NOTEWISE EVALUATION FOR MUSIC SOURCE SEPARATION: A CASE STUDY FOR SEPARATED PIANO TRACKS

Yigitcan Özer[1] **Hans-Ulrich Berendes**[1] **Vlora Arifi-Müller**[1]
Fabian-Robert Stöter[2] **Meinard Müller**[1]
[1]International Audio Laboratories Erlangen, Germany
[2]AudioShake, Inc.
{yigitcan.oezer, meinard.mueller}@audiolabs-erlangen.de

ABSTRACT

Deep learning has significantly advanced music source separation (MSS), aiming to decompose music recordings into individual tracks corresponding to singing or specific instruments. Typically, results are evaluated using quantitative measures like signal-to-distortion ratio (SDR) computed for entire excerpts or songs. As the main contribution of this article, we introduce a novel evaluation approach that decomposes an audio track into musically meaningful sound events and applies the evaluation metric based on these units. In a case study, we apply this strategy to the challenging task of separating piano concerto recordings into piano and orchestra tracks. To assess piano separation quality, we use a score-informed nonnegative matrix factorization approach to decompose the reference and separate piano tracks into notewise sound events. In our experiments assessing various MSS systems, we demonstrate that our notewise evaluation, which takes into account factors such as pitch range and musical complexity, enhances the comprehension of both the results of source separation and the intricacies within the underlying music.

1. INTRODUCTION

Music source separation (MSS) is a key task in Music Information Retrieval (MIR), involving the separation of a musical mixture into individual components like vocals, instruments, and other sound elements [1]. Deep learning techniques have significantly advanced MSS, especially in scenarios with sufficient training data. In particular, this progress is evident in popular music separation, making use of the existence of multitrack recordings inherent in the production process [2–5]. In scenarios with limited training data, systems are often trained using artificially generated mixes through synthesis techniques [6, 7] or data augmentation approaches [8, 9]. An example of such a scenario, also addressed in this paper, is presented in [10], where the goal is to separate piano concertos into piano and orchestra tracks.

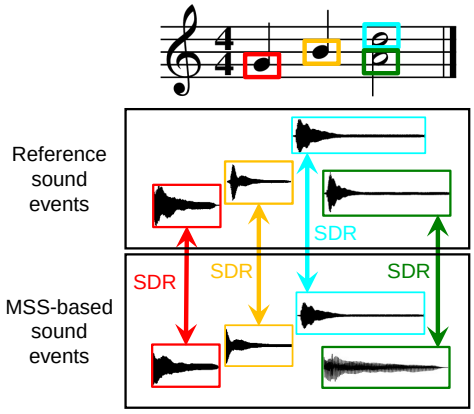

Figure 1: Illustration of the proposed evaluation method for music source separation (MSS), considering signal-to-distortion ratio (SDR) values based on notewise sound events rather than entire recordings.

Extensive efforts have been devoted to evaluating and understanding existing MSS systems. Specifically, in the realm of popular music, evaluation campaigns like the Signal Separation Evaluation Campaign (SiSEC) [11] and the Music Demixing Challenge (MDX) [12] have significantly contributed to the comparison of current systems. In these campaigns, along with evaluations in most approaches described in the literature, one typically relies on quantitative evaluation measures such as the signal-to-distortion ratio (SDR) [13]. These measures are computed and aggregated over audio excerpts or even entire recordings, offering ease of computation and convenience for comparison. However, it is well recognized that such measures provide limited insights into the effectiveness of source separation methods [14, 15]. On the other hand, designing perceptually or musically more relevant measures is challenging, and performing listening tests is often cumbersome and infeasible.

In this paper, we introduce a novel evaluation methodology aimed at attaining a more nuanced understanding of separation quality. This involves comparing a reference signal with a separated signal, utilizing an evaluation

metric based on musically meaningful sound units instead of the entire excerpt. To achieve this, we employ score-informed nonnegative matrix factorization (NMF) [16] to decompose signals into notewise sound events. Then, we calculate SDR values for individual units before aggregating this information in various ways (see Figure 1). This methodology draws conceptual parallels to the evaluation of tasks where automatic speech recognition (ASR) is used as a downstream task. For example, Chen et al. [17] computed word-level and utterance-level metrics to evaluate the quality of the speech separation system.

In a case study, we apply this methodology to the intricate task of separating piano concerto recordings into piano and orchestra tracks. Besides utilizing the Piano Concerto Dataset (PCD) [18], which comprises piano concerto excerpts performed by five pianists in four distinct acoustic settings, we generated piano scores for all the excerpts. We then employed music synchronization techniques [19, 20] to align these scores with all recorded excerpts. As an additional contribution to this paper, we release these annotations, thereby adding a score-based layer to the PCD collection.

In systematic experiments, we apply our evaluation methodology to effectively compare several academic and commercial source separation systems. Our approach uncovers general trends and yields insights into how separation quality is affected by factors like pitch range and musical complexity. In particular, it allows users to explore evaluations in-depth by pinpointing complex passages and challenging sound units where source separation systems tend to fail. Along these lines, we provide qualitative discussions that deepen insights into the behavior of source separation systems and the complexity of the underlying music.

The remainder of the paper is organized as follows. In Section 2, we review relevant literature on source separation and introduce the MSS models used for separating piano concertos. Subsequently, in Section 3, we elaborate on the score-based extension of PCD and outline our evaluation approach, covering NMF-based audio decomposition and notewise SDR-based metrics. In Section 4, we provide details on the experimental settings and report our empirical findings. Finally, in Section 5, we conclude and discuss potential directions for future work.

2. MUSIC SOURCE SEPARATION

As mentioned earlier, the decomposition of music recordings into individual sound components has garnered significant attention in academia and industry in recent years [1–5, 21–23]. While there is a multitude of approaches and architectures proposed in the literature, one can broadly distinguish between spectral-based, waveform-based, and hybrid models. Spectral-based models, such as Open-Unmix (`UMX`) [2] or Spleeter (`SPL`) [3], estimate the magnitude spectrograms of target musical sources given the magnitude spectrogram of an input mixture. Techniques like binary masking, soft masking, or multichannel Wiener filtering are then employed to reconstruct the separated audio signals [24, 25]. Waveform-based models, such as Demucs (`DMC`) [21], process the raw waveform of an input mixture and predict the waveforms of the individual separated sources. Hybrid models integrate complementary information from waveform- and spectrogram-based models, encompassing both spectral and temporal branches. In these architectures, latent representations are combined through the addition of shared layers to leverage the advantages offered by both domains [4, 26, 27]. Examples include the hybrid Demucs model (`HDMC`) introduced in [4] and a system (`AudioShake`) provided by the company AudioShake.

Model ID	Domain	Size (MB)	TS (Hours)
`UMX`	Spectrogram	34	52
`SPL`	Spectrogram	75	52
`DMC`	Waveform	510	52
`HDMC`	Hybrid	319	52
`AudioShake`	Hybrid	N/A	500+

Table 1: MSS models considered in our experiments. TS denotes the size (in hours) of the training set used.

In this paper, we consider the challenging source separation scenario of decomposing piano concerto recordings into distinct piano and orchestral tracks. Piano concertos involve an intricate interplay between the piano and the entire orchestra, resulting in high spectro–temporal correlations among the constituent instruments. Additionally, the absence of multitrack data for training poses an extra challenge for data-driven source separation approaches. To overcome the lack of training data, the approach in [28] proposes generating artificial training data by superimposing randomly chosen audio patches from the solo piano repertoire (e. g., piano sonatas and etudes) and orchestral pieces without piano (e. g., symphonies). The training procedure and comparison of four different models mentioned above are described in [28], including the use of further data augmentation techniques. In our experiments, we employ four pre-trained models from the study [28], shown in Table 1. Additionally, we utilize the commercial system `AudioShake`, trained with over 500 hours of multitrack music recordings spanning various genres, with a focus on popular music. It is important to note that the `AudioShake` system has not been specifically adapted to the piano concerto scenario but is trained on mixtures where the vocal stem is usually dominant.

Finally, we want to emphasize that the implementation details and the reproducibility of the various MSS systems are not the main focus of this paper. Instead, these MSS systems and the piano concerto scenario serve as a framework for illustrating our evaluation methodology, as we will further discuss in Section 4.

3. EVALUATION APPROACH

We now introduce our novel evaluation approach, which we will apply to compare reference piano recordings and separated piano tracks. In Section 3.1, we briefly describe the PCD collection, which will serve as a test dataset, and present our score-based extensions. Then, in Section 3.2,

Room ID	Room Description	Piano	Dur	#Notes
R1	Lecture hall	Yamaha C3	180	1780
R2	Private studio	Yamaha C3X	180	2216
R3	Small concert hall	Seiler	252	2305
R4	Big concert hall	Steinway D	360	3741
		Σ	972	10042

Table 2: Overview of the PCD test set, indicating the four rooms and the piano models employed, and including the duration (in seconds) and the number of notes (piano only).

we revisit the score-informed NMF approach for audio decomposition. Finally, in Section 3.3, we define the SDR-based evaluation metrics, which we use to gain a deeper understanding of the source separation results.

3.1 Piano Concerto Dataset and its Extension

The PCD collection, introduced in [18], is based on piano concerto recordings featuring five different amateur and professional pianists playing along with orchestral recordings provided by the publisher *Music Minus One* [1] . Multitrack recordings with clean piano and orchestra reference tracks were produced from these sessions. The PCD consists of 81 multitrack excerpts, each lasting 12 seconds, selected from 15 piano concertos spanning the Baroque to Post-Romantic period. As summarized in Table 2, the PCD comprises excerpts recorded in four distinct acoustic settings with different grand piano models.

Our novel evaluation approach relies on synchronized score information used for notewise audio decomposition. To this end, we manually generated symbolically encoded sheet music representations using the *Sibelius* software [2] for the piano tracks (and piano-reduced versions of the orchestra tracks, which are not utilized in this study). We employed the *Sync Toolbox* [20] [3] to automatically align the score information with the PCD audio excerpts. To ensure high synchronization accuracy, we computed these alignments in two independent ways: once based on the piano-only tracks and another time based on the piano–orchestra mixes. We then applied fusion techniques to establish the final score annotations. Additionally, expert listeners verified the final results using visual cues provided by the *Sonic Visualizer* [29] and acoustic cues using sonified score annotations overlaid with the audio excerpts. With regard to note onsets, the accuracy of the score annotations for the piano tracks can be expected to lie in the range of 20–40 ms. Additionally, we manually annotated the left-hand (LH) and right-hand (RH) notes, resulting in further musically meaningful note groupings beyond the notewise ones.

We release the symbolically encoded sheet music along with the score-based annotations of the audio excerpts, thereby adding an additional score-based layer to the PCD collection as part of the contributions of this paper. [4]

[1] www.halleonard.com/series/MMONE
[2] www.sibelius.com/
[3] www.github.com/meinardmueller/synctoolbox
[4] www.audiolabs-erlangen.de/resources/MIR/PCD

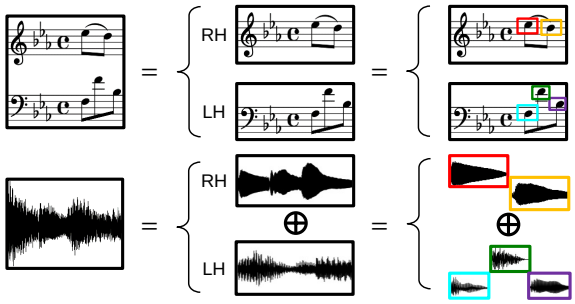

Figure 2: Illustration of the decomposition of the piano track into left-hand (LH), right-hand (RH), and individual note events as indicated by the rectangular windows.

3.2 NMF-Based Audio Decomposition

Nonnegative matrix factorization (NMF) is an algorithm for approximating a nonnegative matrix as the product of two low-ranked nonnegative matrices [30]. In the context of music processing, NMF has been widely applied to decompose a magnitude spectrogram into the product of two nonnegative matrices [31], where the columns of the first matrix encode spectral prototype patterns (called *templates*), and the rows of the second matrix encode their occurrences in time (called *activations*). Thanks to nonnegativity and multiplicative update rules, NMF facilitates the straightforward integration of prior musical knowledge, such as information from an acoustic model or a musical score. For instance, one may constrain the spectral template matrix to enforce a harmonic structure [32] or use aligned score information to constrain the activation matrix [16]. In addition to stabilizing the convergence of the NMF algorithm, such constraints also guide the factorization process to yield decompositions of musical relevance [33].

Following the approach in [34], we adopt a score-informed NMF approach to decompose a given audio signal x into its constituent notewise audio events x^m for $m \in [1:M]$ and a residual signal r such that

$$x = \sum_{m=1}^{M} x^m + r. \tag{1}$$

Here, we assume that we have a score representation with M denoting the number of note events, which are aligned to the audio signal. Note that this alignment does not need to be completely accurate, as it only serves to constrain the NMF algorithm, which can then improve the accuracy in the iteratively learned decomposition process. Besides applying this procedure to obtain a notewise decomposition of the audio signal, one can use the same approach to obtain a decomposition corresponding to note groups, resulting, for example, in the decomposition of the LH and RH notes, as illustrated in Figure 2.

We conclude our description of the NMF-based decomposition approach with some final remarks regarding implementation issues encountered in our experiments based on the PCD test set. Note that, in general, NMF training based on iterative update rules yields more reliable decom-

position results when applied to longer input spectrograms exhibiting a coherent template structure. Therefore, rather than applying the NMF-based decomposition to individual 12-second excerpts, we concatenated all 12-second excerpts recorded in the same room (see Table 2). This strategy is grounded on the assumption that the learned spectral templates, encoding characteristics of the piano and room acoustics, exhibit coherence within each room. Subsequently, we executed the NMF algorithm for 100 iterations on the concatenated data for four subsets with distinct room acoustics, using the same configurations and initialization approach introduced in [16]. This procedure was applied to both the reference piano recordings and the separated piano tracks generated by each MSS model. The resulting notewise decomposition results serve as the basis for our experiments, as reported in Section 4.

3.3 SDR-Based Metrics

The signal-to-distortion ratio (SDR) is a widely used metric in the evaluation of source separation performance, measuring the quality of a separated source by comparing it to the reference source in terms of signal distortion [13]. In our evaluation, when given a reference signal x and a separated signal $\hat{x}$, we use instead the more computationally efficient SDR metric proposed at the recent SDX challange [35], also denoted as SDR:

$$\mathrm{SDR}(x, \hat{x}) := 10 \log_{10} \frac{\|x\|^2}{\|\hat{x} - x\|^2}. \tag{2}$$

Rather than comparing entire excerpts, we use a localized variant referred to as $\mathrm{SDR}_{\mathrm{local}}$ that better accounts for significant level differences within the signal. To this end, we split the reference and separated signals into 1-second segments x_k and $\hat{x}_k$, respectively, defining:

$$\mathrm{SDR}_{\mathrm{local}} := \frac{1}{K} \sum_{k=1}^{K} \mathrm{SDR}(x_k, \hat{x}_k) \tag{3}$$

In our evaluation, we have $K = 12$, as each excerpt in the PCD test set has a duration of 12 seconds.

To obtain a musically more informed evaluation metric, we exploit the decomposition as defined in Equation (1) and consider notewise SDR values:

$$\mathrm{SDR}_{\mathrm{note}} := \mathrm{SDR}(x^m, \hat{x}^m), \tag{4}$$

where x^m and $\hat{x}^m$ denote the notewise sound events of the reference signal and the separated signal, respectively. Note that, using the same score-based activation constraints in the NMF decomposition for x and $\hat{x}$, respectively, the lengths of x^m and $\hat{x}^m$ are identical for a given $m \in [1\!:\!M]$.

4. EXPERIMENTS

In this section, we report on our systematically conducted experiments to highlight the potential of our notewise evaluation methodology. In this context, the piano concerto separation task, along with the five MSS systems described in Section 2, should be considered an illustrative case study of practical relevance. When describing the various experiments, we progress from a coarse to a fine perspective. We start with a more global view of the source separation quality of the MSS systems (Section 4.1). Subsequently, we adopt a more fine-grained perspective, delving into the separation quality depending on the musical pitch (Section 4.2). Finally, we assume an excerptwise view and discuss specific examples to illustrate how separation errors may occur in musically complex situations (Section 4.3). This hierarchical discussion underscores how the notewise evaluation methodology serves as a tool, enabling users to delve into and comprehend not only the separation results but also the intricacies within the underlying music.

Model	Piano	Orchestra
UMX	8.38 ± 4.24	3.61 ± 2.19
SPL	8.16 ± 3.99	3.46 ± 2.25
DMC	7.59 ± 4.38	2.82 ± 2.13
HDMC	9.61 ± 4.42	4.75 ± 2.31
AudioShake	**12.82 ± 4.24**	**8.01 ± 2.97**

Table 3: $\mathrm{SDR}_{\mathrm{local}}$ values (mean and standard deviation) averaged over all PCD excerpts for different MSS systems (see Table 1).

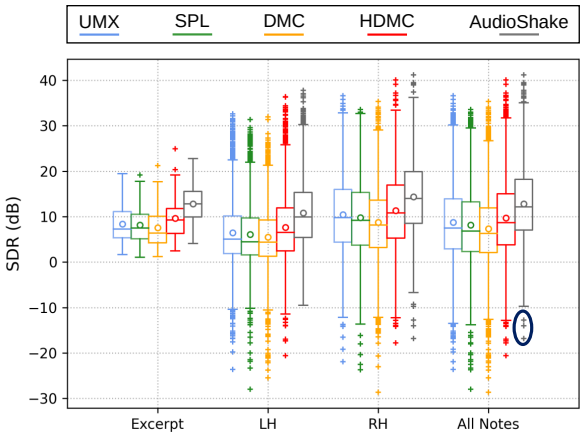

Figure 3: Comparison of different evaluation methodologies for the piano case using boxplots. The three outliers for AudioShake, indicated by the black oval, are shown in Figure 8.

4.1 Global Perspective

To gain an initial understanding of the overall performance of the five MSS systems, Table 3 presents the $\mathrm{SDR}_{\mathrm{local}}$ values averaged across the 81 PCD excerpts for both separated piano tracks and orchestra tracks. For instance, in the piano case, DMC achieves the lowest $\mathrm{SDR}_{\mathrm{local}}$ value at 7.59, while HDMC shows a higher value of 9.61, and AudioShake outperforms all other models with a value of 12.82. Similar trends are evident in the separated orchestra case, although all values are notably lower compared to the piano case. Similar tendencies have been reported in [28].

In the subsequent finer-grained evaluation, we employ notewise evaluation metrics. Since we have the required

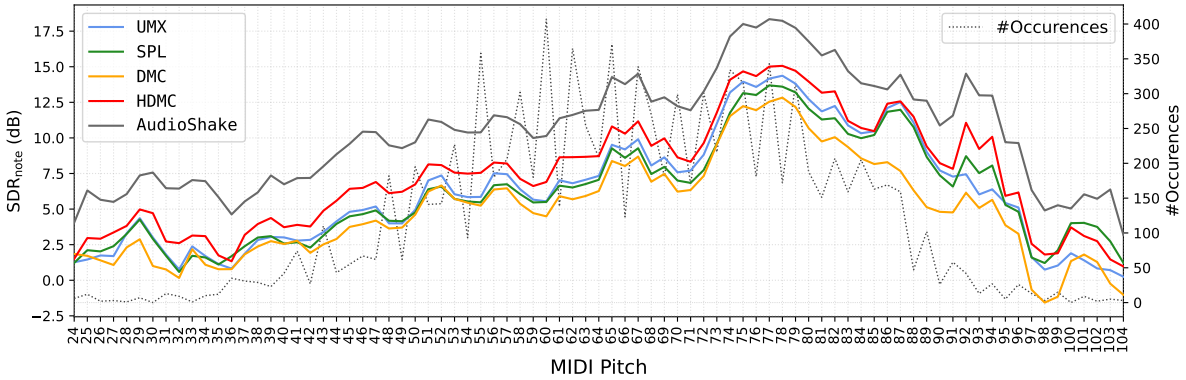

Figure 4: $\mathrm{SDR}_{\mathrm{note}}$ values aggregated by pitch (specified by MIDI note number) shown for five MSS systems.

symbolic score information for the score-based NMF decomposition exclusively for the piano tracks, we confine our analysis to the piano case. [5] Extending the evaluation methodology for the five MSS systems, Figure 3 shows boxplots that indicate the median, first quartile, third quartile, and outliers of differently computed SDR values. The first group of boxplots (Excerpt) provides the $\mathrm{SDR}_{\mathrm{local}}$ values computed as in Table 3. The second (LH) and third (RH) groups show the $\mathrm{SDR}_{\mathrm{note}}$ values for the left-hand and right-hand notes, respectively, and the last group (All Notes) shows the $\mathrm{SDR}_{\mathrm{note}}$ values for all individual notes.

While the general trends for the five MSS systems are similar to those shown in Table 3, the different evaluation methodologies provide additional information. Firstly, being based on notewise aggregation, outliers in the $\mathrm{SDR}_{\mathrm{note}}$-based boxplots offer explicit cues worth further investigation. For instance, outliers such as the three indicated by the black oval in Figure 3 yield interesting examples for musically complex passages as further explored in Section 4.3. The boxplots in Figure 3 also facilitate a comparison of $\mathrm{SDR}_{\mathrm{note}}$ values between the LH and RH notes. Notably, for all MSS systems, a better separation quality can be observed for the right hand compared to the left hand, with a difference of approximately $5\,\mathrm{dB}$. Drawing from these observations, one can formulate various hypotheses regarding the relationship between source separation quality and pitch or musical complexity, as we detail in the subsequent sections. Please visit our demo webpage to find audio examples separated by five MSS models. [6]

4.2 Pitchwise Evaluation

Considering that RH typically contains higher notes than LH, one may conjecture that source separation quality depends on the pitch of the played notes. To test this hypothesis, Figure 4 provides an overview of the $\mathrm{SDR}_{\mathrm{note}}$ values aggregated by pitch (specified by MIDI note number). While the overall trend regarding the MSS systems' performances remains the same (`AudioShake` performing best, `DMC` worst, and `HDMC` being in between), the pitch-dependent $\mathrm{SDR}_{\mathrm{note}}$ values indicate that, overall, source separation quality tends to increase for higher pitch numbers, with the highest values in the pitch range 74–80.

However, such trends, and drawing conclusions from them, need to be taken with care. For example, the curves in Figure 4 may indicate that source separation becomes more difficult for very high pitches in the range 96–104. However, these numbers lack statistical significance due to the limited occurrence (indicated by the dotted line). Also, one may assume that such pitches may rarely occur in the training material used for training the MSS systems, thus leading to poor generalizations on the test set.

4.3 Excerptwise Evaluation

Rather than source separation quality solely being a matter of pitch height, there may be other confounding factors underlying the trend. An alternative hypothesis could be that the LH (or lower-pitched) piano notes are more interwoven with the orchestral track, while the RH (or higher-pitched) piano notes stand out and can be better isolated by MSS systems. To explore aspects of musical complexity, we present in Figure 5 $\mathrm{SDR}_{\mathrm{note}}$ values aggregated by excerpt (specified by PCD ID), this time focusing on the results for the two best-performing MSS systems, `HDMC` and `AudioShake`. Sorting the excerpts, e. g., based on decreasing mean values concerning `AudioShake`, facilitates the identification of challenging excerpts, which are depicted toward the right side of the plot.

Guided by the plot in Figure 5, let us consider some concrete examples. Examining the top three excerpts (PCD IDs 045, 042, and 024), a manual inspection reveals that these excerpts share a common characteristic of relatively low musical complexity, consisting of slower passages drawn from the second movements of piano concertos by Beethoven and Mozart. For such passages, both MSS systems achieve a good separation quality.

Next, let us examine the excerpt with the lowest $\mathrm{SDR}_{\mathrm{note}}$ value. This excerpt has PCD ID 076 and corresponds to measures 18–24 of the first movement of Tchaikovsky's Piano Concerto Op. 23, as shown in Figure 6. Evidently, this passage exhibits a high musical complexity, with both piano and orchestra playing numerous

[5] For the orchestra, we generated only piano-reduced scores due to the considerable effort required for full scores. Additionally, automated synchronization and decomposition approaches present greater challenges for orchestral music compared to piano, extending beyond the scope of the case study presented in this paper.

[6] `www.audiolabs-erlangen.de/resources/MIR/2024-ISMIR-PianoSepEval`

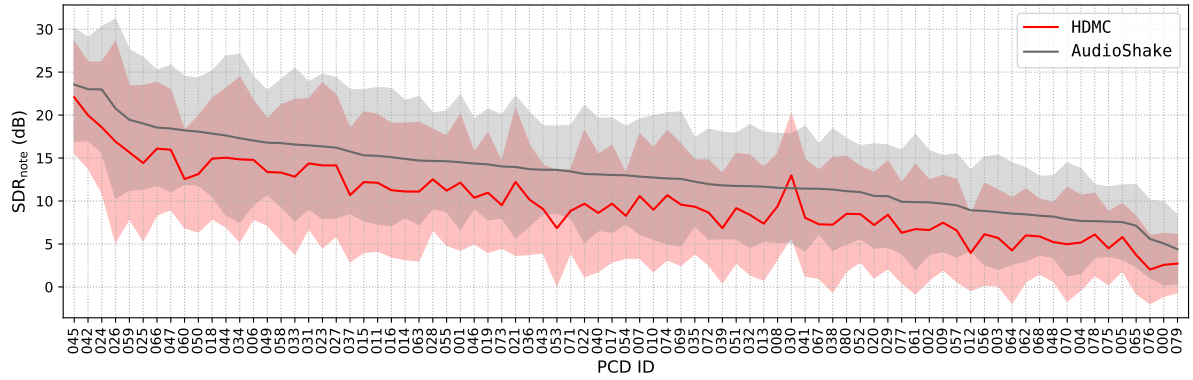

Figure 5: SDR_{note} values aggregated by excerpt (specified by PCD ID) shown for the two best-performing MSS systems, `HDMC` and `AudioShake`. The mean (solid line) and standard deviations (filled regions) are indicated. The excerpts are sorted based on decreasing mean values with regard to `AudioShake`.

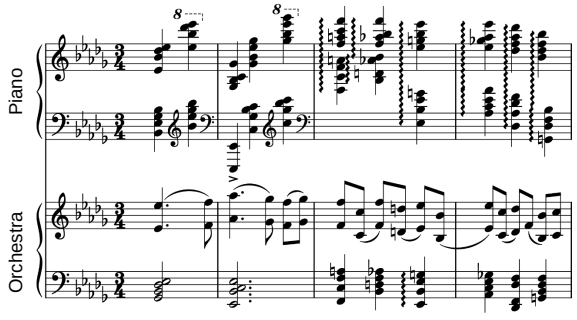

Figure 6: Excerpt with PCD ID 079: Tchaikovsky's Piano Concerto Op. 23, measures 18–24 of the first movement (only four measures are shown here).

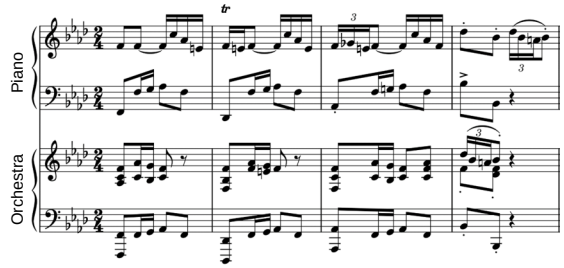

Figure 7: Excerpt with PCD ID 000: Bach's Piano Concerto BWV 1056, measures 1–8 of the first movement (only four measures are shown here).

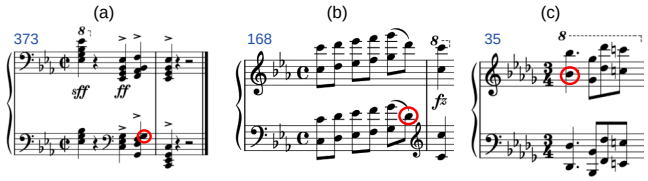

Figure 8: Musical context within the piano scores for the three notewise outliers marked in Figure 3 (here indicated by the red circles). **(a)** PCD ID: 052. **(b)** PCD ID: 061. **(c)** PCD ID: 077.

notes within a wide pitch range. Particularly notable are the fortissimo and broken chords in the piano part, which strongly interfere with the full orchestral sound, not to mention the effects resulting from the application of the sustain pedal. As a second concrete example, let us have a closer look at the excerpt with PCD ID 000, also yielding a low SDR_{note} value. This excerpt corresponds to the first measures of Bach's Piano Concerto BWV 1056, where the piano and orchestra play many notes in unison (see Figure 7). This scenario represents one of the most challenging situations for source separation models to deal with [36, 37].

Finally, we revisit the boxplots shown in Figure 3, where we marked three outliers indicating problematic notewise sound events with low SDR values, poorly separated by `AudioShake`. Figure 8 provides the musical context within the piano scores where these notes occur. A common feature in these examples, which is also typical in piano music in general, is the simultaneous playing of two notes that belong to the same pitch class, contributing to a rich and complex sound texture. Obviously, such instances are difficult for any MSS system, as well as the NMF algorithm to handle.

Overall, these examples show that while MSS systems like `AudioShake` and `HDMC` are capable of achieving impressive separation quality, their efficacy is highly influenced by the intrinsic characteristics of the musical pieces.

5. CONCLUSION

In this paper, we have considered a novel evaluation methodology that compares separated sounds with reference sounds on a notewise basis rather than at the excerpt level. For the challenging piano concerto scenario and employing five MSS systems, we applied this methodology in a case study focusing on the separated piano tracks. This allowed us to gain insights into the separation quality and the complexity of the underlying music. While our focus has been on the piano case, future work may involve evaluating other orchestral instruments and guitars. This could pose additional challenges not only for source separation itself but also for automated synchronization and decomposition approaches. On a meta-level, we hope that our hierarchical discussion, assuming different perspectives, also showcased the potential of musically informed evaluation methodologies, providing a basis for interdisciplinary dialogue between engineering and music experts.

Acknowledgements: This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Grant No. 328416299 (DFG MU 2686/10-2). The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer Institute for Integrated Circuits IIS.

6. REFERENCES

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

[2] F. 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, 2019. [Online]. Available: https://doi.org/10.21105/joss.01667

[3] 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 (JOSS)*, vol. 5, no. 50, p. 2154, 2020. [Online]. Available: https://doi.org/10.21105/joss.02154

[4] A. Défossez, "Hybrid spectrogram and waveform source separation," in *Proceedings of the ISMIR 2021 Workshop on Music Source Separation*, Online, 2021.

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

[6] S. Sarkar, E. Benetos, and M. Sandler, "EnsembleSet: A new high quality dataset for chamber ensemble separation," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Bengaluru, India, 2022, pp. 625–632.

[7] S. Sarkar, L. Thorpe, E. Benetos, and M. Sandler, "Leveraging synthetic data for improving chamber ensemble separation," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2023, pp. 1–5.

[8] S. Uhlich, M. Porcu, F. Giron, M. Enenkl, T. Kemp, N. Takahashi, and Y. Mitsufuji, "Improving music source separation based on deep neural networks through data augmentation and network blending," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, New Orleans, Louisiana, USA, March 2017, pp. 261–265.

[9] H. Kim, J. Park, T. Kwon, D. Jeong, and J. Nam, "A study of audio mixing methods for piano transcription in violin-piano ensembles," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Rhodes Island, Greece, 2023, pp. 1–5.

[10] Y. Özer and M. Müller, "Source separation of piano concertos with test-time adaptation," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Bengaluru, India, 2022, pp. 493–500.

[11] F. Stöter, A. Liutkus, and N. Ito, "The 2018 Signal Separation Evaluation Campaign," in *Proceedings of the International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, ser. Lecture Notes in Computer Science, vol. 10891. Springer, 2018, pp. 293–305.

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

[13] 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.

[14] E. Cano, D. FitzGerald, and K. Brandenburg, "Evaluation of quality of sound source separation algorithms: Human perception vs quantitative metrics," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, 2016, pp. 1758–1762.

[15] M. Torcoli, T. Kastner, and J. Herre, "Objective measures of perceptual audio quality reviewed: An evaluation of their application domain dependence," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1530–1541, 2021.

[16] S. Ewert and M. Müller, "Using score-informed constraints for NMF-based source separation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Kyoto, Japan, 2012, pp. 129–132.

[17] Z. Chen, T. Yoshioka, L. Lu, T. Zhou, Z. Meng, Y. Luo, J. Wu, X. Xiao, and J. Li, "Continuous speech separation: Dataset and analysis," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020, pp. 7284–7288.

[18] Y. Özer, S. Schwär, V. Arifi-Müller, J. Lawrence, E. Sen, and M. Müller, "Piano Concerto Dataset (PCD): A multitrack dataset of piano concertos," *Transactions of the International Society for Music Information Retrieval (TISMIR)*, vol. 6, no. 1, pp. 75–88, 2023.

[19] S. Ewert, M. Müller, and P. Grosche, "High resolution audio synchronization using chroma onset features," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Taipei, Taiwan, 2009, pp. 1869–1872.

[20] M. Müller, Y. Özer, M. Krause, T. Prätzlich, and J. Driedger, "Sync Toolbox: A Python package for efficient, robust, and accurate music synchronization," *Journal of Open Source Software (JOSS)*, vol. 6, no. 64, pp. 3434:1–4, 2021.

[21] 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

[22] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[23] D. Stoller, S. Ewert, and S. Dixon, "Wave-U-net: A multi-scale neural network for end-to-end audio source separation," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Paris, France, 2018, pp. 334–340.

[24] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1830–1840, 2010.

[25] A. Liutkus and R. Badeau, "Generalized Wiener filtering with fractional power spectrograms," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, April 2015, pp. 266–270.

[26] M. Kim, W. Choi, J. Chung, D. Lee, and S. Jung, "KUIELab-MDX-Net: A two-stream neural network for music demixing," in *Proceedings of the ISMIR 2021 Workshop on Music Source Separation*, Online, 2021.

[27] S. Rouard, F. Massa, and A. Défossez, "Hybrid transformers for music source separation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Rhodes Island, Greece, 2023.

[28] Y. Özer and M. Müller, "Source separation of piano concertos using musically-motivated augmentation techniques," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 32, pp. 1214–1225, 2024.

[29] C. Cannam, C. Landone, and M. B. Sandler, "Sonic Visualiser: An open source application for viewing, analysing, and annotating music audio files," in *Proceedings of the International Conference on Multimedia*, Florence, Italy, 2010, pp. 1467–1468.

[30] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proceedings of the Neural Information Processing Systems (NIPS)*, Denver, Colorado, USA, November 2000, pp. 556–562.

[31] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2003, pp. 177–180.

[32] S. A. Raczynski, N. Ono, and S. Sagayama, "Multi-pitch analysis with harmonic nonnegative matrix approximation," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Vienna, Austria, September 2007, pp. 381–386.

[33] S. Ewert, B. Pardo, M. Müller, and M. Plumbley, "Score-informed source separation for musical audio recordings: An overview," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 116–124, April 2014.

[34] J. Driedger, H. Grohganz, T. Prätzlich, S. Ewert, and M. Müller, "Score-informed audio decomposition and applications," in *Proceedings of the ACM International Conference on Multimedia (ACM-MM)*, Barcelona, Spain, 2013, pp. 541–544.

[35] G. Fabbro, S. Uhlich, C.-H. Lai, W. Choi, M. Martínez-Ramírez, W. Liao, I. Gadelha, G. Ramos, E. Hsu, H. Rodrigues, F. Stöter, A. Défossez, Y. Luo, J. Yu, D. Chakraborty, S. Mohanty, R. Solovyev, A. Stempkovskiy, T. Habruseva, N. Goswami, T. Harada, M. Kim, J. H. Lee, Y. Dong, X. Zhang, J. Liu, and Y. Mitsufuji, "The sound demixing challenge 2023 – music demixing track," *arXiv*, 2024.

[36] J. J. Burred, "From sparse models to timbre learning: New methods for musical source separation," Ph.D. dissertation, Technische Universität Berlin, Berlin, Germany, 2009.

[37] C.-B. Jeon, H. Moon, K. Choi, B. S. Chon, and K. Lee, "Medleyvox: An evaluation dataset for multiple singing voices separation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2023, pp. 1–5.