or multitimbral (orange line) duets.

In order to identify potential disparities in the behavior of the metrics that can be attributed to timbral similarities in the mixture components, we simulated imperfect estimates of a reference signal x_1 by creating additive synthetic mixtures of the signals x_1, x_2 as:

$$m = \alpha x_1 + (1 - \alpha)x_2, \tag{2}$$

with $\alpha \in (0, 1)$, and measured the values of the SDR, SI-SDR metrics between these mixtures and x_1. We examined two cases using signals derived from Track 29 of the GuitarDuets(S): i) a monotimbral mixture, where both x_1 and x_2 constitute guitar signals, and ii) a multitimbral mixture, where x_2 was synthesized from the second guitar's notes using a piano virtual instrument plugin. To guarantee a fair comparison across all tests, we performed amplitude normalization between the two tracks for each experiment.

The results, displayed in Figure 4, indicate that both metrics for the guitar mixtures are consistently higher than those obtained from mixtures of different instruments. For instance, the mixing ratio α required to reach an SDR value of 5 approaches 0.8 for the multi-timbral case, while 0.6 for the mono-timbral case. This suggests that the timbral similarity between the two guitars introduces a challenge for the metrics to accurately assess the quality of separation.

5. CONCLUSIONS

In this paper, we introduced GuitarDuets, a dataset consisting of both real and synthesized classical guitar duets. We exhibit that our dataset can be utilized for developing monotimbral source separation algorithms within both traditional and score-informed frameworks. We further developed a joint permutation-invariant framework for transcription and separation of monotimbral mixtures, which we show that can lead to improved performance in separation of real guitar duets. In the future, we plan to extend the recordings of both the real and synthesized subsets of GuitarDuets, and provide note-level annotations for its real subset. Furthermore, regarding the joint transcription-separation architecture, we intend to explore more sophisticated ways for integrating the predicted guitar transcripts into the separator [47, 48]. Finally, we aim to conduct extensive listening tests, which will help in further shedding light into both the performance of the various approaches we compare, and the significance of objective metrics within the context of monotimbral audio source separation.

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 Acronym: i-MRePlay, Project Number: 7773).

6. REFERENCES

[1] R. Liu and S. Li, "A review on music source separation," in *Proc. IEEE Youth Conf. on Information, Computing and Telecommunication*, 2009.

[2] Y. Mitsufuji, G. Fabbro, S. Uhlich, F.-R. Stöter, A. Defossez, M. Kim, W. Choi, C. Y. Yu, and K. W. Cheuk, "Music demixing challenge 2021," *Frontiers in Signal Processing*, vol. 1, Jan., 2022.

[3] E. Cano, D. FitzGerald, A. Liutkus, M. D. Plumbley, and F.-R. Stöter, "Musical source separation: An introduction," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 31–40, 2019.

[4] 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 Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 8, pp. 1307–1335, 2018.

[5] S. Rouard, F. Massa, and A. Défossez, "Hybrid transformers for music source separation," in *Proc. Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[6] F.-R. Stöter, S. Uhlich, A. Liutkus, and Y. Mitsufuji, "Open-unmix-a reference implementation for music source separation," *Journal of Open Source Software*, vol. 4, no. 41, p. 1667, 2019.

[7] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "Musdb18-a corpus for music separation," 2017.

[8] E. Vincent, T. Virtanen, and S. Gannot, *Audio Source Separation and Speech Enhancement*. John Wiley & Sons, 2018.

[9] J.-L. Durrieu, A. Ozerov, C. Févotte, G. Richard, and B. David, "Main instrument separation from stereophonic audio signals using a source/filter model," in *Proc. 17th European Signal Processing Conf.*, 2009.

[10] A. Défossez, N. Usunier, L. Bottou, and F. R. Bach, "Music source separation in the waveform domain," 2019. [Online]. Available: http://arxiv.org/abs/1911.13254

[11] Y. Luo and J. Yu, "Music source separation with band-split RNN," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 31, pp. 1893–1901, 2023.

[12] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," in *Proc. Int'l Society for Music Information Retrieval Conf. (ISMIR)*, 2018.

[13] W. Zai El Amri, O. Tautz, H. Ritter, and A. Melnik, "Transfer learning with jukebox for music source separation," in *Proc. Int'l Conf. on Artificial Intelligence Applications and Innovation*, 2022.

[14] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 27, no. 8, p. 1256–1266, Aug. 2019.

[15] C. Garoufis, A. Zlatintsi, and P. Maragos, "HTMD-Net: A Hybrid Masking-Denoising Approach to Time-Domain Monaural Singing Voice Separation," in *Proc. 29th European Signal Processing Conf. (EUSIPCO)*, 2021.

[16] 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.

[17] D. Samuel, A. Ganeshan, and J. Naradowsky, "Meta-learning extractors for music source separation," in *Proc. Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[18] O. Slizovskaia, L. Kim, G. Haro, and E. Gomez, "End-to-end sound source separation conditioned on instrument labels," in *Proc. Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[19] D. Petermann, P. Chandna, H. Cuesta, J. Bonada, and E. Gómez, "Deep learning based source separation applied to choir ensembles," *Proc. Int'l Society for Music Information Retrieval Conf. (ISMIR)*, 2020.

[20] M. Schwabe and M. Heizmann, "Improved separation of polyphonic chamber music signals by integrating instrument activity labels," *IEEE Access*, vol. 11, pp. 42 999–43 007, 2023.

[21] M. Miron, J. Janer, and E. Gómez, "Monaural score-informed source separation for classical music using convolutional neural networks," in *Proc. Int'l Society for Music Information Retrieval Conf. (ISMIR)*, 2017.

[22] M. Gover and P. Depalle, "Score-informed source separation of choral music," in *Proc. Int'l Society for Music Information Retrieval Conf. (ISMIR)*, 2020.

[23] L. Lin, Q. Kong, J. Jiang, and G. G. Xia, "A unified model for zero-shot music source separation, transcription and synthesis," in *Proc. Int'l Society for Music Information Retrieval Conf. (ISMIR)*, 2021.

[24] K. W. Cheuk *et al.*, "Jointist: Simultaneous improvement of multi-instrument transcription and music source separation via joint training," 2023. [Online]. Available: https://arxiv.org/pdf/2206.10805

[25] J. Shi, X. Chang, S. Watanabe, and B. Xu, "Train from scratch: Single-stage joint training of speech separation and recognition," *Computer Speech & Language*, vol. 76, p. 101387, Apr., 2022.

[26] I. Pereira, F. Araújo, F. Korzeniowski, and R. Vogl, "Moisesdb: A dataset for source separation beyond 4-stems," in *Proc. Int'l Conf. of International Society for Music Information Retrieval (ISMIR)*, 2023.

[27] B. Li *et al.*, "Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications," *IEEE Trans. on Multimedia*, vol. 21, pp. 522–535, 2018.

[28] E. Manilow *et al.*, "Cutting music source separation some slakh: A dataset to study the impact of training data quality and quantity," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019.

[29] S. Sarkar, E. Benetos, and M. Sandler, "Ensembleset: A new high-quality synthesised dataset for chamber ensemble separation," in *Proc. Int'l Society for Music Information Retrieval Conf. (ISMIR)*, 2022.

[30] Q. Xi, R. M. Bittner, J. Pauwels, X. Ye, and J. P. Bello, "Guitarset: A dataset for guitar transcription," in *Proc. Int'l Society for Music Information Retrieval Conf. (ISMIR)*, 2018.

[31] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *Journal of the Acoustical Society of America*, vol. 25, pp. 975–979, 1953.

[32] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[33] T. Yoshioka, H. Erdogan, Z. Chen, and F. Alleva, "Multi-microphone neural speech separation for far-field multi-talker speech recognition," in *Proc. Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[34] S. Sarkar, E. Benetos, and M. B. Sandler, "Vocal harmony separation using time-domain neural networks," in *Proc. Interspeech Conf.*, 2021.

[35] C.-B. Jeon, H. Moon, K. Choi, B. S. Chon, and K. Lee, "Medleyvox: An evaluation dataset for multiple singing voices separation," in *Proc. Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[36] R. Schramm and E. Benetos, "Automatic transcription of a cappella recordings from multiple singers," in *AES International Conference Semantic Audio*, 2017.

[37] A. Draguns, E. Ozolins, A. Sostaks, M. Apinis, and K. Freivalds, "Residual shuffle-exchange networks for fast processing of long sequences," in *Proc. AAAI Conf. on Artificial Intelligence*, 2020.

[38] R. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. Bello, "Medleydb: A multitrack dataset for annotation-intensive mir research," in *Proc. Int'l Society for Music Information Retrieval Conf. (ISMIR)*, 2014.

[39] Y. Zang, Y. Zhong, F. Cwitkowitz, and Z. Duan, "Synthtab: Leveraging synthesized data for guitar tablature transcription," in *Proc. Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 1286–1290.

[40] H. Pedroza, W. Abreu, R. Corey, and I. Roman, "Leveraging real electric guitar tones and effects to improve robustness in guitar tablature transcription modeling," 2024. [Online]. Available: https://arxiv.org/abs/2405.14679

[41] A. Défossez, "Hybrid spectrogram and waveform source separation," in *Proc. Music Demixing Workshop (MDX)*, 2021.

[42] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, and H. Jégou, "Going deeper with image Transformers," in *Proc. Int'l Conf. on Computer Vision (ICCV)*, 2021.

[43] J. Thickstun, Z. Harchaoui, and S. Kakade, "Learning features of music from scratch," in *Proc. Int'l Conf. on Learning Representations (ICLR)*, 2017.

[44] K. Freivalds, E. Ozolins, and A. Sostaks, "Neural Shuffle-Exchange Networks - Sequence processing in O(n log n) time," in *Advances in Neural Information Proc. Systems (NeurIPS)*, 2019.

[45] 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.

[46] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?" in *Proc. Int'l Conf. on Acoustics Speech and Signal Processing (ICASSP)*, 2019.

[47] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *Proc. AAAI Conf. on Artificial Intelligence*, 2018.

[48] G. Meseguer-Brocal and G. Peeters, "Conditioned-u-net: Introducing a control mechanism in the u-net for multiple source separations," in *Proc. Int'l Society for Music Information Retrieval Conf. (ISMIR)*, 2019.