A Computational Approach for Creating orchestra tracks from Piano Concerto Recordings

Yigitcan Özer, Meinard Müller
International Audio Laboratories Erlangen, Germany
{yigitcan.oezer, meinard.mueller}@audiolabs-erlangen.de

Introduction

The piano concerto, composed for a pianist accompanied by an orchestra, is a genre of great importance in Western classical music. Even though most pianists practice piano concertos (also as an essential part of their piano education) in their careers, only first-class pianists have the opportunity to actually play with an orchestra. In this contribution, we propose a computational pipeline that allows pianists of any level to create their own mixes with an orchestra track coming from an existing recording. In particular, this pipeline consists of four components using techniques from music information retrieval (MIR) (see Figure 2). First, starting with a complete piano concerto recording, we apply data-driven source separation techniques to separate the piano and the orchestra. Second, we alleviate separation artifacts (e.g., musical noise) in a post-processing step. Third, we use music synchronization techniques to temporally align the separated orchestra track with the pianist's own recording. Finally, we apply time-scale modification to warp the orchestra track and create the final mix. While introducing a novel dataset used for training and testing our overall procedure, we discuss the various MIR techniques involved.

Source Separation of Piano Concertos

As a first step of the pipeline, our goal is to separate piano and orchestra tracks in a music recording using music source separation (MSS). Being an essential task in music information retrieval (MIR), MSS seeks to recover individual musical sources in audio recordings. Generally, a musical source can refer to singing, an instrument, or an entire group of instruments providing an accompaniment [1]. Here, we consider the separation of piano concertos into piano and orchestra tracks, which can be regarded as a lead-accompaniment separation task [8].

MSS proves to be a challenging task in music processing due to the non-stationary spectro–temporal characteristics of musical signals, as well as their high correlation in both time and frequency. In the last years, deep neural networks (DNNs) have led to substantial improvements in separating musical sources. One disadvantage of data-driven deep models is their need for a large training dataset, which in the case of MSS consists of multitrack recordings with (isolated) individual sources or stems. Most of the open-source datasets involving isolated stems are limited to popular music, e.g., MUSDB18 [9]. However, professionally produced multitrack recordings are rare for Western classical music.

Figure 1: While practicing piano concertos is a crucial aspect of a pianist's education, it is only the first-class pianists who have the chance to perform alongside an orchestra. In this contribution, we propose a computational approach to create orchestra tracks for piano concertos.

For training deep MSS models, generating random mixes of solo instrument recordings may improve the separation quality [11, 13]. In case multitrack recordings are not available, random mixing for data generation and augmentation has opened up new paths for separating instrument mixtures. In this paper, we closely follow our previous work [6] to address the separation of existing piano concerto recordings into piano and orchestra tracks. Our MSS model is trained using an artificial training dataset through randomly mixing samples from the solo piano repertoire (e.g., piano sonatas, mazurkas, etc.) and orchestral pieces without piano (e.g., symphonies) to simulate piano concertos. As an example, Figure 3 shows the separation of an excerpt from the first movement of Piano Concerto in D minor (KV 466) by Wolfgang Amadeus Mozart.

While random mixes cannot simulate the harmonic and rhythmic relationships between various instruments in a real recording, it guides the model to distinguish timbral characteristics of the constituent musical sources. However, the acoustic properties of recordings (including reverberation, and background noise) play an essential role when upmixing and separating different musical tracks. For instance, in the case of poor recording conditions, (e.g., historical recordings) the properties of the test data may not be reflected well in the training set (as known as *inductive bias*), thus leading to a poor separation quality. Finetuning a pre-trained MSS model in the testing phase using a few samples drawn from the test

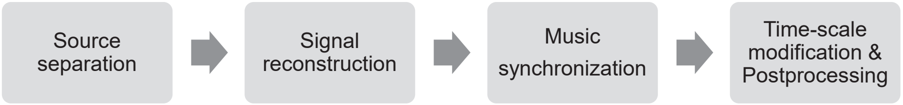

Figure 2: The proposed pipeline for creating orchestra tracks of piano concertos.

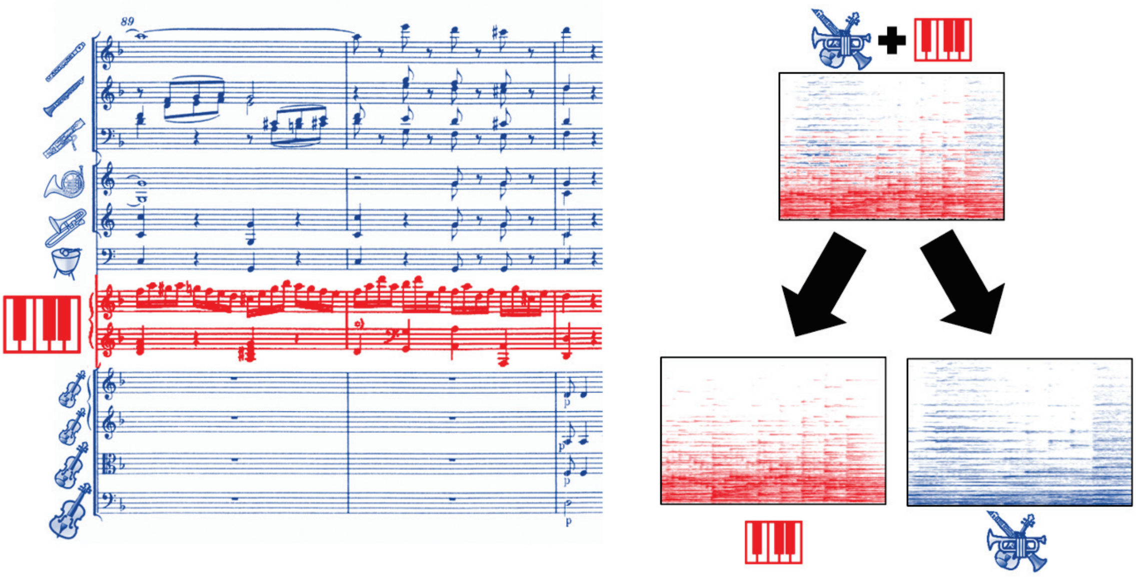

Figure 3: An excerpt from the first movement of the Piano Concerto in D minor (KV466) by Wolfgang Amadeus Mozart. The spectral-based MSS model estimates the magnitude spectrograms of the piano (red) and orchestra tracks (blue). (Figure taken from [6].)

data (also called *test-time adaptation (TTA)* [12]) can improve the separation quality by capturing the specific acoustic features found in a music recording. Depending on the period in which a piano concerto was composed, these compositions often comprise long piano-only (e.g., in the cadenza) and orchestra-only parts (e.g., in the exposition, also called *opening ritornello*). Using these sections, one can create artificial mixes which are extracted from the audio material of the given test item. As a result, the mixes share the same recording conditions as the test data. For further details about the improvement of qualitative and subjective separation quality via TTA, we refer to [6].

Signal Reconstruction

Separated sources need to be clean and acoustically satisfactory to achieve a high-quality orchestra track for the pianists. However, MSS models may introduce undesired or distorted sounds that are introduced during the process of isolating individual musical sources from a mixture. These artifacts can manifest in various ways, such as residual sounds from other sources, musical noise, phase cancellation, spectral smearing, and unnatural-sounding audio. In our current pipeline, we use softmasking to obtain the waveforms from learned magnitude spectrograms, which may lead to phase inconsistencies in the reconstructed piano and orchestra tracks (see Figure 4).

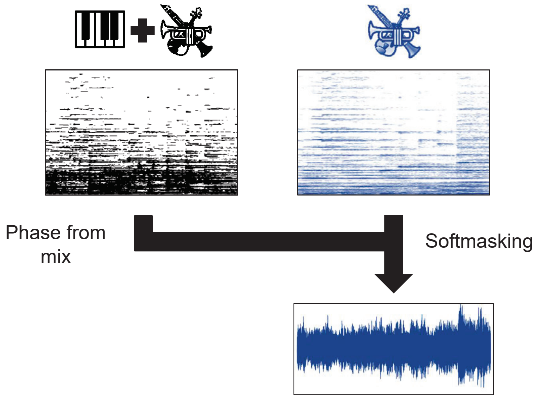

Figure 4: Signal reconstruction from the predicted magnitude spectrogram to audio waveform via softmasking.

In future work, we aim to enhance the non-optimal separation results as the second step of our pipeline. To this end, we are planning to investigate waveform-based MSS models (e.g., [2, 10]), which are able to capture fine temporal details in the input signal that are important for separating sources with fast transients, such as piano onsets. Furthermore, we consider using Generative Adversarial Networks (GANs) as post-processing following source separation. Following the approach by SEGAN [7], we can improve the quality of the separated sources and alleviate the artifacts, such as musical noise and interference from other instruments.

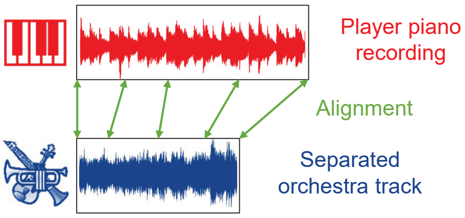

Figure 5: The alignment between the played piano recording and separated orchestra track is computed using music synchronization techniques.

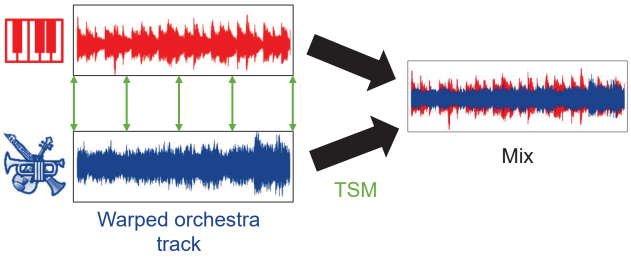

Figure 6: The piano recording by the pianist (red) is mixed with the warped orchestra track (blue) using time-scale modification (TSM).

Music Synchronization

The first two steps of our pipeline allow pianists to select a piano concerto recording and extract the orchestra track to play along with. However, this is insufficient for a pleasant user experience since classical music interpretations can vary greatly in tempo and dynamics. In particular, the performers' global or local tempo choices make their interpretations unique and enrich their performances. In a real-life recording process, the pianist and conductor interact for optimal synchronization and cohesion between the piano and orchestra. In our scenario, however, playing along with a pre-recorded accompaniment would be challenging for the performer.

To address this issue, we propose the following solution. Pianists play and record the piano part of their desired piano concerto freely. Then, we align the recorded piano by the pianists with the separated orchestra track, which comes from the original piano concerto recording (see Figure 5). For synchronization, we use the open-source Python package *Sync Toolbox* [5][1], which provides all components needed to realize a music synchronization pipeline that is robust, efficient, and accurate.

Time-Scale Modification and Postprocessing

As our pipeline's fourth and final step, we use time-scale modification (TSM) to align the separated orchestra track with the piano recording. Using the alignment path acquired from the synchronization algorithm as an input for the TSM Algorithm, we speed up or slow down the separated orchestra track without affecting the frequency content. Figure 6 shows the creation of the final mix using the piano recording by the pianist and separated, clean, and warped orchestra track.

For TSM, we use the approach by Driedger et al. [3], which combines harmonic–percussive source separation (HPSS) and classical algorithms, such as phase vocoder [4] and WSOLA [14]. The TSM approaches are available as open-source Python[2] and MATLAB[3] packages.

Following the TSM, we apply a postprocessing step to create the final mix. To this end, we first apply equalization to the piano recordings to ensure consistent timbral qualities without overcompensating the differences between piano and orchestra. Then, we apply artificial reverberation to both tracks simultaneously using the *FabFilter Pro-R2* algorithmic reverb software to increase the coherence between the piano part and the orchestra track.

Piano Concerto Dataset (PCD)

In [6], we reported on the separation results using random mixes of piano-only and orchestra-only parts sampled from publicly-available piano concertos as test data. In this scenario, the lack of multitrack recordings made a realistic quantitative evaluation of the MSS model difficult. To enable the subjective and quantitative evaluation of MSS models addressing the separation of piano concertos, we proposed a multitrack dataset: Piano Concerto Dataset (PCD). The dataset comprises a collection of excerpts with separate piano and orchestra tracks from piano concertos ranging from the Baroque to the Post-Romantic era.

For the creation of PCD, we used the backing tracks provided by Music Minus One (MMO)[4] and recorded excerpts from 14 different piano concertos played by five different performers on various instruments in diverse acoustic environments. In this scenario, achieving precise synchronization between the performer and pre-recorded orchestra accompaniments poses a significant challenge. For guiding the pianists to obtain high synchronization accuracy, we incorporated additional click tracks using measure and beat annotations of the orchestral tracks, which are also included in the PCD.

PCD is relevant for various MIR tasks, including music source separation, automatic accompaniment, music synchronization, editing, and upmixing. We released PCD via an interactive web-based interface[5] to provide a convenient access.

[1] `https://github.com/meinardmueller/synctoolbox`
[2] `https://github.com/meinardmueller/libtsm`
[3] `https://www.audiolabs-erlangen.de/resources/MIR/TSMtoolbox/`
[4] `https://www.halleonard.com/series/MMONE`
[5] `https://www.audiolabs-erlangen.de/resources/MIR/PCD/`

Conclusion

In this paper, we proposed a computational approach that will allow pianists of any level to create their own piano concerto mixes using existing recordings. The pipeline consists of four essential tasks in MIR: source separation, signal reconstruction and enhancement of separated sources, music synchronization, and TSM. We also presented the PCD, which constitutes a dataset for various applications in MIR, particularly for quantitative and subjective evaluation of source separation models. Our pipeline offers a concrete and practical application and paves the way for exploring new research questions in various MIR techniques.

References

[1] Cano, E., FitzGerald, D., Liutkus, A., Plumbley, M. D. & Stöter, F.: Musical Source Separation: An Introduction. In: IEEE Signal Processing Magazine (2019), **36**, 1: 31–40.

[2] Défossez, A.: Hybrid Spectrogram and Waveform Source Separation. In: Proceedings of the ISMIR 2021 Workshop on Music Source Separation. Online (2021).

[3] Driedger, J., Müller, M. & Ewert, S.: Improving Time-Scale Modification of Music Signals using Harmonic–Percussive Separation. In: IEEE Signal Processing Letters (2014), **21**, 1: 105–109.

[4] Flanagan, J. L. & Golden, R. M.: Phase Vocoder. In: Bell System Technical Journal (1966), **45**: 1493–1509.

[5] Müller, M., Özer, Y., Krause, M., Prätzlich, T. & Driedger, J.: Sync Toolbox: A Python Package for Efficient, Robust, and Accurate Music Synchronization. In: Journal of Open Source Software (JOSS) (2021), **6**, 64: 3434:1–4.

[6] Özer, Y. & Müller, M.: Source Separation of Piano Concertos with Test-Time Adaptation. In: Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), 493–500. Bengaluru, India (2022).

[7] Pascual, S., Bonafonte, A. & Serrà, J.: SEGAN: Speech Enhancement Generative Adversarial Network. In: Lacerda, F. [Hrsg.]: Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech), 3642–3646. ISCA, Stockholm, Sweden (2017).

[8] Rafii, Z., Liutkus, A., Stöter, F., Mimilakis, S. I., FitzGerald, D. & Pardo, B.: An Overview of Lead and Accompaniment Separation in Music. In: IEEE/ACM Transactions on Audio, Speech, and Language Processing (2018), **26**, 8: 1307–1335.

[9] Rafii, Z., Liutkus, A., Stöter, F.-R., Mimilakis, S. I. & Bittner, R. (2017): The MUSDB18 Corpus for Music Separation. URL https://doi.org/10.5281/zenodo.1117372.

[10] Stoller, D., Ewert, S. & Dixon, S.: 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), 334–340. Paris, France (2018).

[11] Stöter, F., Uhlich, S., Liutkus, A. & Mitsufuji, Y.: Open-Unmix – A Reference Implementation for Music Source Separation. In: Journal of Open Source Software (2019), **4**, 41. URL https://doi.org/10.21105/joss.01667.

[12] Sun, Y., Wang, X., Zhang, L., Miller, J., Hardt, M. & Efros, A. A.: Test-Time Training with Self-Supervision for Generalization under Distribution Shifts. In: Proceedings of the International Conference on Machine Learning (ICML) (2020).

[13] Uhlich, S., Porcu, M., Giron, F., Enenkl, M., Kemp, T., Takahashi, N. & Mitsufuji, Y.: 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), 261–265. New Orleans, Louisiana, USA (2017).

[14] Verhelst, W. & Roelands, M.: An overlap–add technique based on waveform similarity (WSOLA) for high quality time–scale modification of speech. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Minneapolis, USA (1993).