Source Separation of Piano Music Recordings

Quellentrennung von Klaviermusikaufnahmen

Dissertation

Der Technischen Fakultät

der Friedrich-Alexander-Universität Erlangen-Nürnberg

zur

Erlangung des Doktorgrades

Doktor der Ingenieurwissenschaften (Dr.-Ing.)

vorgelegt von

Yigitcan Özer

aus

İzmir / Türkei

Als Dissertation genehmigt
von der Technischen Fakultät
der Friedrich-Alexander-Universität Erlangen-Nürnberg

Tag der mündlichen Prüfung:	12. Juni 2024
1. Gutachter:	Prof. Dr. rer. nat. Meinard Müller
2. Gutachter:	Prof. Dr. Gerhard Widmer

Abstract

In the field of Music Information Retrieval (MIR), the decomposition of a music recording into its constituent sound sources, referred to as source separation, is relevant to a broad spectrum of tasks and applications, such as supporting music analysis, creating karaoke systems, aiding in music production, and facilitating music transcription. In this thesis, we address the novel and previously unexplored source separation task of decomposing piano concerto recordings into individual piano and orchestral tracks. As a genre of central importance in Western classical music, practicing and performing piano concertos is an essential aspect of a pianist's education and professional journey. However, only first-class pianists have the opportunity to actually perform with an orchestra. Addressing the lack of orchestral accompaniments for pianists of any level, we aim to extract orchestral tracks from piano concerto recordings. As one main contribution of this thesis, we adapt deep-learning-based source separation techniques, initially designed for the separation of popular music recordings or speech signals. In particular, we address the challenge of higher spectro–temporal correlations between piano and orchestra compared to popular music tracks, and the lack of multitrack datasets for training, by introducing musically motivated data augmentation approaches. Another main contribution of this thesis is the creation of a multitrack dataset of piano concertos. This dataset encompasses a collection of excerpts with separate orchestral and piano tracks, performed by both professional and amateur pianists. To create these temporally synchronized multitrack recordings, we used pre-existing orchestral accompaniments provided by Music Minus One (MMO) and applied semi-automatic techniques, such as beat tracking and music synchronization. The broad scope of our dataset not only serves as a valuable resource for both quantitative and subjective evaluation of source separation models but also opens up various possibilities for other MIR applications, including score following, downbeat estimation, and music synchronization. As a third main contribution, we split the separated piano tracks into notewise events using score-informed nonnegative matrix factorization (NMF). In particular, we apply this audio decomposition technique for evaluating source separation results of piano tracks introducing a notewise signal-to-distortion ratio (SDR) measure to gain deeper insights into various source separation artifacts. Overall, this thesis not only addresses a novel challenge in MIR but also enhances the way pianists can interact with classical music performances.

Zusammenfassung

Im Forschungsgebiet *Music Information Retrieval (MIR)* ist die Zerlegung einer Musikaufnahme in ihre einzelnen Klangquellen, die so genannte Quellentrennung, für eine Vielzahl von Anwendungen relevant, beispielsweise zur Unterstützung der Musikanalyse oder bei der Musikproduktion, zur Erleichterung von Musiktranskriptionen oder für die Entwicklung von Karaoke-Systemen. Diese Dissertation befasst sich mit der neuartigen und bislang unerforschten Aufgabenstellung der Zerlegung von Klavierkonzertaufnahmen in einzelne Klavier- und Orchesterspuren. Als eine Gattung von zentraler Bedeutung in der westlichen klassischen Musik ist das Einstudieren und Aufführen von Klavierkonzerten ein wesentlicher Aspekt der Ausbildung und des beruflichen Werdegangs eines Pianisten. Allerdings verfügen nur erstklassige Pianisten über die Möglichkeit, gemeinsam mit einem Orchester aufzutreten. Um für Pianisten aller Niveaus die Möglichkeit einer Orchesterbegleitung zu schaffen, ist unser Ziel, Orchesterspuren aus Klavierkonzertaufnahmen zu extrahieren. Ein Hauptbeitrag dieser Arbeit ist die Anpassung der Techniken des maschinellen Lernens zur Quellentrennung, die ursprünglich für die Trennung von Aufnahmen populärer Musik oder Sprachsignale konzipiert wurden. Wir gehen die Herausforderungen von im Vergleich zu Popularmusik oft höherer spektro-temporaler Korrelationen zwischen dem Klavier und Orchester und dem Mangel an mehrspurigen Datensätzen für das Training an, indem wir musikalisch motivierte Ansätze zur Datenaugmentierung einführen. Ein weiterer Hauptbeitrag dieser Arbeit ist die Erstellung eines Mehrspur-Datensatzes von Klavierkonzerten. Der Datensatz umfasst eine Sammlung von Ausschnitten mit separaten Orchester- und Klavierspuren, gespielt von sowohl professionellen als auch Amateurpianisten. Um diese zeitlich synchronisierten Mehrspuraufnahmen zu erstellen, verwendeten wir vorhandene Orchesterbegleitungen, die von Music Minus One (MMO) bereitgestellt wurden, und wandten halbautomatische Techniken, wie Beat-Tracking und Musiksynchonisierung, an. Der breite Umfang dieses Datensatzes dient nicht nur als wertvolle Grundlage für die quantitative und subjektive Evaluierung von Verfahren zur Quellentrennung, sondern eröffnet auch verschiedene Möglichkeiten für andere MIR-Anwendungen, einschließlich Partiturverfolgung, Downbeat-Schätzung und Musiksynchronisation. Als dritten Hauptbeitrag zerlegen wir die getrennten Klavierspuren in notenweise Events unter Verwendung von partiturinformierter nichtnegativer Matrix Faktorisierung (NMF). Insbesondere wenden wir diese Technik der Audiozerlegung zur Bewertung von Quellentrennungsergebnissen von Klavierspuren an, indem wir ein notenweises Signal-to-Distortion-Ratio (SDR) einführen, um tiefere Einblicke in die verschiedenen Quellentrennungsartefakte zu gewinnen. Zusammenfassend präsentiert diese Arbeit nicht nur eine neue Fragestellung auf dem Gebiet des MIRs, sondern verbessert auch die Art und Weise, wie Pianisten mit Aufnahmen im Bereich der klassischen Musikaufführung interagieren können.

Contents

1 Introduction

Music is a language of artistic expression, creating beauty and resonating with human emotions while transcending cultural and linguistic boundaries to unite listeners in a shared experience of sound. The digital revolution in musical distribution and storage, for instance through mobile devices and the internet, has made music a ubiquitous part of daily life for billions of people worldwide. However, there still remains a considerable potential to enhance human interaction with musical content. For example, humans possess the innate ability to focus on specific instruments or voices, despite the complex overlay of acoustic source signals from various instruments. The development of techniques enabling the interaction with constituting audio components within a music recording, such as isolating the vocals, instruments, or instrument groups, opens up numerous possibilities for a variety of applications.

In the field of Music Information Retrieval (MIR), the decomposition of a music recording into its constituent sound sources is commonly referred to as *source separation*. Within the context of music, a source might refer to a melody, a bass line, a drum track, a general instrumental voice, a group of instruments, or even single note events played on these instruments. Music source separation (MSS) aims at decomposing a musical mixture into its constituent sources, ideally, as if they were played in an isolated fashion.

Whereas research on MSS is mostly limited to separating popular music recordings into vocals, drums, bass, and other sources, we address in this thesis the novel and rarely considered source separation task of decomposing piano concerto recordings into separate piano and orchestral tracks. These compositions constitute a genre of great importance in Western classical music, renowned for their rich, dynamic sound and distinctive, contrasting musical elements. With a substantial repertoire throughout music history, classical music archives are rich in historical, public-domain recordings of piano concertos, which can be useful for numerous applications in MIR, including source separation [42, 83, 185], audio editing [48, 86], upmixing [117, 180], music alignment [57, 146], automatic accompaniment [28, 37, 39, 203], audio decomposition [47, 58], and automatic transcription [7, 55, 100, 127].

Even though practicing piano concertos is a fundamental aspect of a pianist's education and career, only first-class pianists have the opportunity to actually perform with an orchestra. To address the lack of orchestral accompaniments for pianists of any level, we aim to extract orchestral tracks from public-domain recordings of piano concertos, as visualized in Figure 1.1. From a technical perspective, our goal is related to MSS, which is the task of recovering individual musical sources in audio recordings.

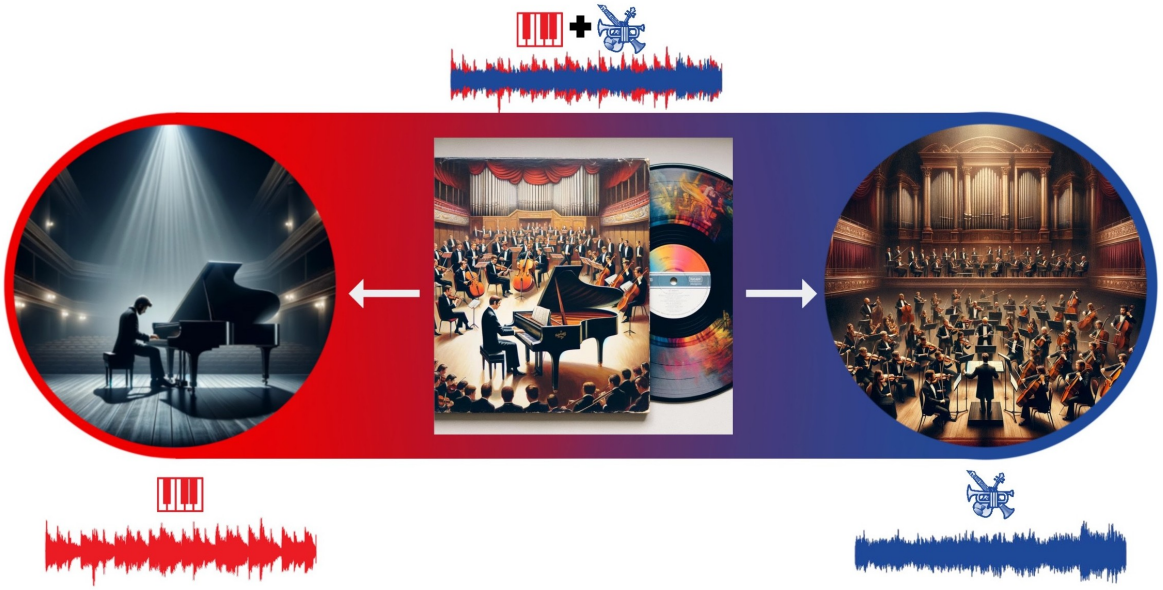

Figure 1.1: In this thesis, the focus is the separation of piano music recordings, particularly piano concertos. Our primary goal is to decompose public-domain recordings of piano concertos into separate piano and orchestral tracks. These separated orchestral accompaniments can then be synchronized with piano solo recordings by any performer, allowing them to create their own unique mixes (the images involved are partly created with the assistance of DALL-E 3).

As our main technical tool for source separation, we employ data-driven methods, particularly deep learning (DL) techniques. DL-based approaches have emerged as the preeminent paradigm in audio processing and have led to a breakthrough in MSS [42, 83, 114, 122, 185]. In this thesis, we adapt several existing DL techniques, mainly used for the separation of popular music recordings or speech signals. A key challenge 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 containing isolated stems are limited to popular music. However, professionally produced multitrack recordings are rare for Western classical music. To circumvent the problem of missing multitrack training samples, we consider the generation of artificial training examples by random mixing and introduce musically motivated data augmentation approaches to enhance the separation performance.

Furthermore, for a fair quantitative and subjective evaluation of the MSS models, the use of a multitrack dataset of real recordings is essential. Recognizing this issue, we generated the Piano Concerto Dataset (PCD), which involves a collection of excerpts with separate piano and orchestral tracks from piano concertos ranging from the Baroque to Post-Romantic era. In particular, using existing backing tracks by the music publisher Music Minus One (MMO), we recorded excerpts from different piano concertos played by five interpreters on various pianos under different acoustic conditions.

Our MSS approaches allow pianists to select a piano concerto recording and extract the orchestra track to play along with. This allows pianists to play and record the piano part of their desired piano concerto freely and then add the separated orchestra track in a post-processing step using alignment techniques in

combination with time-scale modification (TSM) [49]. While introducing a novel dataset used for training and testing our overall approach, we discuss the various MIR techniques involved in this pipeline.

1.1 Structure and Main Contributions of this Thesis

The thesis is structured as follows. We begin with foundational concepts of music audio signal processing that are used throughout this thesis in **Chapter 2**. In particular, we elaborate on audio signals, time–frequency representations, and chroma-based audio features.

Then, we focus on high-resolution music synchronization in **Chapter 3**, which plays a crucial role for dataset curation, audio pre- and post-processing in the subsequent chapters. Our approach employs an efficient implementation of dynamic time warping (DTW) [146] to align different versions of the same musical piece. Furthermore, we incorporate additional information to the synchronization pipeline to increase the temporal accuracy, including onset cues, beat, and downbeat activation functions, in addition to chroma features [57, 135]. Our findings indicate that the integration of a combined version of all three activation functions significantly improves the synchronization accuracy while maintaining the robustness of the chroma-based synchronization approach.

In **Chapter 4**, we address the separation of piano concertos as one main contribution of this thesis, introducing our separation approach. Our initial experiments involve training with artificial mixes which are randomly generated by sections from solo piano pieces (e.g., piano sonatas, mazurkas, etc.) and orchestral works without piano (e.g., symphonies) [131]. For model finetuning, we propose a test-time adaptation (TTA) procedure [107, 188], which exploits random mixtures of the piano-only and orchestra-only parts in the test data to further improve the separation quality. Our experiments demonstrate that exploiting the compositional structures of piano concertos through TTA substantially improves the quantitative and subjective evaluation results, both for the piano and orchestra. Note that these first experiments use a synthetically created test dataset for quantitative and subjective evaluation due to the lack of multitrack piano concerto recordings at the time of research.

To address the lack of a realistic multitrack dataset of piano concertos, we introduce the Piano Concerto Dataset (PCD) [136] in **Chapter 5**, which comprises a collection of excerpts with separate piano and orchestral tracks from piano concertos ranging from the Baroque to Post-Romantic era. This dataset, curated using backing tracks from MMO, involves excerpts from 15 concertos performed by five interpreters on various instruments under different acoustic conditions. The broad musical scope of PCD enables various applications for MIR research, particularly for quantitative and subjective evaluation of source separation models.

In **Chapter 6**, we build on the separation approach discussed in **Chapter 4**, investigating spectrogram- and waveform-based MSS approaches, as well as hybrid models operating in both spectrogram and waveform domains. Furthermore, we introduce a novel, musically motivated data augmentation strategy for training

based on artificially generated samples and conduct a thorough analysis of different augmentation methods on DL models using the PCD (**Chapter 5**) for our quantitative and subjective evaluations. In particular, we generate a dataset to simulate unison passages utilizing the synchronization approach from **Chapter 3**. To this end, we use recordings of Beethoven symphonies and their renowned piano transcriptions by Franz Liszt. A key finding from these experiments is that the hybrid model, when trained with a full suite of augmentation techniques, achieves the best source separation performance [132] in view of the quantitative and subjective evaluations.

When evaluating the separation results, commonly used metrics include the signal-to-distortion ratio (SDR), computed over entire excerpts or songs. Departing from this conventional approach, in **Chapter 7**, we introduce a novel evaluation method that decomposes an audio track into musically meaningful sound events and applies the evaluation metric based on these units. To assess piano separation quality, we use a score-informed nonnegative matrix factorization (NMF) approach to decompose the reference and separated 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.

In **Chapter 8**, we use score-informed NMF as a baseline for efficient audio decomposition and extend this strand of research to include nonnegative autoencoders (NAEs) in combination with gradient projection and structured dropout techniques. Conducting experiments based on piano recordings, we compare the decomposition results of NAE-based approaches with those obtained using a variant of score-informed NMF. In this context, we explore various gradient descent methods, employing both fixed and adaptive learning rates, to optimize the encoder and decoder parameters of NAEs.

Finally, we conclude this thesis in **Chapter 9** with a summary and detailed discussion of prospects for future work.

1.2 Publications Related to Ph.D. Thesis

The main chapters of this thesis are based on articles that have previously been published or accepted to appear in peer-reviewed journals and conference proceedings within the fields of audio signal processing and MIR. I am the first author and main contributor to all these publications.

[132] Yigitcan Özer and Meinard Müller. Source separation of piano concertos using musically-motivated augmentation techniques. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 32:1214–1225, 2024. doi: 10.1109/TASLP.2024.3356980

[137] Yigitcan Özer, Hans-Ulrich Berendes, Vlora Arifi-Müller, Fabian-Robert Stöter, and Meinard Müller. Notewise evaluation for music source separation: A case study for separated piano tracks. *Submitted for publication*, 2024

[136] Yigitcan Özer, Simon Schwär, Vlora Arifi-Müller, Jeremy Lawrence, Emre Sen, and Meinard Müller. Piano Concerto Dataset (PCD): A multitrack dataset of piano concertos. *Transactions of the International Society for Music Information Retrieval (TISMIR)*, 6(1):75–88, 2023. doi: 10.5334/tismir.160

[130] Yigitcan Özer and Müller. A computational approach for creating orchestral accompaniments from piano concerto recordings. In *Proceedings of the Deutsche Jahrestagung für Akustik (DAGA)*, pages 1370–1373, Hamburg, Germany, 2023

[131] Yigitcan Özer and Meinard Müller. Source separation of piano concertos with test-time adaptation. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 493–500, Bengaluru, India, 2022

[135] Yigitcan Özer, Matěj Ištvánek, Vlora Arifi-Müller, and Meinard Müller. Using activation functions for improving measure-level audio synchronization. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 749–756, Bengaluru, India, 2022

[134] Yigitcan Özer, Jonathan Hansen, Tim Zunner, and Meinard Müller. Investigating nonnegative autoencoders for efficient audio decomposition. In *Proceedings of the European Signal Processing Conference (EUSIPCO)*, pages 254–258, Belgrade, Serbia, 2022. doi: 10.23919/EUSIPCO55093.2022.9909787

1.3 Additional Publications

Aside from the main articles that make up this thesis, I contributed to the following additional publications in the field of audio and music processing.

[138] Yigitcan Özer, Leo Brütting, Simon Schwär, and Meinard Müller. libsoni: A Python toolbox for sonifying music annotations and feature representations. *Journal of Open Source Software (JOSS)*, 9(96):1–6, 2024. doi: 10.21105/joss.06524

[186] Sebastian Strahl, Yigitcan Özer, Hans-Ulrich Berendes, and Meinard Müller. Hearing your way through music recordings: A text alignment and synthesis approach. *Submitted for publication*, 2024

[198] T. J. Tsai, Kavi Dey, Yigitcan Özer, and Meinard Müller. Customizing piano concerto accompaniments using hybrid dense-sparse dynamic time warping. *Submitted for publication*, 2024

[191] Nazif Can Tamer, Yigitcan Özer, Meinard Müller, and Xavier Serra. High-resolution violin transcription using weak labels. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 223–230, Milan, Italy, 2023

[192] Nazif Can Tamer, Yigitcan Özer, Meinard Müller, and Xavier Serra. TAPE: An end-to-end timbre-aware pitch estimator. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Rhodes Island, Greece, 2023. doi: 10.1109/ICASSP49357.2023.10096762

[133] Yigitcan Özer, Michael Krause, and Meinard Müller. Using the sync toolbox for an experiment on high-resolution music alignment. In *Demos and Late Breaking News of the International Society for Music Information Retrieval Conference (ISMIR)*, Online, 2021. URL `https://archives.ismir.net/ismir2021/latebreaking/000025.pdf`

[126] Meinard Müller, Yigitcan Özer, Michael Krause, Thomas Prätzlich, and Jonathan Driedger. Sync Toolbox: A Python package for efficient, robust, and accurate music synchronization. *Journal of Open Source Software (JOSS)*, 6(64): 3434:1–4, 2021. doi: 10.21105/joss.03434

[74] Prachi Govalkar, Ahmed Mustafa, Nicola Pia, Judith Bauer, Metehan Yurt, Yigitcan Özer, and Christian Dittmar. A lightweight neural TTS system for high-quality German speech synthesis. In *Proceedings of the ITG Conference on Speech Communication*, pages 39–43, 2021

[112] Patricio López-Serrano, Christian Dittmar, Yigitcan Özer, and Meinard Müller. NMF toolbox: Music processing applications of nonnegative matrix factorization. In *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, Birmingham, UK, 2019

1.4 Acknowledgments

In his celebrated text "Letters to a Young Poet", Rilke poses the question *"In the deepest hour of the night, confess to yourself that you would die if you were forbidden to write. And look deep into your heart where it spreads its roots, the answer, and ask yourself, must I write?"*. This question resonates profoundly with me, capturing the essence of my devotion to music, particularly the piano. Driven by this passion, I searched for a harmonious intersection between my dual loves of engineering and music. This is the very story of how I met my Ph.D. supervisor Meinard Müller, and landed in the field of MIR.

First and foremost, I would like to express my gratitude to my supervisor, Meinard, for his unwavering support, guidance, and mentorship throughout the course of my Ph.D. His expertise and insights have been invaluable to my research. His advice was always just an email away, and over time, our professional relationship has evolved into a friendship. Moreover, it has always been a great pleasure to run into Meinard during my evening practice sessions on the grand piano at the institute and share our enthusiasm. Thank you, Meinard, for opening the doors of AudioLabs for me; your acceptance has been both an honor and a defining moment in my career and academic journey.

Secondly, I would like to thank my former manager at Fraunhofer IIS, Christian Dittmar. Working with him was not only rewarding as an engineer but also greatly enjoyable. His ideas have been a constant source of inspiration, and the knowledge I gained under his guidance has been priceless. Thank you, Christian, for your great support and friendship.

Thirdly, I also would like to express my gratitude to Gerhard Widmer, the second reviewer of this thesis, for his valuable time and effort in reviewing my work. Additionally, I am grateful for the opportunity to have visited his lab in Linz in 2015, an experience that offered both inspiration and hospitality.

Joining Fraunhofer IIS in 2019 as a research associate in the Spoken Language Processing Group was a significant milestone for my career. Then, in 2021, I joined AudioLabs Erlangen as a Ph.D. student, which is a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer IIS. Situated between both the university and the Audio & Multimedia division of Fraunhofer IIS, the AudioLabs offer an amazing infrastructure that fosters close scientific collaborations in an ideal setting. I am grateful to Bernhard Grill and Frederik Nagel as representatives for all those who have established this excellent research environment.

During my Ph.D. at AudioLabs, I have had the honor and pleasure of collaborating with exceptionally skilled colleagues. This first includes the members of *GroupMM*. In alphabetical order: Jakob Abeßer, Vlora Arifi-Müller, Judith Bauer, Hans-Ulrich Berendes, Ching-Yu Chiu, Michael Krause, Ben Maman, Peter Meier, Sebastian Rosenzweig, Simon Schwär, Sebastian Strahl, Christof Weiß, Frank Zalkow, and Johannes Zeitler. Besides GroupMM, it has been a privilege to team up in the inspiring atmosphere of AudioLabs and to be surrounded with great colleagues, including the administrative staff, professors, Ph.D. students, and all other colleagues, Ahmad Aloradi, Carlotta Anemüller, Leo Brütting, Srikanth Raj Chetupalli, Pablo Delgado, Bernd Edler, Muhammad Elminshawi, Richard Füg, Ünal Ege Gaznepoğlu, Ning Guo, Philipp Götz, Emanuël A. P. Habets, Tracy Harris, Jürgen Herre, Adrian Herzog, Thorsten Kastner, Kishor Kayyar Lakshminarayana, Jeremy Lawrence, Wolfgang Mack, Frederik Nagel, Nils Peters, Day-See Riechmann, Thomas Robotham, Lorenz Schmidt, Neeraj Kumar Sharma, Fabian-Robert Stöter, Martin Strauß, Matteo Torcoli, Stefan Turowski, Julian Wechsler, Elke Weiland, Nils Werner, Niklas Winter, and many others. Moreover, I thank my former colleagues from Fraunhofer IIS, not only for their support, but also keeping the friendship.

Being a part of the MIR community has been and surely will be a great privilege for me. Throughout my Ph.D., besides meeting numerous inspiring researchers, I had very fruitful scientific collaborations with several colleagues. Special thanks to Matěj Ištvánek from Brno Institute of Technology for both his collaboration and hospitality during my research visit to Brno. I am also grateful to Nazif Can Tamer and Xavier Serra from Universitat Pompeu Fabra (UPF) for guiding me into the realm of strings music! A special thanks goes to TJ Tsai from Harvey Mudd College, for our collaboration on generating piano concerto accompaniments and sharing our enthusiasm! My appreciation extends to Yiğit Aydın, Barış Demirezer, and Tolga Yayalar for fruitful discussions and organizing the recording sessions at Bilkent University, Ankara. Additionally, I thank Juhan Nam, Hyemi Kim, Heyyoon Cho, Jiyun Park, Taegyun Kwon, and Yonghyun Kim from Korea Advanced Institute of Science & Technology (KAIST), for their shared passion and efforts in extending our piano concerto dataset. I also thank Cynthia Liem for the privilege of playing Schubert's Fantasia in F minor in a four-hand performance at the Dagstuhl Seminar 22082 in 2022 – that was a life-time experience for me!

Music has been a driving force behind this thesis, and I must express my gratitude to those who have touched my life musically. Douze points pour Emre Şen, who not only had a great contribution to the dataset we created during my Ph.D., but has also been a supportive friend and a phenomenal piano teacher. Emre, your support and musicianship have been a great source of inspiration! My gratitude extends to my piano teachers, Bianca Bodalia, Berrin Beycan, Gülnara Azizova, Edna Golandsky, Jin Jeon, and Gamze Kırtıl.

My sincere thanks go to my study advisor at Bilkent University, Orhan Arıkan. His door was always open for a conversation, even after my graduation. Beyond academic guidance, his enduring support and persuasive encouragement were crucial toward the path of my Ph.D. journey. I am also deeply grateful to my former roommate Efe Çötelioğlu for his enduring friendship and the spirit of comradeship we shared

over the years. My gratitude also extends to my colleague and friend, Çağdaş Tuna for his support and inspiration during my Ph.D. I also want to thank my other dear friends, who have been a great support during this highly enjoyable but intensive phase of my life.

Last but not least, my deepest gratitude goes to my family – your constant love and support have both been my anchor and my sail, allowing me to pursue my dreams. Thank you, Mom, Dad, Pınar, Kemal, and Defne, for always being there for me. I love you all from the bottom of my heart. Bu tezi size adıyorum.

2 Audio Signal Processing

In this chapter, we introduce key concepts that will be used in all the subsequent sections of the thesis. In particular, we present audio signals in Section 2.1 and cover time–frequency representations, including the discrete Fourier transform (DFT), short-time Fourier transform (STFT), and spectrograms in Section 2.2. We then explore chroma-based audio representations in Section 2.3, which are widely used to tackle several MIR tasks. For a detailed overview, we refer to the textbook by Müller [123], the notation of which is consistently employed throughout this thesis.

2.1 Audio Signals

Sound signals are detected by humans as oscillations in air pressure, which originate from vibrating objects. For example, upon pressing a key on a piano, a hammer strikes one or more strings, which induces vibrations. The resulting change in air pressure at a certain location can be graphically represented through a plot of pressure over time, commonly referred to as the *waveform* of the produced sound. One can mathematically describe a *continuous-time*, or *analog* audio waveform as a function $f : \mathbb{R} \to \mathbb{R}$, mapping a time point $t \in \mathbb{R}$ measured in physical units (e. g., seconds) to an amplitude $f(t) \in \mathbb{R}$, serving as an indicator for the relative strength of sound waves. As an example, Figure 2.1 shows the waveform of an excerpt from Mozart's Piano Concerto No. 21 in C Major, KV. 467, 2$^{\text{nd}}$ movement, which serves as our running example throughout this chapter.

When using digital technology, only a discrete number of parameters can be stored and processed. Therefore, analog audio signals need to be converted into discrete representations. In the process of analog-to-digital conversion, two main steps are involved: *sampling* and *quantization*. Sampling defines the process of exclusively storing a finite set of amplitudes at discrete time positions, whereas quantization refers to the process of mapping these real-valued amplitudes to a finite set of possible amplitudes $\Gamma \subset \mathbb{R}$, such that these values may be represented on digital devices.

In particular, *equidistant sampling* only retains values of an analog signal at time positions, which are integer multiples of a sampling period $T \in \mathbb{R}_{>0}$. Given the analog signal $f \in \mathbb{R} \to \mathbb{R}$, we define a function $x : \mathbb{Z} \to \mathbb{R}$ by setting

$$x(n) = f(n \cdot T), \tag{2.1}$$

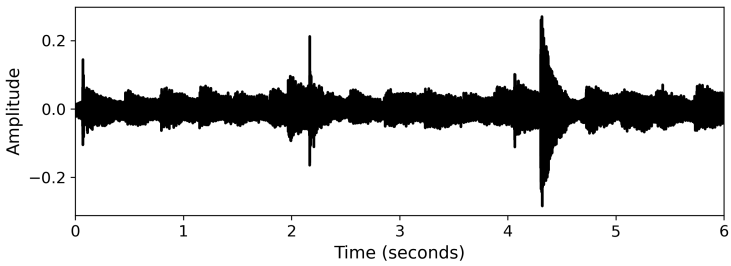

Figure 2.1: Waveform of an excerpt from Mozart's Piano Concerto No. 21 in C Major, KV. 467, 2nd movement, which serves as a running example in this chapter.

where $n \in \mathbb{Z}$ is the time index. Since x is defined at discrete time points, it is termed a *discrete-time* signal. $x(n)$ denotes the *sample* taken at time $t = n \cdot T$. This procedure is also known as *T-sampling*. Its reciprocal yields the *sampling rate* F_s of this process. The sampling rate indicates the number of samples per second and is measured in Hertz (Hz):

$$F_\mathrm{s} = 1/T. \tag{2.2}$$

Sampling is a lossy operation, implying that the original analog signal cannot be perfectly recovered from its sampled version. For example, the industry standard for CD recordings employs a sampling rate of $F_\mathrm{s} = 44.1\,\mathrm{kHz}$, capturing frequencies up to 22.05 kHz according to the *Nyquist Theorem*. For further details about sampling and quantization, please refer to [123].

2.2 Fourier Transform and Spectrograms

The Fourier transform is a fundamental tool in audio signal processing, which converts a time-domain signal into a function of frequency. Following [123], we define the *discrete Fourier transform (DFT)* of length $N \in \mathbb{N}$ for a discrete-time signal $x = (x(0), ..., x(N-1))^\mathrm{T}$ as

$$\mathcal{X}(k) = \sum_{n=0}^{N-1} x(n) \exp(-2\pi i k n / N) \tag{2.3}$$

for the frequency index $k \in [0 : K]$. Due to the real-valued nature of the signal x, the number of frequency bins K corresponds to the frequency index at the Nyquist frequency $K = \lfloor N/2 \rfloor$. Therefore, the frequency bins beyond this upper half of the spectrum are eliminated, which represent the negative frequencies. The complex value $\mathcal{X}(k)$ denotes the *Fourier coefficient*. Its magnitude $|\mathcal{X}(k)|$ intuitively reflects the degree to which the signal contains a periodic oscillation of a particular frequency

$$F_{\mathrm{coef}(k)} = \frac{k \cdot F_\mathrm{s}}{N}. \tag{2.4}$$

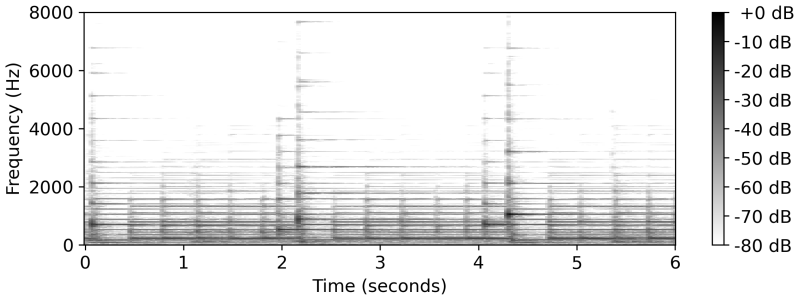

Figure 2.2: The magnitude spectrogram of our running example. This spectrogram exhibits magnitudes on a logarithmic decibel scale, where +0 dB represents the peak magnitude within the selected excerpt.

Given that the DFT averages the frequency information over the entire signal length N, the details regarding the temporal occurrence of the constituent frequencies remain hidden in the transformed representation. To recover the time information, the short-time Fourier transform (STFT) [66] is used, which basically applies the Fourier transform on small sections extracted from the original signal.

To compute the discrete STFT, the original discrete-time signal x is divided into overlapping analysis frames of length $N \in \mathbb{N}$ using a hop size of $H \in [1 : N-1]$. In each of these frames, the waveform is multiplied with a window function $w : [0 : N-1] \to \mathbb{R}$ and correlated with complex exponentials at different frequency indices k. Formally, we define the discrete STFT $\mathcal{X} \in \mathbb{C}^{M \times K}$ of the discrete-time signal x by

$$\mathcal{X}(m,k) = \sum_{n=0}^{N-1} x(n+mH)w(n)\exp(-2\pi ikn/N), \tag{2.5}$$

with the spectral frame index $m \in \mathbb{Z}$. The number of spectral frames $M \in \mathbb{N}$ is determined by the number of discrete signal samples.

From the complex-valued spectrogram $\mathcal{X}$, the *magnitude spectrogram* $\mathcal{Y} \in \mathbb{R}_{\geq 0}^{M \times K}$ is derived by

$$\mathcal{Y}(m,k) = |\mathcal{X}(m,k)|\,. \tag{2.6}$$

Figure 2.2 depicts the magnitude spectrogram of our running example. The vertical axis represents the frequency axis, whereas the horizontal axis corresponds to the temporal axis of the recording. To compute the spectrogram, we employ a sampling rate of $F_s = 44.1$ kHz, a window size of $N = 4096$, a hop size of $H = 1024$, and a standard Hann window. Variations in magnitude levels are reflected by the grayscale intensity in the spectrogram.

For a more in-depth exploration of the Fourier transform and discrete-time signal processing, please refer to [123, 129, 147].

2.3 Chroma-Based Audio Features

In most MIR tasks, the crucial information about the musical context is often encoded implicitly within the audio signal. For processing and analyzing music signals, an important step is thus to extract features which capture aspects of music signals that are relevant for a given task. For example, focusing on the harmonic content of a music signal is essential for tasks such as structure analysis [140, 141], music segmentation [174, 201], chord recognition [5, 99], or music synchronization [46, 57, 64].

Chroma features, also known as pitch class profiles [65, 71], are frequently used due to their effectiveness and robustness in capturing the harmonic content of music signals [123]. These features transform a signal into a time–chroma representation, which reduces all occuring pitches into the twelve pitch classes $\mathrm{C}, \mathrm{C}^{\#}, \mathrm{D}, \mathrm{D}^{\#}, \mathrm{E}, \mathrm{F}, \mathrm{F}^{\#}, \mathrm{G}, \mathrm{G}^{\#}, \mathrm{A}, \mathrm{A}^{\#}, \mathrm{B}$ by disregarding the octave information (and assuming enharmonic equivalence).

In the following, we first introduce a pitch-based log frequency spectrogram which serves as an intermediate step for the chroma computation in Section 2.3.1. Then, we elaborate on chroma features in Section 2.3.2.

2.3.1 Log-Frequency Spectrogram and Pitch Features

The relationship between frequency and perceived musical pitch is inherently logarithmic. However, the frequency axis of the spectrogram introduced in Equation (2.6) is linearly spaced. In this section, we explain how a magnitude spectrogram can be transformed into a *time–pitch representation* through a log-frequency spectrogram.

The core idea of the log-frequency spectrogram lies in redefining the frequency axis to correspond to the logarithmically-spaced frequency bands of the equal-tempered scale. Identifying pitches with musical instrument digital interface (MIDI) note numbers (with the pitch A4 corresponding to MIDI note number $p = 69$), the center frequency of a pitch $p \in [0 : 127]$ is given by:

$$F_{\text{pitch}}(p) = 2^{(p-69)/12} \cdot 440. \tag{2.7}$$

Now, we define a set of associated frequency bins for each pitch p:

$$P(p) = \{k : F_{\text{pitch}}(p - 0.5) \leq F_{\text{coef}}(k) \leq F_{\text{pitch}}(p + 0.5)\}. \tag{2.8}$$

Using this, we obtain a log-frequency spectrogram $\mathcal{Y}_{\text{LF}} : \mathbb{Z} \times [0 : 127] \to \mathbb{R}_{\geq 0}$, defined as:

$$\mathcal{Y}_{\text{LF}}(m, p) = \sum_{k \in P(p)} |\mathcal{X}(m, k)|^2. \tag{2.9}$$

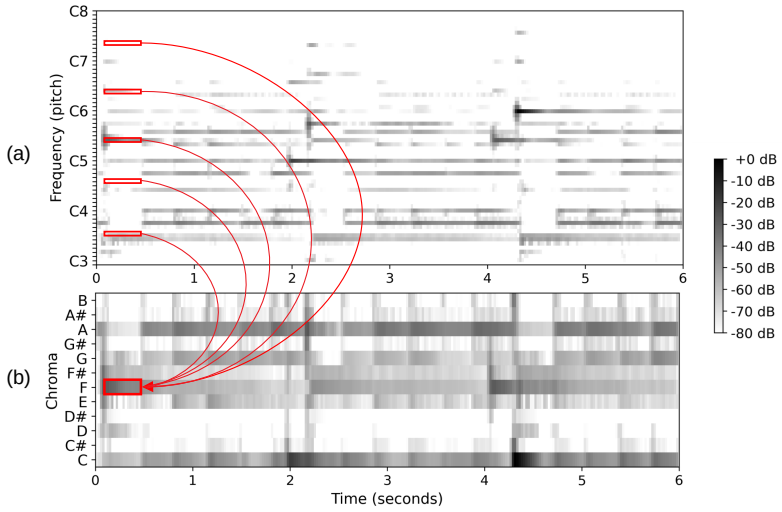

Figure 2.3: Feature representations of our running example, Mozart's Piano Concerto No. 21 in C Major, KV. 467, 2nd movement. **(a)** Pitch-based log-frequency spectrogram $\mathcal{Y}_{\mathrm{LF}}$, **(b)** Chromagram C.

Note that $\mathcal{Y}_{\mathrm{LF}}$ now comprises a frequency bin for each MIDI pitch. See Figure 2.3a for an illustration of the log-frequency spectrogram of our running example.

2.3.2 Chroma Features

Humans perceive pitch in a periodic manner; pitches separated by an octave – equivalent to 12 semitones in the equal-tempered scale – are perceived as possessing a similar quality or "color." This phenomenon is also described as sharing the same chroma[1] or pitch class. Chroma features exploit this observation, by aggregating all energy from pitches that belong to the same pitch class.

Following [123], we can derive a chromagram $C : \mathbb{Z} \times [0:11] \to \mathbb{R}$ from the pitch-based log-frequency spectrogram by:

$$C(m,c) = \sum_{\{p \in [0:127] : p \bmod 12 = c\}} \mathcal{Y}_{\mathrm{LF}}(m,p), \tag{2.10}$$

for $c \in [0:11]$.

Figure 2.3 visualizes the derivation of chroma features from a pitch-based log-spectrogram. In Figure 2.3a, the F pitches are highlighted by red rectangles in the pitch-based log-frequency spectrogram. Note that the pitches lower than C3 are not shown in this figure. Figure 2.3b shows the chromagram, where the F chroma bin is also indicated by a red rectangle. Throughout this excerpt, it is observable that F, A, and C chromas are dominant, as anticipated, given that the musical passage being analyzed is in F major.

[1] From Ancient Greek $\chi\rho\omega\mu\alpha$ (khrôma, "color")

3 High-Resolution Music Synchronization

This chapter is based on [135]. The first author Yigitcan Özer is the main contributor to this article. Yigitcan Özer and Matěj Ištvánek made equal contributions to the implementation and design of experiments. Matěj Ištvánek and Vlora Arifi-Müller curated the dataset. Meinard Müller closely supervised this work and contributed with Matěj Ištvánek to the article's writing.

Audio synchronization aims at aligning multiple recordings of the same piece of music. Traditional synchronization approaches are often based on DTW using chroma features as an input representation. Previous work has shown how one can integrate onset cues into this pipeline for improving the alignment's temporal accuracy [57, 77]. Furthermore, recent work based on deep neural networks has led to significant improvements for learning onset, beat, and downbeat activation functions. However, for music with soft onsets and abrupt tempo changes, these approaches may be unreliable, leading to unstable results. As the main contribution of this chapter, we introduce a combined approach that integrates activation functions into the synchronization pipeline. We show that this approach improves the temporal accuracy thanks to the activation cues while inheriting the robustness of the traditional synchronization approach. Conducting experiments based on string quartet recordings, we evaluate our combined approach where we transfer measure annotations from a reference recording to a target recording.

The remainder of this chapter is organized as follows. Following the introduction in Section 3.1, in Section 3.2, we introduce our combined synchronization approach, explore conventional and DL-based activation cues, and show how to integrate activation functions into the synchronization pipeline. In Section 3.3, we present our dataset, the measure transfer between string quartet recordings as our application scenario, and report on our systematic experiments and empirical results. Finally, we conclude in Section 3.4 with prospects on future work.

3.1 Background

In MIR, synchronization techniques are essential for several applications including score following [172], content-based retrieval [62], automatic accompaniment [37], or performance analysis [105, 162]. Beside these applications, music synchronization has a great potential to simplify data augmentation, data annotation, and model evaluation. For example, one can use music synchronization to obtain additional training data for deep learning methods semi-automatically by transferring annotations from one recording to another recording. Furthermore, using music synchronization, one can transfer measure positions between audio recordings for navigation purposes, structural segmentation, and cross-version analysis [98, 209].

While traditional synchronization approaches typically rely on alignment algorithms such as DTW and conventional chroma features used as the input representation [38, 124, 199], the integration of additional onset-related information has proven to enhance the synchronization accuracy [3, 57, 128]. Inspired by the combined approach from [57], where *decaying locally adaptive chroma onset (DLNCO)* features are integrated into the synchronization pipeline, we incorporate in this work onset, beat, and downbeat activation functions to obtain a better temporal accuracy while retaining the robustness of the original chroma-based synchronization approach (see Figure 3.1 for an illustration of the overall approach). The addition of activation functions results in a grid-like structure in the cost matrix, which guides the alignment through activation cues that point to note onsets or other musical events.

While the integration of DLNCO and spectral flux (SF) have led to substantial improvement of synchronization results [57, 77], the detection of soft onsets constitutes a challenging problem due to their long attack phase with a slow rise in energy. To adapt the onset detection task to music recordings which comprise soft onsets and temporal–spectral modulations such as vibrato (e. g., string music), Böck and Widmer [13] introduced the superflux (SF*) feature. Furthermore, deep learning (DL) methods such as bidirectional long short-term memory (BLSTM) networks [59] and convolutional neural networks (CNNs) [168] have led to significant improvements compared to conventional onset detectors.

As the main contribution of this chapter, we show how one can integrate conventional and DL-based activation functions into the synchronization pipeline. Different from the approach in [57], we do not apply any hard peak picking but directly use onset-related activation cues. Furthermore, we go beyond onsets by integrating activation functions that indicate onset, beat, and downbeat positions. In particular for music with noisy and unreliable onset cues, we show that beat and downbeat cues are more reliable and better suited for improving the synchronization accuracy. For extracting beat and downbeat activation functions, we build on recent work by Böck et al. [12, 15], using recurrent neural network (RNN) models for extracting beat and downbeat activation functions.

To better understand our improved synchronization pipeline, we compare several synchronization approaches where we transfer measure annotations from a reference recording to a target recording, similar to [208]. In particular, we conduct systematic experiments based on three versions of the String Quartet

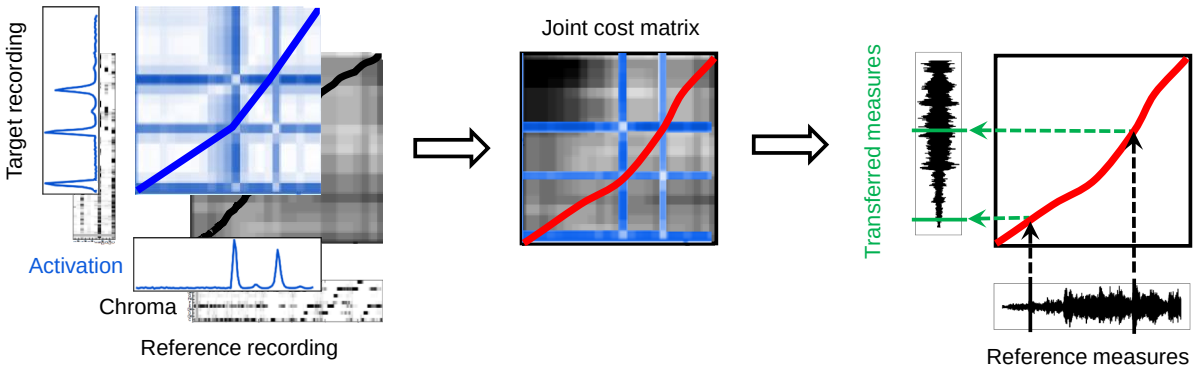

Figure 3.1: Overview of the combined approach integrating activation functions into a conventional chroma-based synchronization pipeline. The cost matrix computed with activation cues (blue) yields a grid-like structure to guide the alignment of musical events, whereas the chroma-based cost matrix (black) accounts for the robustness of the overall synchronization. The resulting warping path (red) is used for transferring measure positions.

No. 12 in F major, Op. 96, composed by Antonín Dvořák. As string music generally comprises vibrato, tremolo, rubato, and abrupt tempo changes, which increase the musical complexity, the synchronization of string quartets is a challenging scenario. We show that integrating DL-based activation functions significantly improves the temporal accuracy while retaining the robustness of chroma features.

3.2 Combined Synchronization Approach

In this section, we show how activation functions can be integrated into the synchronization pipeline to enhance the temporal accuracy of traditional chroma-based synchronization approaches. Here, we regard the activation function as a function that yields a value between 0 and 1. Each entry in the function indicates the likelihood of a certain musical event, e. g., onsets, beats, or downbeats, for each frame (time position). In the ideal case, the value of the activation function is one when an event occurs and zero otherwise. Note that we do not apply any peak picking in our approach, but only use the activation functions as temporal cues. This is opposed to onset detection or beat tracking where one needs to apply a temporal decoding method to obtain an explicit representation of onset and beat positions from the activation functions.

In the following, we first explore conventional onset-based activation functions in Section 3.2.1. Then, in Section 3.2.2, we investigate DL-based onset, beat and downbeat detectors. Finally, in Section 3.2.3, we explain how we integrate activation functions into the synchronization pipeline.

3.2.1 Conventional Onset-Based Activation Functions

3.2.1.1 DLNCO

DLNCO features are 12-dimensional pitch-based onset features, which combine the robustness of the chroma features with the accuracy of one-dimensional onset features. To compute DLNCO features, we first apply a pitch-wise audio decomposition. Then, we derive pitch-wise onset cues by considering points of energy increase (see Figure 3.2c for an illustration). DLNCO features are particularly suited for the music with clear note attacks such as piano music. For further details about the computation of DLNCO features, we refer to [57].

3.2.1.2 SF

SF captures the changes in the spectral content of an audio signal, and is widely used for onset detection [6, 124]. For the computation of SF, we apply a first-order differentiator on the log-compressed magnitude spectrogram of a music recording. Half-wave rectification follows the differentiation to keep only the positive differences between subsequent frames. As a final step, we subtract a local average function to enhance the peak structure (see Figure 3.2d).

3.2.1.3 SF*

Superflux (SF*) is a modified version of SF for detecting soft onsets [13]. These features are suitable for music recordings with vibrato, such as strings quartets. Similar to the SF algorithm, SF* also relies on the detection of positive changes in the energy over time. However, it includes a trajectory-tracking stage through maximum filtering, instead of simply calculating the difference between spectral bins over time. Trajectory tracking helps to suppress spurious spectral peaks, especially arising from vibrato. For further information, we refer to [13] (see also Figure 3.2e).

3.2.2 DL-Based Activation Functions

3.2.2.1 CNN Onset Detector

Schlüter and Böck [168] approach the onset detection task as a computer vision problem, where magnitude spectrograms of the audio recordings are used as the input to a CNN. Onsets are often characterized by rapid transient changes in the spectrum, resulting in sharp edges that are clearly visible in a spectrogram. Using convolutional kernels, one can easily detect these sharp edges of onsets. Similar to SF-based methods, the proposed CNN model computes spectro–temporal differences and captures percussive and pitched onsets. The resulting activation function is referred to as `DL-0` (see Figure 3.2f for an example).

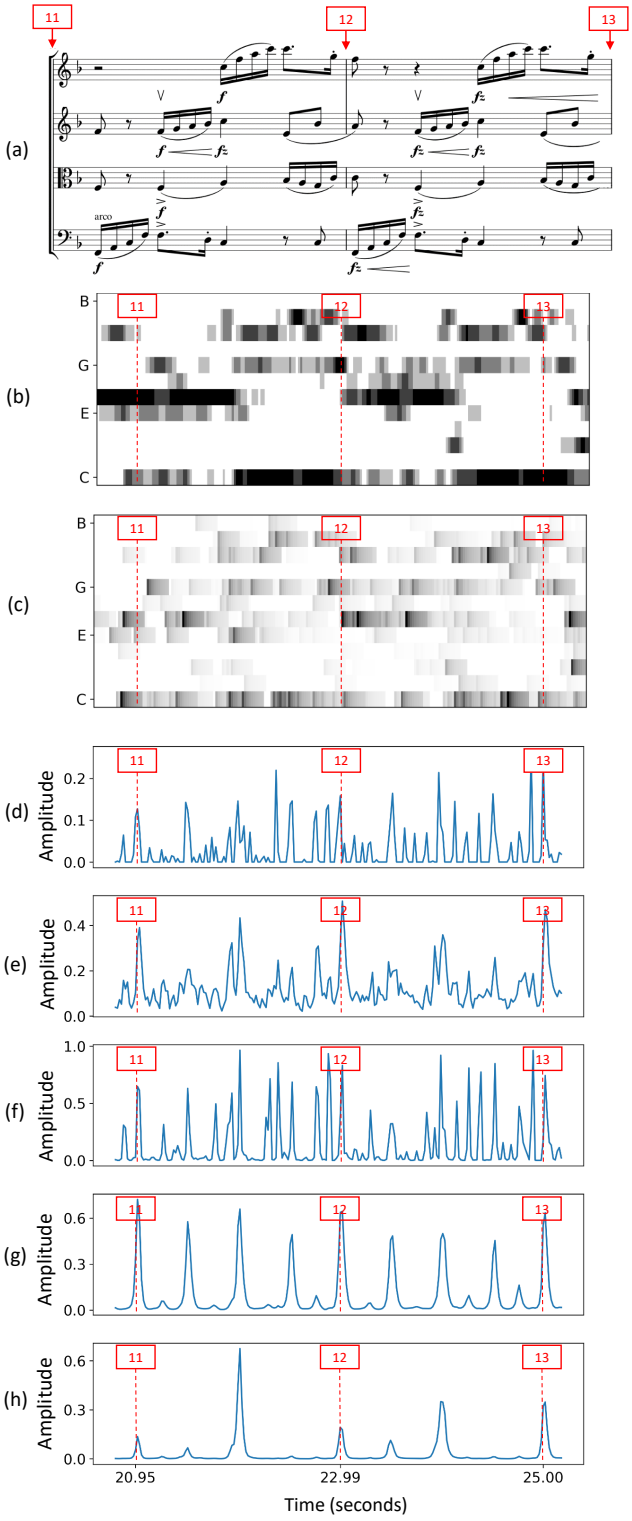

Figure 3.2: Chroma features and activation functions computed for an excerpt of the String Quartet No. 12 in F major, Op. 96 (first movement) composed by Antonín Dvořák, performed by the Borromeo Ensemble. Activation functions are shown in blue and ground-truth measure positions in red. **(a)** Sheet music representation of the measures 11–13. **(b)** Chroma **(c)** DLNCO **(d)** SF **(e)** SF* **(f)** Onsets (DL-O) **(g)** Beats (DL-B) **(h)** Downbeats (DL-D)

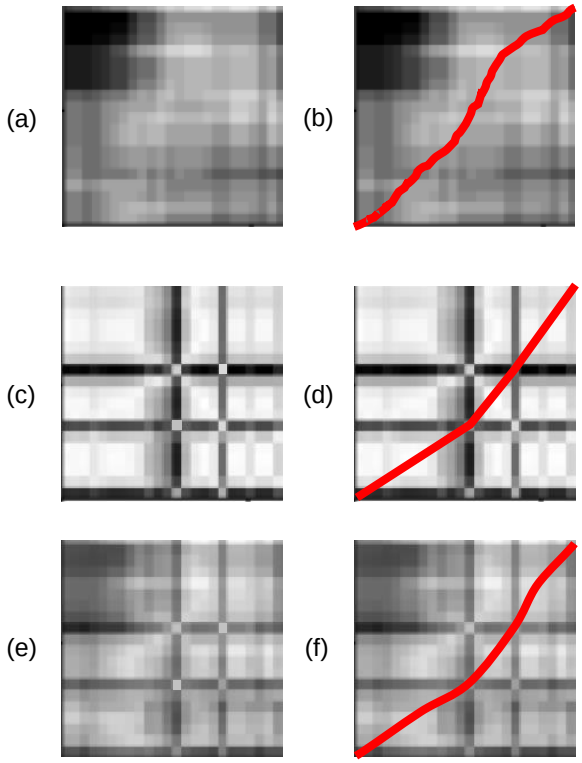

Figure 3.3: Excerpts from cost matrices and corresponding warping paths computed with DTW **(a)** / **(b)**: $\mathbf{C}_{\mathtt{CHROMA}}$, **(c)** / **(d)**: $\mathbf{C}_{\mathtt{ACT}}$, **(e)** / **(f)**: $\alpha\mathbf{C}_{\mathtt{CHROMA}} + (1 - \alpha)\,\mathbf{C}_{\mathtt{ACT}}$, with $\alpha = 0.5$.

3.2.2.2 RNN Beat Detector

In the case of unreliable and noisy onset cues, using beat activation functions constitutes a more feasible solution to improve the temporal alignment. To compute such activation functions, we use the BLSTM model by Böck and Schedl [12] for framewise beat detection. BLSTMs can effectively model the temporal context of the data and is therefore suitable for beat tracking. In the proposed approach, magnitude spectrograms computed with three different window lengths, and their first order differences are used as the input to the network. The network outputs encode the likelihoods of beat positions, as illustrated by Figure 3.2g. In our experiments, the resulting beat activation functions are denoted as `DL-B`.

3.2.2.3 RNN Downbeat Detector

Böck et. al [15] present an RNN model to jointly detect beat and downbeats. Like the previously mentioned onset and beat detectors, this model also operates on magnitude spectrograms. The downbeat detector uses an RNN similar to the proposed network in [12] to model beats and downbeats. In our experiments, we only use the probability of downbeats as the activation cues, which we denote as `DL-D` (see Figure 3.2h for an illustration).

3.2.3 Combined Synchronization with Activation Functions

To find the optimal alignment between two feature sequences $X := (x_1, \ldots, x_N)$ and $Y := (y_1, \ldots, y_M)$, where $n \in \{1, \ldots, N\}$ and $m \in \{1, \ldots, M\}$, we rely on DTW. By comparing each pair of elements in the feature sequences, we obtain a cost matrix $\mathbf{C}(n, m) := c(x_n, y_m)$ of size $N \times M$, where c defines a local cost measure. Then, an *optimal warping path* is determined via dynamic programming. We refer to [124] for a detailed account on DTW for music synchronization. For efficient implementations, we refer to [126, 197].

We now adapt the combined synchronization idea by Ewert et al. [57], integrating conventional onset-based and DL-based activation functions into the synchronization pipeline. Building upon this approach, we introduce three cost matrices. The first one $\mathbf{C}_{\texttt{CHROMA}}$ is a cost matrix based on the normalized chroma features and the cosine distance, see Figure 3.3a for an example. The second cost matrix $\mathbf{C}_{\texttt{ACT}}$ is computed using the Euclidean distance and conventional onset-based or DL-based activation functions as introduced in Section 3.2.1 and Section 3.2.2, respectively. Note that $\mathbf{C}_{\texttt{ACT}}$ exhibits grid-like structures. The visible horizontal and vertical grid lines of high cost (shown in black) correspond to high values in the first and second function, respectively. Only when a horizontal grid line intersects with a vertical grid line, the cost matrix has a small cost value at this intersection point (and also in a small neighborhood). This is where a high activation value of the first sequence meets another high activation value of the second sequence. In other words, these *intersection cells* encode a pair of time positions where two musical events (onsets, beats, downbeats) meet (see Figure 3.3c). Furthermore, the sections in the activation functions which have low values lead to homogeneous, zero-cost regions in the cost matrix $\mathbf{C}_{\texttt{ACT}}$. The third matrix is the sum of two cost matrices

$$\mathbf{C} = \alpha \mathbf{C}_{\texttt{CHROMA}} + (1 - \alpha)\mathbf{C}_{\texttt{ACT}}, \tag{3.1}$$

where $\alpha \in [0, 1]$ is a weighting parameter. The sum $\mathbf{C}$ accounts for both harmonic or melodic information of the representations via $\mathbf{C}_{\texttt{CHROMA}}$ and additional activation cues via $\mathbf{C}_{\texttt{ACT}}$. Figure 3.3e illustrates an example using $\alpha = 0.5$.

Comparing the resulting optimal warping paths using $\mathbf{C}_{\texttt{CHROMA}}$ in Figure 3.3b, $\mathbf{C}_{\texttt{ACT}}$ in Figure 3.3d, and $\mathbf{C}$ in Figure 3.3f, we can observe an enhancement of the temporal alignment. The inclusion of DL-based onset, beat, and downbeat cues leads to an improvement of the warping path guided by grid structure's intersection points. Note that $\mathbf{C}_{\texttt{ACT}}$ remains zero in the regions without any novel events, and the overall alignment of $\mathbf{C}$ is mainly guided by $\mathbf{C}_{\texttt{CHROMA}}$.

Table 3.1: Overview of four movements, including number of measures, time signature, and global tempo in BPM for each movement.

Movement	#Measures	Time signature	Global Tempo
M1	239	4/4	100
M2	97	6/8	84
M3	244	3/4	185
M4	382	2/4	140

Table 3.2: Version, identifier, recording year, and duration of each movement.

			Duration (seconds)				
Version	ID	Year	M1	M2	M3	M4	Σ
Alban Berg	A	1991	599	410	238	323	1570
Borromeo	B	2012	540	465	228	338	1571
Prague	P	1973	584	424	250	322	1580
		Σ	1723	1299	716	983	4721

3.3 Experiments

3.3.1 Dataset

The genre of string quartet is composed for a small conductor-less ensemble, which consists of two violins, a viola and a violoncello. In our experiments, we use three versions (performances) of the String Quartet No. 12 in F major, Op. 96, by Antonín Dvořák, which comprises four movements. To give an insight of the musical properties of the string quartet, Table 3.1 provides the number of measures, time signature, and global tempo for each movement based on the recordings in our dataset. Note that the global tempo does not reflect any local tempo deviations. Its purpose is to indicate at what pace (average of three performances) a given movement is performed. In each recording, the repetitions are played as notated in the sheet music, thus ensuring structural consistency.

For each recording (in the following referred to as version), we manually annotated the measure positions. In Table 3.2, the name of the version, the identifier, the recording year for each version, and the duration of each movement are listed. Note that each of the three performances last around 26 minutes in total, whereas durations of the movements may vary across different versions.

3.3.2 Beat and Downbeat Tracking

As a baseline, we first introduce how DL-based beat [12] and downbeat trackers [15] perform in the detection of measure positions, where a dynamic Bayesian network (DBN) was used for post-processing (peak picking). Table 3.3 provides precision, recall, and F-measure values using a tolerance of $\tau = 70$ ms. Here, each entry indicates the mean value over different performances. Due to the higher density of beats, the beat tracker reveals a higher recall than the downbeat tracker for each movement, leading to a low precision and F-measure. Furthermore, each movement has a different time signature, which results in a different number of beats per measure and needs to be taken into consideration as prior information for the DBN.

Table 3.3: Precision (P), Recall (R), and F-measure (F) based on methods DL-B for beat, and DL-D for downbeat tracking with DBN post-processing, evaluated on reference measure annotations and tolerance $\tau = 70$ ms. ϕ denotes the average accuracy over four movements M1, M2, M3, and M4.

	DL-B & DBN			DL-D & DBN		
	P	R	F	P	R	F
M1	0.194	0.806	0.312	0.648	0.586	0.612
M2	0.138	0.820	0.236	0.657	0.643	0.650
M3	0.327	0.539	0.407	0.524	0.224	0.314
M4	0.517	0.908	0.655	0.876	0.647	0.742
ϕ	0.294	0.768	0.403	0.676	0.525	0.580

3.3.3 Synchronization Results

In this section, we describe our experimental setting and evaluate audio alignments obtained using chroma features and different activation functions. In our experiments, we use the resulting warping path to transfer the measure positions annotated for the reference recording to the target recording. Given two versions of the same music piece with the time-continuous axes $[0, T_1]$ and $[0, T_2]$, the monotonous alignment can be modeled as a function

$$\mathcal{A} : [0, T_1] \to [0, T_2].$$

The pairwise alignment error ϵ_P for a given alignment of two recording is specified as the mean over the values

$$\epsilon_P(g_1) := |\mathcal{A}(g_1) - g_2|,$$

where $(g_1, g_2) \in [0, T_1] \times [0, T_2]$ denotes the ground-truth pairs of measure annotations. As an evaluation metric, we use *accuracy*, which is defined as the proportion of correctly transferred measure positions with a pairwise alignment error below a given tolerance τ [145].

In the following, we use eight different synchronization approaches based on conventional chroma features and the combination of chroma features with DLNCO features, SF, SF*, DL-based onsets (DL-O), DL-based beats (DL-B), DL-based downbeats (DL-D), and finally a 3-dimensional stacked activation function combining DL-based onsets, beats, and downbeats (DL-OBD) (see Section 3.2 for a detailed overview of activation functions). We use a feature rate of 50 Hz for the computation of chroma, and conventional onset features. To generate DL-based features, we use the *madmom* [14] library, for which we utilize the default setting 100 Hz as the feature rate, and downsample generated features to 50 Hz (after low-pass filtering).

3.3.3.1 Overall Result

To get a first impression of the alignment behavior of different approaches, Figure 3.4 illustrates an overview of average accuracy values (averaged over all movements and all pairs of different performances) and for different tolerances τ. Obviously, one can observe that the synchronization accuracy improves with increasing threshold. For example, using $\tau = 30$ ms, the average synchronization accuracy is 0.454 when

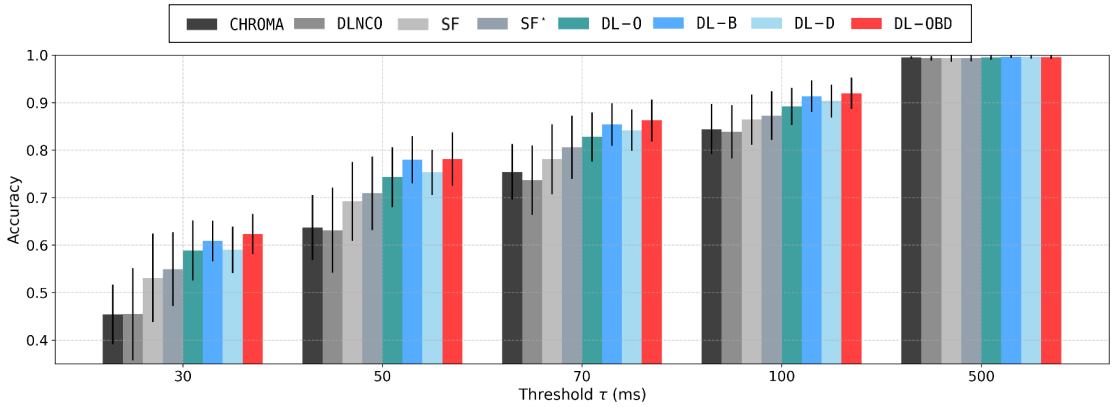

Figure 3.4: Comparison of the average accuracy values for different synchronization approaches and different threshold parameters τ. The accuracy denotes the proportion of correctly transferred measure positions having an error below a given tolerance τ.

Table 3.4: Accuracy values based on chroma and DL-OBD for different tolerances $\tau = 30, 70, 100$ ms. ϕ denotes the average accuracy over four movements M1, M2, M3, and M4.

	$\tau = 30$ ms		$\tau = 70$ ms		$\tau = 100$ ms	
	CHROMA	DL-OBD	CHROMA	DL-OBD	CHROMA	DL-OBD
M1	0.408	0.576	0.723	0.813	0.817	0.877
M2	0.396	0.600	0.694	0.842	0.789	0.917
M3	0.528	0.655	0.827	0.912	0.910	0.956
M4	0.485	0.662	0.773	0.884	0.861	0.930
ϕ	0.454	0.624	0.754	0.863	0.844	0.920

using only chroma features, and the accuracy increases to 0.624 when integrating the DL-OBD activation cues. This is a substantial improvement.

Next, we focus on the comparison of different approaches for $\tau = 70$ ms, which is a common tolerance value for the evaluation of music synchronization and beat tracking procedures. The inclusion of DLNCO slightly worsens the alignment since DLNCO features are not suited for soft onsets as occuring in string music. However, the integration of SF and SF* into the synchronization pipeline results in a better accuracy. Among conventional onset features, SF* shows a better performance than SF and DLNCO, owing to the fact that SF* features account for the detection of soft onsets and are therefore more suitable for string music. Furthermore, using DL-based methods leads to a better synchronization result compared to the conventional methods. Note that the integration of DL-OBD, which combines DL-O, DL-B, DL-D, reveals the best accuracy among all the synchronization approaches. It is also interesting to observe that DL-B, which is based on beats, is the second-best among DL-based approaches and leads to better accuracy values than downbeat-based DL-D, whereas DL-O reveals the lowest accuracy among the DL-based approaches.

In general, similar trends can be observed when using other thresholds. Using $\tau = 500$ ms, all synchronization approaches yield nearly perfect results.

Table 3.5: Accuracy based on chroma and DL-OBD across different synchronization pairs, for different tolerances τ. The first column indicates the pair of versions (see Section 3.3.1).

	τ = 30 ms		τ = 70 ms		τ = 100 ms	
	CHROMA	DL-OBD	CHROMA	DL-OBD	CHROMA	DL-OBD
AP	0.494	0.636	0.774	0.864	0.839	0.928
PA	0.500	0.632	0.778	0.864	0.847	0.928
AB	0.434	0.576	0.722	0.849	0.830	0.907
BA	0.451	0.580	0.743	0.863	0.857	0.920
BP	0.429	0.662	0.762	0.870	0.850	0.919
PB	0.417	0.657	0.746	0.866	0.843	0.915
ϕ	0.454	0.624	0.754	0.863	0.844	0.920

3.3.3.2 Dependency on Movement

In our next experiment, we analyze the synchronization accuracy across different movements. Table 3.4 provides a comparison of alignment results based on the conventional chroma-based approach and our proposed combined method DL-OBD per movement for different tolerances τ. For example, considering the first movement and τ = 70 ms, the accuracy of 0.723 for the chroma-based approach increases to 0.813 when using our combine approach DL-OBD. One can observe a similar trend across different movements for different tolerance parameters τ. The second movement tends to yield the lowest accuracy values when using only chroma features, while the integration of DL-OBD significantly improves the synchronization accuracy from 0.694 to 0.842 for the second movement. One reason may be that the beat and downbeat information leads to a significant improvement in synchronization accuracy of slower sections.

3.3.3.3 Dependency on Performance

As a final experiment, we provide a comparison of the accuracy values across different performances for different tolerance parameters τ in Table 3.5. In general, using a combination of chroma and activation functions significantly improves the accuracy for all the synchronization pairs and tolerances. For τ = 70 ms, chroma reveals an average accuracy of 0.774 for AP and DL-OBD improves the accuracy to 0.864. Remarkably, across other synchronization pairs the synchronization accuracy values are similar. Nonetheless, deviations may occur due to soft onsets, slight inconsistencies in ground-truth annotations, and linear interpolation while measure transfer. Note that the synchronization accuracy rather depends on the musical complexity and structure, e. g., across different movements, but not on the performances.

3.4 Conclusion

In this chapter, we investigated the incorporation of conventional onset features, and activation cues obtained by recent DL-based onset, beat, and downbeat detectors to a conventional chroma-based synchronization pipeline. Our results reveal that the integration of a combined version of onset, beat, and downbeat activation functions significantly improves the synchronization accuracy while maintaining the robustness

of the original chroma-based approach. This improvement is particularly notable in scenarios involving music with noisy and unreliable onset cues, e. g., string quartets, where beat and downbeat cues proved more reliable and better suited for improving the alignment accuracy. Synchronization, as detailed here, is a key component for the subsequent chapters of this thesis. It plays a crucial role in the creation of our multitrack dataset, as will be introduced in Chapter 5, and also significantly contributes to the data augmentation processes that are of central importance for the methodologies presented in Chapter 6.

4 Source Separation with Test-Time Adaptation

This chapter is based on [131]. The first author Yigitcan Özer is the main contributor to this article. In collaboration with his supervisor Meinard Müller, he devised the ideas, developed the formalization, and wrote the paper. Furthermore, Yigitcan Özer implemented all approaches and conducted the experiments.

Audio signals are typically complex mixtures of various sound sources, which can refer to several people talking simultaneously in a room, different musical instruments playing together, or a speaker talking in the foreground with music being played in the background. Decomposing a complex sound mixture into its constituent components is one of the fundamental research areas in audio signal processing, which is often referred to as *source separation*. A classical separation scenario is the *cocktail party problem*, where the objective is to isolate a specific speaker's voice from a mixture of conversations with multiple speakers and background noise [18]. Within the context of music, a source might refer to a melody, a bass line, a drum track, a general instrumental voice, or a group of instruments. MSS aims at decomposing a musical mixture into its constituent sources, as if these were recorded in an isolated fashion [123]. The practical importance of separating these individual sources from a sound mixture can be seen in diverse applications, such as creating karaoke systems, aiding in music production, facilitating music transcription, and supporting music analysis.

In this chapter, we address the novel and rarely considered source separation task of decomposing piano concerto recordings into separate piano and orchestral tracks. Being a genre written for a pianist typically accompanied by an ensemble or orchestra, piano concertos often involve an intricate interplay of the piano and the entire orchestra, leading to high spectro–temporal correlations between the constituent instruments. We generate artificial training material by randomly mixing sections of the solo piano repertoire (e. g., piano sonatas) and orchestral pieces without piano (e. g., symphonies) to train state-of-the-art DNN models for MSS. As our main contribution, we propose a test-time adaptation (TTA) procedure, which exploits random mixtures of the piano-only and orchestra-only parts in the test data to further improve the separation quality.

The remainder of the chapter is organized as follows. Following the introduction in Section 4.1, in Section 4.2, we describe our MSS approach, explore the recent state-of-the-art spectrogram-domain

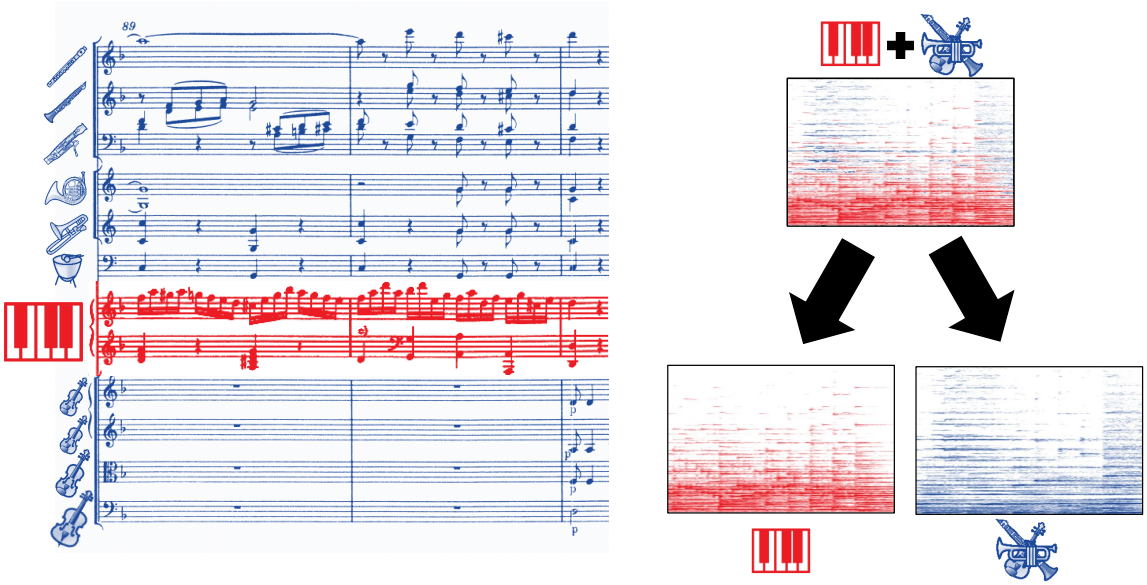

Figure 4.1: An excerpt from the first movement of the Piano Concerto in D minor (KV466) by Wolfgang Amadeus Mozart. Our goal is to estimate the magnitude spectrograms of the piano part (red) and orchestral part (blue).

DNN model Spleeter to address the MSS task and describe our experimental setting. In Section 4.3, we introduce the TTA procedure to improve the separation quality of piano concertos and present our dataset. In Section 4.4, we report on the quantitative empirical results, including a subjective evaluation. Finally, we conclude in Section 4.5.

4.1 Background

From the Baroque era onward, numerous composers have written piano concertos, which are compositions highlighting the virtuosity of pianists. As an example, Figure 4.1 shows an excerpt from the first movement of Piano Concerto in D minor (KV 466) by Wolfgang Amadeus Mozart. In addition to the large number of compositions, there also are many prominent historical recordings of piano concertos in classical music archives. In this context, separation of piano and orchestra can enable applications such as editing and upmixing historical and modern piano concerto recordings.

As the piano is the lead instrument and the orchestra takes over the accompaniment, separation of piano concertos can be regarded as a lead instrument and accompaniment separation task [22, 90, 152]. The piano has distinct timbral characteristics, e. g., clear onsets, which intuitively may help a separation model in distinguishing it from orchestral instruments such as strings, woodwinds, and brass. Nevertheless, the high spectro–temporal correlations between the piano and orchestral parts in concertos constitute a challenging problem.

Since musical signals often exhibit non-stationary spectro–temporal properties and may be highly correlated in time and frequency, MSS proves to be a challenging task in music processing [24]. In the last years, DNNs have led to substantial improvements in separating musical sources [42, 43, 83, 88, 95, 109, 165, 181, 185, 189]. Despite their effectiveness, 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 with isolated stems are limited to popular music, e. g., MUSDB18 [151] and MoisesDB [142]. However, for western classical music, professionally produced multitrack recordings are quite rare [10, 166].

In this chapter, we adapt the spectral-based Spleeter model [83] to address the separation of piano and orchestra in piano concertos. Spleeter has achieved impressive results for the separation of four stems (vocals, drums, base, and others) in the Signal Separation Evaluation Campaign (SiSEC) challenge [184]. Building upon its standard architecture, which is an encoder–decoder CNN, we train our baseline model using a proprietary dataset.

When training deep MSS models, generating random mixes of solo instrument recordings may improve the separation quality [185, 202]. Random mixing for data generation and augmentation has opened up new paths for separating instrument mixtures, for which multitrack recordings are not available. For example, Chiu et al. [30] created their own synthetic dataset comprising classical violin and pop piano solo recordings, which serve as training material of an MSS model for the separation of piano and violin duos. Inspired by the recent advances in deep learning and data augmentation, we generate in our setting 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.

Whereas one can improve the performance of data-based models using artificially generated data, a supervised machine learning model necessitates a representative training set to ensure its robustness during the testing phase. In the case of MSS, many recordings have specific acoustic properties (e. g., historical recordings) that are not reflected in training datasets, thus leading to a poor separation performance. To overcome this issue, one can exploit the occurence of repeating patterns in the same recording [150], use bootstrapping to improve separation results [45, 103], or adapt a pre-trained model to one specific target mixture [25, 206]. In this chapter, our approach is based on the latter strategy. To this end, we first train on our artificial dataset and then finetune the model at the testing stage. As our main contribution, we propose a TTA method similar to [188], where we exploit that a piano concerto typically has relatively long piano-only and orchestra-only passages. Generating random mixes of these sections, we adapt the separation model at the test time individually for each piano concerto in our test dataset. Our systematic experiments highlight the benefits of TTA trained with the spectrogram-domain MSS model Spleeter [83]. To evaluate the performance of our models, we use the widely-used SDR [205], and the *2f-score* [92], which is an objective quality measure. Furthermore, we conduct multiple stimulus with hidden reference and anchors (MUSHRA) listening tests [87] to assess the perceptual separation quality.

4.2 Source Separation Approach

In this chapter, our focus is the spectrogram-domain Spleeter model [83], which learns to approximate the magnitude spectrogram of a target source. To reconstruct the separated audio signals, spectrogram-based models typically use binary masking, soft masking or multichannel Wiener filtering [51, 110].

In particular, we adapt the Spleeter model [83] for separating piano concertos. The default model architecture is based on a *U-Net* [157], which has recently been a widely-used architecture in the MIR community to address the MSS task [34, 88, 119, 181, 206]. Following this trend of research, we adapt a U-Net model to predict the magnitude spectrograms of the constituent piano and orchestral parts in a piano concerto. In the following, we revisit the U-Net architecture in Section 4.2.1 and present our experimental setting in Section 4.2.2.

4.2.1 U-Net Model

The U-Net model is composed of a convolutional encoder–decoder architecture with skip connections, which account for the resurrection of fine-grained details in the reconstructed representation. Following the default setting of the U-Net model in [88], we use a 12-layer network (6 layers for the encoder and 6 for the decoder). Each encoder layer uses a strided 2D convolution with a kernel size of 5×5 and a stride size of 2, preceded by a leaky rectified linear unit (ReLU) activation function, and batch normalization. The decoder is composed of strided deconvolution layers with a kernel size of 5×5 and a stride size of 2, as in the encoder. The decoder uses ReLU as the activation function, different from the encoder. To avoid overfitting, we use here dropout with a probability of 0.5 in the first three layers of the decoder. The final layer of the network is a sigmoid activation function, yielding a soft mask for each target source, which contains values between 0 and 1. As the loss function, we use ℓ^1-norm between masked input mixture and target spectrograms. For further details about network architecture, we refer to [83, 88].

4.2.2 Experimental Setting

In this chapter, we use monaural recordings, which are sampled at 22.05 kHz. We generate the magnitude spectrograms using a Hann window size of 2048 and hop size of 512. In a first step, we train our models using an artificial dataset which contains 20-second random chunks from the mixtures of solo piano recordings (e. g., piano sonatas) and orchestral pieces without piano (e. g., symphonies) by 16 different composers from different periods. The total duration of our randomly generated proprietary dataset is circa 45 hours. We regard this model as our pre-trained model, which we denote as `PT`.

We train all our models on a single NVIDIA GeForce 1080 Ti graphics processing unit (GPU), using a batch size of 8, and a learning rate of 1e−4 with adaptive moment optimization (ADAM) optimizer. To

Composer	Full Name	Performer	Work ID	Year	M.	Dur. (T)	Dur. (E)	Dur. (C)
Beethoven	Piano Concerto No.1 in C major, Op.15	Schnabel	`BeetOp015`	1932	1	1020	170	277
Beethoven	Piano Concerto No.4 in G major, Op.58	Gulda	`BeetOp058`	1960	1	1116	159	197
Brahms	Piano Concerto No.1 in D minor, Op.15	Arrau	`BrahOp015`	1958	3	728	`N/A`	65
Haydn	Piano Concerto No.11 in D major, Hob. XVIII:11	Gulda	`HaydnHob018`	1962	1	486	83	52
Mozart	Piano Concerto No.20 in D minor, KV. 466	Renzi	`MozKV466`	`N/A`	1	862	129	161
Mozart	Piano Concerto No.21 in C major, KV. 467	`N/A`	`MozKV467`	1962	1	833	136	76
Mozart	Piano Concerto No.27 in B–flat major, KV. 595	Casadesus	`MozKV595`	1963	1	778	128	88
					Σ	5823	805	916

Table 4.1: Composer, full name of the work, performer, identifier, recording year, movement (M.), duration (Dur.) in seconds of total recording (T), exposition (E), and cadenza (C).

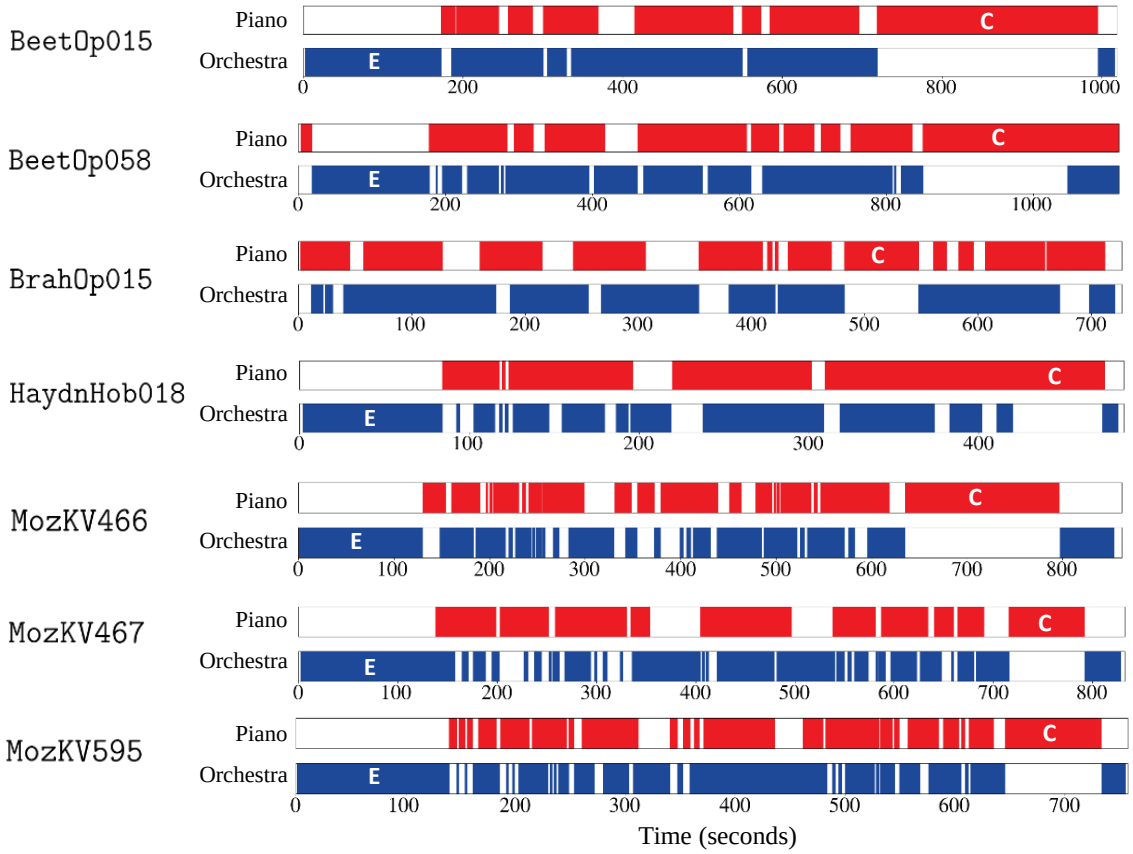

Figure 4.2: Annotation of the piano concertos in our dataset into the piano (red), orchestral (blue) parts. To finetune the pre-trained model with the TTA approach, we generate random mixtures of the piano-only (e. g., in the *Cadenza*, denoted as **C**) and orchestra-only (e. g., in the *Exposition*, denoted as **E**) sections.

improve the separation quality of real piano concerto recordings, we finetune the model with TTA, which we describe in the next section.

4.3 Test-Time Adaptation

Supervised deep learning models addressing the MSS task typically require a large dataset that consists of isolated recordings. As a data augmentation method, one can use random mixes to provide training material for an MSS model in the case isolated stems are missing [30, 202]. While this approach cannot simulate the harmonic and rhythmic relationships between various instruments in a real recording, it helps the model to distinguish timbral characteristics of the concurrent 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, thus resulting in a poor separation quality. Finetuning a pre-trained MSS model in the testing phase using a few samples drawn from the test data (also called *test-time adaptation (TTA)* [107, 188]) can improve the separation quality by capturing the specific acoustic features found in a music recording.

From this perspective, separation of piano concertos is a particularly suitable scenario for applying TTA thanks to their compositional form. Depending on the period in which the work 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 come from the audio material of the given test item. As a result, the mixes share the same recording conditions as the test data.

To investigate this approach, we consider seven piano concerto recordings (see Table 4.1). The selected movements of these piano concertos have a long cadenza, which contains only the piano (see Figure 4.2). Note that, with the exception of `BrahOp015`, these musical pieces also comprise a long exposition in which only the orchestra plays. For our experiments, we annotate the piano-only and orchestra-only sections, which are publicly available.[2] Exploiting the structural characteristics of piano concertos, we create random mixes of piano-only and orchestra-only sections, which serve as further training data for model adaptation for each piano concerto individually. In the next section, we investigate the improvement of qualitative and subjective separation quality via TTA.

4.4 Evaluation

In this section, we report on the separation results acquired by our pre-trained model `PT` and the finetuned models `TTA`. In Section 4.4.1 we describe our test dataset. We discuss the quantitative empirical results in Section 4.4.2 and present the subjective evaluation in Section 4.4.3.

[2] https://www.audiolabs-erlangen.de/resources/MIR/2022-PianoSep/

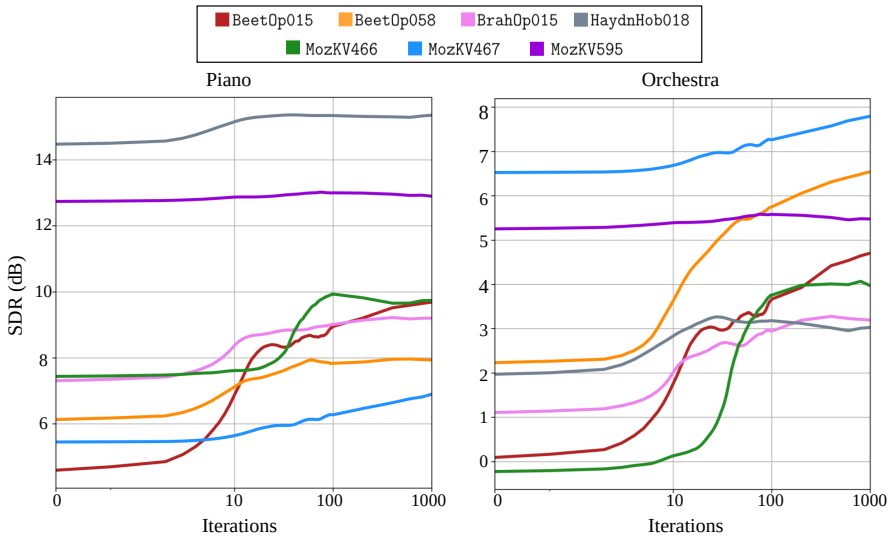

Figure 4.3: Evolution of SDR values based on our test dataset, applying TTA on the pre-trained model PT.

4.4.1 Test Dataset

For the evaluation of our models, we generate 30-second random mixes of piano-only and orchestra-only parts sampled from the annotated piano concertos (see Figure 4.2). These are different from the artificial training set, which we use for the pre-trained model, as they share harmonic and acoustic properties originating from the same recording. Note that we ensure that the samples used for training do not overlap with the test mixtures which we use for the evaluation purposes.

4.4.2 Quantitative Evaluation

To get a first impression of the performance of the models PT and TTA, we use the SDR [205] as our quantitative evaluation metric for the separation (see Section 7.3.3 for a detailed account of SDR-based evaluation). Table 4.2 provides a comparison of the resulting SDR values between a baseline for the SDR values (denoted as BL), which we compute using the test mixture as the target signal and ground-truth sources as the reference, pre-trained PT, and finetuned TTA after 100 iterations. One can observe that PT leads to a substantial improvement in SDR values compared to BL over the whole dataset, both for the piano and orchestra. It is interesting to observe that PT improves the SDR value of BeetOp015 for the separation of piano from 4.48 to 4.60, which is a relatively low improvement compared to other piano concertos in the dataset. Note that BeetOp015 is a historical recording (see Figure 4.1 for the recording year of the piano concertos), whose inadequate recording conditions may not be well represented in the random mixes used for the training of the PT, thus leading to a relatively poor separation performance.

Now, we focus on the comparison between PT and TTA. In general, the SDR-based results demonstrate that TTA enhances the separation of PT across all the piano concertos, for both the piano and the orchestra.

Table 4.2: Comparison of the SDR (dB) values between the baseline BL, and the separated sources by the pre-trained model PT and the finetuned model TTA after 100 iterations. The average SDR values are denoted with ϕ.

Work ID	Piano			Orchestra		
	BL	PT	TTA	BL	PT	TTA
BeetOp015	4.48	4.60	8.95	−4.36	0.09	3.67
BeetOp058	1.62	6.13	7.83	−1.58	2.23	5.75
BrahOp015	4.75	7.31	9.02	−4.60	0.09	3.67
HaydnHob018	10.99	14.47	15.34	−10.93	1.97	3.18
MozKV466	5.01	7.44	9.93	−5.06	−0.23	3.76
MozKV467	−0.72	5.45	6.28	0.73	6.52	7.26
MozKV595	6.64	12.74	13.00	−6.89	5.25	5.58
ϕ	4.67	8.31	10.05	−4.67	2.27	4.70

For example, in the case of BeetOp015, PT yields an SDR value of 4.60 for the separated piano. After finetuning with TTA for 100 iterations, this improves to 8.95. For the separated orchestra of BeetOp015, TTA also improves the SDR from 0.09 to 3.67. In the case of better quantitative separation results by PT, e. g., MozKV595, we observe that the improvement via TTA is relatively lower. Here, the SDR values improve from 12.74 to 13.00 for the piano and from 5.25 to 5.58 for the orchestra. Furthermore, our analysis reveals that the SDR value for the separated orchestra is generally lower than piano for both PT and TTA over the whole test dataset, except for MozKV467. An informal inspection states that the TTA leads to a significant improvement in the separation performance for historical recordings, which are not well-reflected in the training dataset of the pre-trained model PT.

In our next experiment, we investigate the performance of the finetuned models TTA per iteration. Figure 4.3 illustrates the evolution of the SDR values for each piano concerto in our test dataset. The overall convergence behavior exhibits a general trend of improvement of SDR values through TTA over PT for the separation of the piano and the orchestra. In particular, the SDR values for the separation of BeetOp015 depict a rapid improvement within the first 10 iterations. For the other piano concertos, the improvement of SDR values generally accelerates after the 10^{th} iteration. After the 100^{th} iteration, the separation performance remains steady for most of the piano concertos. Furthermore, after the 100^{th} iteration, the SDR values constantly increase in the case of BeetOp015 and MozKV467 for both piano and orchestra.

Although SDR is widely used as a quantitative evaluation metric for MSS, it is well known that it may not be suitable for determining the perceptual sound quality of separated musical sources [23]. The work by Torcoli et al. [195] provides a comparison of objective quality measures in the source separation domain. Their analysis indicates that a quantitative evaluation using the metric called *2f-score* exhibits the best correlation with ground-truth data based on the subjective ratings from MUSHRA listening tests. For a detailed account of the 2f-score, we refer to [92]. Note that the 2f-score values range from 0 to 100 following MUSHRA rating scores (see Section 4.4.3). Table 4.3 provides the resulting 2f-score values for the separated sources by PT and TTA using 100 iterations. In general, one can observe here a similar trend as for the SDR. PT mostly reveals better 2f-score scores than the baseline BL, except for the piano separation of BeetOp015, which presumably suffers from its poor recording conditions that are not well-represented in the artificial training set.

Table 4.3: Comparison of the 2f-score values between the baseline BL, and the separated sources using the pre-trained model PT and finetuned model TTA after 100 iterations. The average 2f-score values are denoted with ϕ.

Work ID	Piano			Orchestra		
	BL	PT	TTA	BL	PT	TTA
BeetOp015	21.60	21.50	28.39	15.51	25.85	29.52
BeetOp058	22.08	27.13	36.19	27.02	38.65	38.68
BrahOp015	24.01	30.79	36.43	22.63	35.36	33.20
HaydnHob018	19.27	34.57	38.31	27.10	41.19	40.35
MozKV466	19.25	32.30	39.49	26.07	35.62	40.18
MozKV467	15.61	28.80	31.52	28.43	40.26	41.21
MozKV595	14.88	27.82	31.52	18.08	30.36	31.49
ϕ	19.53	28.99	34.55	23.55	35.33	36.38

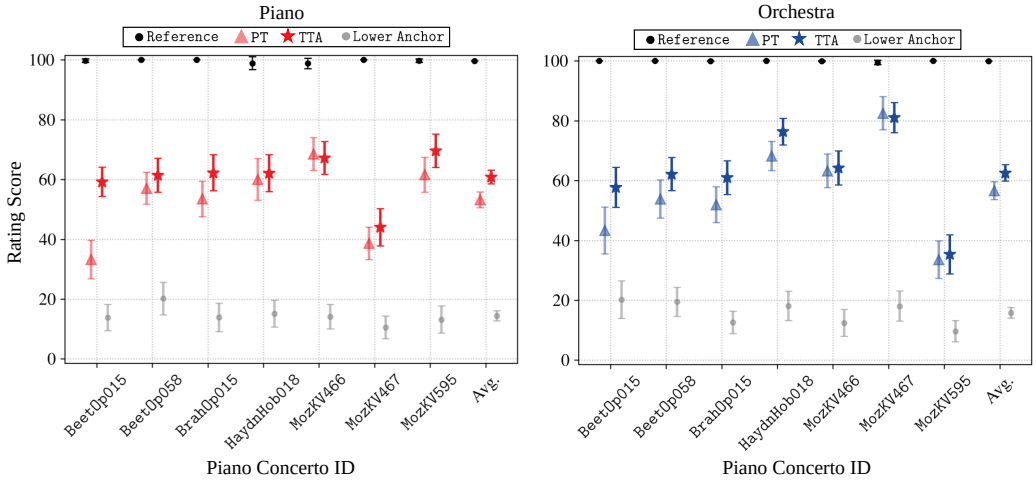

Figure 4.4: Results of our subjective evaluation based on MUSHRA listening tests for the piano (left) and orchestra (right). The colored markers indicate the average rating scores enclosed by 95% confidence intervals (indicated by the vertical lines).

As for the SDR-based results, 2f-score values increase via TTA after 100 iterations for both the piano and the orchestra. For example, in the case of BeetOp015, PT yields a 2f-score value of 21.50 for the separated piano, improving to 28.39 after applying TTA. Interestingly, the separation of the orchestral part yields better results than the piano according to 2f-score values. This is opposed to the evaluation based on the SDR, where the separation results are significantly better in the case of piano separation (see Table 4.2).

4.4.3 Subjective Evaluation

In this section, we describe the experimental setting for our subjective evaluation to assess the perceived separation quality. We carried out two listening tests using the MUSHRA methodology following the ITU-R BS.1534-3 recommendation [87]. It is a double-blind multi-stimulus test method with a hidden reference and an additional lower anchor signal. The rating scores range from 0 to 100, involving five categories: Bad (0-20), Poor (20-40), Fair (40-60), Good (60-80), and Excellent (80-100).

In total, 34 participants were involved in our listening tests (31 experienced listeners and 3 inexperienced listeners). The MUSHRA methodology suggests a post-screening of the participants stating that participants

should be excluded from the listening test if they give the hidden reference a score lower than 90 for more than 15% of the test items. Concerning our tests, one of the participants was excluded after post-screening.

Each of the two listening tests contains seven test items. For each test item, we generated four different signals with a duration of 12 seconds (maximum allowed signal duration for MUSHRA listening tests), which are excerpts from the test mixtures that we use for our quantitative evaluation. In our first listening test, we asked the participants to rate the overall audio quality for each of the four signals (also called *conditions*) with respect to a reference signal, which is a clean piano-only section. Similar to [52], we created the lower anchor signals by low-pass filtering the test mixtures to a 3.5 kHz cut-off frequency and by adding musical noise, i.e., randomly setting 20% of the remaining time/frequency coefficients to zero. The other two signals involve separated piano parts by `PT` and `TTA`. Similarly, our second listening test evaluates the overall quality of the separated orchestral parts following the same procedure as the first listening test.

Figure 4.4 summarizes the results from our listening tests. First, one can observe that the participants rated the reference signal with an average MUSHRA rating score of 99 and the lower anchor is significantly below the conditions `PT` and `TTA`. Remarkably, the general trend of the performances by `PT` and `TTA` support our quantitative analyses, inferring that the TTA procedure generally improves the separation of both the piano and orchestra. When observing the rating score of the piano concertos individually, one can observe that the rating of the historical recording `BeetOp015` is significantly lower than other items for `PT`. Intuitively, this is due to its poor recording conditions. After applying TTA, the average rating score of `BeetOp015` improves from 33 to 59 for the piano and from 43 to 58 for the orchestra. Furthermore, the orchestra separation led to a lower MUSHRA score in the case of `MozKV595`, both for `PT` and `TTA`. One reason may be the audible clipping artifacts in the reference signal and hidden separated orchestra, which a subset of the participants noted during the listening test.

As a final remark, the subjective results demonstrate that the average separation quality of the orchestra is better than for the piano, which is in favor of the results based on the 2f-score (see Table 4.3). This again illustrates that quantitative and subjective evaluations need to be carefully interpreted.

4.5 Conclusion

In this chapter, we investigated the separation of piano and orchestra in piano concertos. We trained our model using a U-Net architecture based on the Spleeter implementation with random mixes of solo piano and orchestral recordings, which we regarded as our baseline pre-trained model. We proposed a TTA procedure to enhance the separation quality using the random mixes created from the samples found in the test data. In particular, we showed that TTA substantially improved the quantitative and subjective evaluation results, both for the piano and orchestra. The test data used for evaluation were based on synthetic mixes of piano-only and orchestra-only sections from piano concerto recordings. For a

more realistic and fair evaluation, we introduce a multitrack dataset of piano concertos in the following chapter (Chapter 5), which involves a collection of excerpts with separate piano and orchestral tracks from piano concertos ranging from the Baroque to Post-Romantic era. Furthermore, the considered scenario in this chapter indicates the potential of musically motivated mixing approaches for training and finetuning deep neural network (DNN) models. In Chapter 6, we provide a more thorough exploration of musically motivated data augmentation methods that simulate more realistic mixtures.

5 A Multitrack Dataset of Piano Concertos

This chapter is based on [136]. Yigitcan Özer is the main contributor to this article's recording process, dataset creation, and writing. Simon Schwär was the sound engineer responsible for the recording and post-production process. Emre Sen, Jeremy Lawrence, Yigitcan Özer, and Meinard Müller were the main performers of the dataset. Vlora Arifi-Müller contributed to the dataset curation. Meinard Müller closely supervised this work and contributed with Simon Schwär to the article's writing.

The piano concerto is a genre of central importance in Western classical music, often consisting of a virtuoso solo part for piano and an orchestral accompaniment. In this chapter, we introduce the Piano Concerto Dataset (PCD), which comprises a collection of excerpts with separate piano and orchestral tracks from piano concertos ranging from the Baroque to the Post-Romantic era. In particular, using existing backing tracks by the music publisher Music Minus One (MMO), we recorded excerpts from 15 different piano concertos played by five interpreters on various instruments under different acoustic conditions. The key challenge of playing along with pre-recorded orchestral accompaniments lies in the exact synchronization of the performer. For guiding the pianists for obtaining a high synchronization accuracy, we used additional click tracks generated with measure and beat annotations of the orchestral tracks, which also are provided in the PCD. The dataset is relevant for a variety of MIR tasks, including music source separation, automatic accompaniment, music synchronization, editing, and upmixing.

The remainder of the chapter is organized as follows. Following the introduction in Section 5.1, in Section 5.2, we give an overview of the existing multitrack datasets. In Section 5.3, we address the role and significance of piano concertos in Western classical music and review their form and compositional structure. As the main contribution of this chapter, we describe the content of PCD in Section 5.4 and outline its recording process and challenges. In Section 5.5, we describe the different interfaces for accessing the dataset. Finally, we conclude in Section 5.6 with prospects on the potential applications of the PCD.

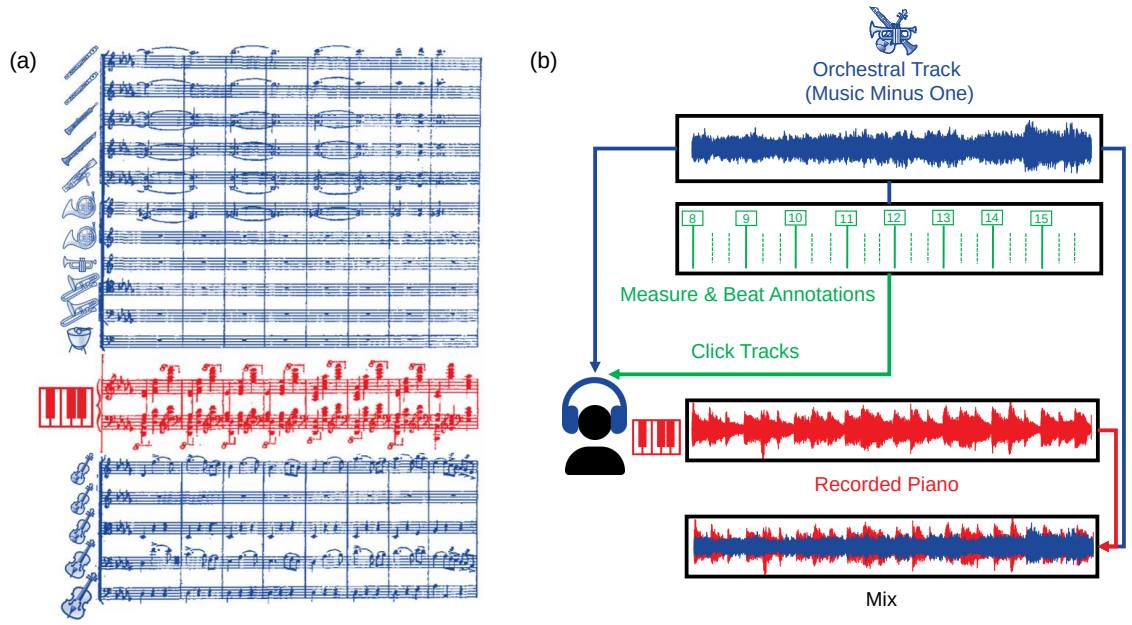

Figure 5.1: Overview of the recording process. **(a)** The sheet music of measures 8–15 from Tchaikovsky's Piano Concerto No. 1 in B Flat Minor, Op. 23, 1st movement. **(b)** During the recording process, the pianist is supposed to play synchronously with the backing track. In a real-life recording process, 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 is a difficult task for the performer. To address this challenge, they listen to metronome-like click tracks sonified using measure (solid green) and beat (dashed green) annotations in addition to the orchestral accompaniments. As the result of a final mastering step, PCD comprises dry and reverberant *synchronous* recordings of piano and orchestra accompaniments selected from 15 different piano concertos and their mixes.

5.1 Background

Data-driven models for MSS typically require clean, isolated target sources (also called *stems*) for training and evaluation. Research in MSS mainly focuses on separating vocals, bass, and drums from mixtures of popular music songs, mainly due to the availability of multitrack datasets such as MUSDB18 [151] for this task. Separating classical music recordings into individual sound sources has also recently received attention [19, 29, 30, 75, 121, 131, 143, 163]. Compared to popular music, the constituent sources of classical music recordings often reveal a higher spectro–temporal correlation, which makes the separation task more challenging. Furthermore, the availability of sufficiently large multitrack datasets for Western classical music is a limiting factor for research in this area. In this chapter, we introduce a novel multitrack dataset called PCD, which enables both quantitative and subjective evaluation for separating piano concertos.

PCD contains 81 excerpts of multitrack piano and orchestra recordings, each having a duration of 12 seconds. These are selected from 15 different piano concertos from the Baroque to the Post-Romantic period. The

variety in the works' complexity, the recordings' acoustical settings, its orchestral instrumentation, and five different performers contributes to the diversity of PCD.

The piano concerto is an essential genre in Western classical music from the Baroque era onward. These compositions are generally written for pianists to demonstrate their virtuosity. Furthermore, the piano concerto has a rich and dynamic sound that is distinctive to this type of music, characterized by opposing musical elements [35]. Besides a large number of compositions throughout music history, classical music archives comprise numerous prominent historical, public-domain recordings of piano concertos, which can be useful for various applications in MIR, including source separation, editing, and upmixing [117], music alignment [57, 146], automated accompaniment [2, 39], and audio decomposition [58]. We refer to Section 6.2 for a more detailed account on related work for general source separation.

To create a multitrack dataset of piano concertos, we use MMO[3]. MMO provides recordings of backing tracks, in which the lead instrument or the vocal part is omitted, typically the soloist. This allows musicians to practice or perform the solo part along with the pre-recorded accompaniment in case they do not have access to other musicians to play with them. The main difficulty of performing with a pre-recorded accompaniment lies in the absence of any interaction between the player and other musicians. This is particularly problematic for classical music since interpretations can vary greatly in terms of tempo and dynamics. Moreover, piano concertos often contain long sections which only involve orchestral accompaniment. The lack of guidance for the pianist, as typically provided by a conductor, can result in being asynchronous or missing the cue after a long rest. To address this issue, we annotated the measure and beat positions of the backing track of each piano concerto. During the recording sessions, the pianists simultaneously listened to the orchestral accompaniments and sonified click tracks, which were generated using these annotations. Notably, in case of abrupt tempo changes or long piano-solo sections, the additional click tracks have proven helpful for the interpreters while playing along with the pre-recorded accompaniments. Figure 5.1 displays an excerpt from the Piano Concerto in B Flat Minor, Op. 23, 1st Movement by Peter Ilyich Tchaikovsky and its recording process. The recording sessions are followed by post-production for generating cohesive mixtures of newly recorded piano tracks and existing MMO accompaniments. As a main contribution of PCD, we provide dry and reverberant recordings of piano and orchestra stems and their mixtures.

5.2 Related Work

For the training and evaluation of data-driven models, datasets constitute an essential component of MIR research. In particular, the availability of multitrack datasets has led to impressive results of data-driven MSS approaches that focus on separating popular music recordings. In the Western classical music domain, several datasets have been introduced for polyphonic vocal music [36, 118, 158, 169], which

[3] https://www.halleonard.com/series/MMONE

Table 5.1: Multitrack instrumental datasets in the Western classical music domain, indicating the number of recordings (#R) and their total duration (Dur) in hh:mm:ss.

Name & Author	# R	Dur
WWQ [4]	1	00 : 09 : 00
TRIOS [144]	5	00 : 03 : 12
PHENICX [166]	4	00 : 10 : 36
Bach10 [50]	10	00 : 05 : 30
URMP [106]	44	01 : 36 : 00
EnsembleSet [163]	9	01 : 03 : 34
PCD	81	00 : 16 : 12

comprise isolated recordings of vocal ensembles. For instrumental music, however, there are only a few multitrack datasets (see Table 5.1). Bay et al. [4] presented the Woodwind Quintet (WWQ) dataset, which includes separate tracks of a woodwind adaptation of Beethoven's String Quartet, Op. 18 No. 5. The TRIOS dataset [144] involves multitrack recordings of four classical pieces and one jazz piece, as well as their transcriptions. The PHENICX-Anechoic dataset [166] comprises annotations and audio material of anechoic multitrack recordings of four orchestral works, which differ in terms of the number of instruments per instrument class. Bach10 [50] consists of multitrack recordings of ten chamber music pieces where each work comprises four parts (SATB) played by violin, clarinet, saxophone, and bassoon. Li et al. [106] introduced the University of Rochester Multimodal Performance (URMP) dataset, which addresses the music performance as a multi-modal art form and provides the musical score, as well as the audio recordings of the individual stems of 44 ensemble pieces. Their work also describes the challenges of maintaining synchronization and musical expressiveness while creating a multitrack dataset of classical music pieces. Sarkar et al. [163] presented the EnsembleSet, which consists of synthesized multitrack recordings of strings, woodwind instruments, and brass, generated by using MIDI files from RWC [72] and Mutopia[4]. For an overview of a variety of publicly available datasets in MIR, please refer to [11][5].

5.3 Piano Concertos in Western Classical Music

As it is a central theme in PCD, we highlight in this section the compositional structure and evolution of piano concerto as a genre of central importance in Western classical music. A piano concerto is a musical composition written for piano and orchestra. It typically consists of multiple movements, with the piano playing the primary role and the orchestra providing the accompaniment. Since the Baroque period, piano concertos have been composed by many composers from all epochs until today. As a result, piano concertos are an enduring and popular form of classical music and continue to be enjoyed by audiences around the world.

In the seventeenth century, the earliest use of the term *concerto* in Western classical music referred to its literal meaning *combined effort*. The "combined effort" sense persisted until Johann Sebastian Bach, whose keyboard concertos depend on the *reconciliation* of cembalo or harpsichord and other

[4] https://www.mutopiaproject.org
[5] https://mirdata.readthedocs.io/en/latest/

instruments [35]. One has to consider that in J. S. Bach's time, the keyboard did not yet have the status of a virtuoso instrument as it does today. When it appeared in association with other instruments, it was initially associated with the term *continuous bass instrument* [167]. Nowadays, pianists often perform baroque keyboard concertos on the modern piano.

Whereas the high Baroque period cultivated various kinds of concertos, the *solo concerto*, which comprises a lead instrument accompanied by an orchestra, emerged as the preeminent type of this form in the high Classical period. In the late eighteenth century, the classical concerto evolved to an independent form, incorporating form-functional elements associated with the Baroque period, e. g., the ritornello, and the Classical period, e. g., the classical sonata form [26]. The pioneers of the Vienna Classic, Haydn, Mozart, and Beethoven, wrote piano concertos that involve a dialogue between orchestra and solo instrument [167].

During the course of the nineteenth century, romanticism brought a new interest in orchestral color, and composers explored a variety of sounds obtained by closely intertwining the solo instrument and the orchestra. Additionally, the piano had grown in tonal capabilities compared to its usage in the Baroque and Classical periods. As a result, romantic piano concertos diverged from the classical form [69]. For example, the focus of the interaction between orchestra and piano shifted in favor of the soloists in the case of piano concertos by Frédéric Chopin. In contrast, while renouncing the mere virtuoso display of the soloist, Robert Schumann's Piano Concerto in A minor, Op. 54, is considered a masterpiece of thematic and melodic integration of piano and orchestra. Romanticism reached a climax in Brahms' piano concertos, interchangeably splitting the themes between orchestra and the piano. Finally, the virtuoso style in the Romantic period witnessed its best examples in Tchaikovsky's famous first piano concerto (Op. 23), but even more by the post-romanticism in the piano concertos by Rachmaninov.

5.4 Piano Concerto Dataset (PCD)

In this section, we present the PCD as our main contribution of this chapter. In Section 5.4.1, we cover the details of the musical content and characteristics of the dataset. We define the naming conventions of the included files in Section 5.4.2. In Section 5.4.3, we explain our approach for the alignment of pianists with the pre-recorded orchestral tracks. We elaborate on the recording process in Section 5.4.4, describe the required pre-processing steps in Section 5.4.5, and finally the post-production in Section 5.4.6.

5.4.1 Dataset Content and Characteristics

This section describes several aspects concerning the content and characteristics of PCD. The dataset consists of 81 excerpts selected from 15 piano concertos by 10 different composers, as shown in Table 5.2. Here, the *WorkID* specifies the prefix of each filename in the dataset encoding the composer, assigned work number (i. e., Op, BWV, and KV), and the movement, respectively. For further information on the

WorkID	Composer	Full Name	Mvm	PID	#V	#E	Dur
Bach_BWV1056-01	J. S. Bach	Piano Concerto in F Minor, BWV 1056	1	YO	2	10	120
Beethoven_Op015-01	Beethoven	Piano Concerto No. 1 in C major, Op. 15	1	MM	1	6	72
Beethoven_Op019-01	Beethoven	Piano Concerto No. 2 in B Flat Major, Op. 19	1	ES	2	4	48
Beethoven_Op037-01	Beethoven	Piano Concerto No. 3 in C Minor, Op. 37	1	ES	2	4	48
Beethoven_Op037-02	Beethoven	Piano Concerto No. 3 in C Minor, Op. 37	2	LR	1	1	12
Beethoven_Op058-02	Beethoven	Piano Concerto No. 4 in G Major, Op. 58	2	ES	2	2	36
Chopin_Op021-03	Chopin	Piano Concerto No. 2 in F Minor, Op. 21	3	ES	1	5	60
Grieg_Op016-01	Grieg	Piano Concerto in A Minor, Op. 16	1	ES	1	1	12
Mendelssohn_Op025-01	Mendelssohn	Piano Concerto No. 1 in G Minor, Op. 25	1	ES	2	2	24
Mozart_KV414-01	Mozart	Piano Concerto No. 12 in A Major, KV.414	1	YO	1	2	24
Mozart_KV467-01	Mozart	Piano Concerto No. 21 in C Major, KV.467	1	YO	1	5	60
Mozart_KV467-02	Mozart	Piano Concerto No. 21 in C Major, KV.467	2	YO	2	6	72
Rachmaninoff_Op018-01	Rachmaninov	Piano Concerto No. 2 in C Minor, Op. 18	1	JL	1	5	60
Rachmaninoff_Op018-02	Rachmaninov	Piano Concerto No. 2 in C Minor, Op. 18	2	JL	1	5	60
Rachmaninoff_Op018-03	Rachmaninov	Piano Concerto No. 2 in C Minor, Op. 18	3	JL	1	5	60
Rachmaninoff_Op030-01	Rachmaninov	Piano Concerto No. 3 in D Minor, Op. 30	1	ES	2	6	72
Saint_Op022-01	Saint-Saëns	Piano Concerto No. 2 in G Minor, Op. 22	1	ES	1	2	24
Schumann_Op054-01	Schumann	Piano Concerto in A Minor, Op. 54	1	ES	2	4	48
Tschaikovsky_Op023-01	Tchaikovsky	Piano Concerto No. 1 in B Flat Minor, Op. 23	1	ES	2	6	72
					Σ	81	972

Table 5.2: Overview of the dataset indicating the work identifier (WorkID), composer, full name of the work, movement (Mvm), performer identifier (PID), number of versions (#V), number of excerpts (#E), and total duration in seconds (Dur). The versions here refer to distinct performances recorded under different acoustic conditions and played on different pianos.

naming conventions of the audio and annotation files in the dataset, see Section 5.4.2. In addition to various compositional styles ranging from the Baroque to the Post-Romantic era, PCD includes different difficulty levels of piano concertos. For example, J. S. Bach's Piano Concerto in F Minor, BWV 1056, is classified as moderately difficult, whereas Rachmaninov's Piano Concerto No. 3 in D Minor, Op. 30, is a very challenging virtuoso work for pianists.

Although we recorded longer sections, including the exposition, development, or sometimes entire movements of piano concertos, we decided to extract and provide only shorter excerpts of the recordings for several reasons. First, practicing and performing entire movements can be difficult for pianists. Second, it is time-consuming for sound engineers to edit and process longer recordings. Third, depending on the compositional style, piano concertos may involve long sections where the piano and orchestra do not play together, which does not serve the multitrack dataset. We will make the raw piano recordings (also of longer sections) available, upon request.

The choice of excerpts is mainly based on musical coherence and a balance between piano and orchestra. Besides passages where both parts play together, we also included sections where the piano and orchestra follow a conversational style, such as in Beethoven's Piano Concerto No. 4 in G Major, Op. 58. In order to account for a suitable duration of the excerpts, we regarded two guidelines. First, the excerpts need to be long enough to involve a complete musical phrase. Second, they should be relatively short for their usability in a subjective listening test. Based on these criteria, we decided on a duration of 12 seconds.

Room ID	Room Description	Piano	#E	Dur
R1	Lecture Hall (Fraunhofer IIS)	Yamaha C3	15	180
R2	Private Studio (Jeremy Lawrence)	Yamaha C3X	15	180
R3	Music Academy (Emre Şen)	Seiler	21	252
R4	Saygun Concert Hall (Bilkent University)	Steinway D	30	360
		Σ	81	972

Table 5.3: Overview of different rooms where the recordings took place, number of excerpts recorded in each room (#E), and their total duration in seconds (Dur). Note that the piano model is different in each acoustic environment.

This audio length has been a good compromise for musicality while being the longest recommended duration for Multiple Stimulus with Hidden Reference and Anchors (MUSHRA) listening tests [87].

For a wide range of interpretations, five pianists participated in the curation of the dataset: Emre Şen (ES), Jeremy Lawrence (JL), Lisa Rosendahl (LR), Meinard Müller (MM), and Yigitcan Özer (YO). All the performers have provided their consent to publish the recorded material for research purposes under a Creative Commons license. The performers' skills range from amateurs (LR, MM) to semi-professional players (JL, YO) to a concert pianist (ES), and their experiences differ accordingly. LR is a historian and musicologist, and MM is a full-time professor in MIR, playing the piano as a hobby. Among the semi-professional performers, JL is a Master's student in electrical engineering with a strong musical background and experience as a pianist, whereas YO is a Ph.D. candidate working on MIR, with a background in electrical engineering and piano performance. ES is a concert pianist who regularly performs recitals and plays piano concertos with orchestras.

Furthermore, the room acoustics vary among the performances, ranging from a small and relatively dry domestic space (R2), via small recital halls (R1 and R3), to a spacious concert hall environment (R4). Each room is also associated with a different grand piano model. Table 5.3 summarizes the differences in recording conditions for each room.

In addition to distinct acoustic conditions, PCD includes recordings that vary in quality and orchestral accompaniments. The recordings of Rachmaninov's Piano Concerto No. 2 in C Minor, Op. 18, performed by JL are considered the highest quality recordings in the dataset. These performances were recorded in multiple sessions and underwent exhaustive post-processing. Moreover, this is the only instance where the orchestral accompaniment is synthetic (as provided by MMO), whereas other backing tracks are real recordings. Note that we also provide recordings of multiple movements from the same piece for three works: Beethoven's Piano Concerto No. 3 in C Minor, Op. 37, Mozart's Piano Concerto in C Major, KV.467, and Rachmaninov's Piano Concerto No. 2 in C Minor, Op. 18. Furthermore, there are two versions of certain excerpts, providing different piano recordings using the same underlying orchestral accompaniments.

To gain a more comprehensive understanding of the statistics of the dataset, the distribution of the number of pieces per composer is presented in Figure 5.2a. In PCD, Rachmaninov is the most prominent composer,

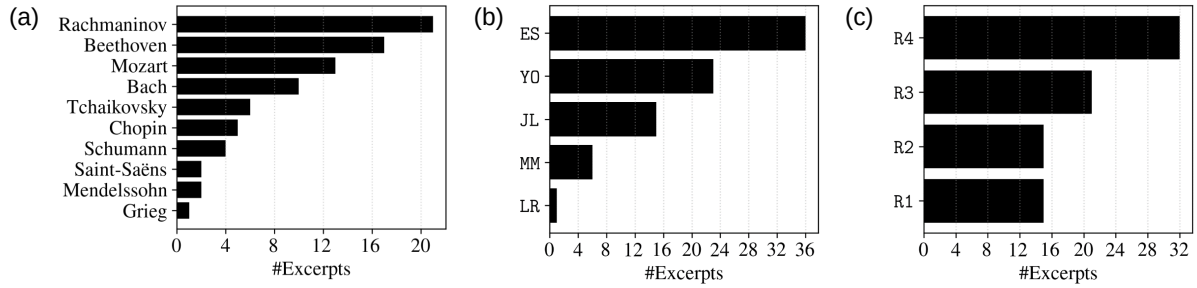

Figure 5.2: Various bar plots describing the dataset. The number of selected 12-second excerpts is indicated by the horizontal axis per **(a)** composer, **(b)** performer, and **(c)** acoustic environment.

with 21 excerpts and a total duration of 252 seconds. Beethoven comes in second place, with 17 excerpts, followed by Mozart, J. S. Bach, and Tchaikovsky. Figure 5.2b provides an overview of the number of excerpts played by each performer. Most of the performances are by the concert pianist ES. Note that several pieces were performed by the same performer in different rooms. For example, ES performed Tchaikovsky's Piano Concerto No. 1 in B Flat Minor, Op. 23 both in R3 and R4. Figure 5.2c illustrates the number of excerpts per room. The majority (32) of the recordings took place in R4, roughly a quarter of them (21) in R3, 15 in R1, and 15 in R2.

5.4.2 Naming Conventions

PCD offers a variety of musical dimensions as summarized in Table 5.4. These dimensions, referred to as ComposerID, WorkNo, MeasRange, PID, VersionID, StemType, and Reverb, are used in the filenames of the provided WAV audio files. The ComposerID specifies the composer (see Figure 5.2a). WorkNo indicates the Opus, BWV, or KV number of the work, and MovementNo denotes the number of the movement from which the excerpt was selected. The MeasRange dimension specifies range of the excerpt in measures. PID identifies the performer, as introduced in Section 5.4.1, and VersionID the version. StemType refers to the post-processing configurations presented in Section 5.4.6. Reverb refers to the presence of artificial reverb added in the post-production. The audio filename using the instances in the **Example** column in Table 5.4 is `Bach_BWV1056-01-mm001–008_YO-V2_OP_reverb`.wav. It represents Bach's (ComposerID) Piano Concerto in F Minor, BWV 1056 (WorkNo), 1st Movement (MovementNo), Measures 1–8 (MeasRange), played by YO (PID), second version (VersionID), which includes piano part plus orchestral accompaniment (StemType) with artificial reverb (Reverb).

5.4.3 Synchronization

Similar to the development of other multitrack datasets, the PCD curation encountered several challenges regarding the alignment of separate tracks. The missing interaction between the performer and other musicians constitutes a key challenge in a multitrack recording setting. As Li et al. [106] suggest,

Table 5.4: PCD dimensions, encoded in filenames.

Dimension	Description	Example
ComposerID	Composer Identifier	`Bach`
WorkNo	Op. / BWV / KV	`BWV1056`
MovementNo	Movement Number	`01`
MeasRange	Measure Range	`mm001–008`
PID	Performer Identifier	`YO`
VersionID	Version Identifier	`V2`
StemType	(O)rchestra / (P)iano	`OP`
Reverb	Presence of Reverb	`reverb`

audio-visual cues may help the musicians when playing along with a given audio track. In the recording process of the PCD, we used only audio cues, which served as a guide for the performers alongside the pre-recorded orchestral accompaniments.

The main objective of PCD is to provide piano recordings, which are synchronous to the *original* backing tracks by MMO. This design choice enables the dataset's reproducibility while restricting the freedom of interpretations since the pianists *must* steadily adapt their tempo to the orchestral track. To overcome the challenges posed by the recording settings, we provided metronome-like click tracks in the form of sonified measure and beat annotations. We first manually annotated the measure positions in the backing track where the orchestra is active. Note that piano concertos often involve relatively long piano solo passages. In these sections, we employed linear interpolation to estimate the measure positions. This approach guaranteed that the sonified metronome-like click tracks retained consistent tempo in sections where the backing track is silent.

For the beats, we initially experimented with manual annotations. However, we found that using manual annotations based on the orchestral accompaniments was ineffective for the performers, as the tempo changes within measures were often inconsistent. As an alternative, we again utilized linear interpolation to estimate the beats within the manually annotated measure positions based on the time signature of the piece. This approach resulted in equidistant beats interpolated between the manual measure annotations, which were more helpful for the pianists than manual beat annotations.

Only for the recordings of Rachmaninov Piano Concerto No. 2 in C Minor, Op. 18, we adopted a more involved iterative approach for the generation of beat annotations. For example, piano-only sections meant to be played with rubato (rather than a consistent tempo) were annotated by the performer such that the click tracks would match the tempo fluctuations in their interpretation. This facilitated a more natural-sounding recording of solo sections with larger variations in tempo.

Finally, we sonified measure and beat positions with different frequencies to aid the pianists during the recording process. Depending on the preference of the musician, we either activated or deactivated the click tracks during recording to allow for more agogical playing.

Figure 5.3: An impression from the recording process in R4 with stereo spot microphones (2× *Schoeps MK4*). To synchronously play with the orchestra, the performer listens to the MMO orchestral accompaniment via *Beyerdynamic DT 770 Pro* headphones (superimposed by click tracks).

5.4.4 Recording Process

In this section, we outline the technical details about the recording process. The performances in rooms R1, R3, and R4 were recorded using a stereo spot microphone with *Schoeps MK4* cardioid microphones, placed near the bend of the grand piano body (see Figure 5.3). Comparable high-end microphones like the MK4 are often used in similar professional recording setups. The exact position of the microphones was individually adjusted to the acoustics of each recording space and the characteristics of the instrument, roughly following an *ORTF (Office de Radiodiffusion Télévision Française)* setup. The microphone signals were recorded using a *RME Babyface Pro FS* audio interface and the *REAPER*[6] digital audio workstation. Recordings were initially stored in the *WAV* format with a sampling rate of 44.1 kHz and 24 bits per sample. The orchestral accompaniment and sonified click tracks were presented to the musicians via headphones (*Beyerdynamic DT 770 Pro*) and played back from the same *REAPER* session to ensure a synchronous recording of the piano part. The pianists had the possibility to record the movements in shorter segments and repeat individual sections, as is common in a studio recording process. This typically results in multiple takes for the same section, which are later edited to form a coherent performance (see Section 5.4.6). The performances in room R2 (the recordings of Rachmaninov's Piano Concerto No. 2 in C Minor, Op. 18) were captured in a similar fashion, only differing in the utilized equipment. These performances were recorded using a stereo pair of *Sennheiser MKH 8020* omnidirectional microphones in

6 https://www.reaper.fm/

AB configuration with a spacing of 35 cm. The microphones were placed approximately 1 m from the bend of the grand piano body at the height of 145 cm. A *Steinberg UR22mkII* audio interface and the *Cubase* [7] digital audio workstation were used.

5.4.5 MMO Pre-Processing

The backing tracks provided by MMO vary in recording quality and format. To provide consistent orchestral accompaniments suitable for recording the piano parts, we modified the original tracks in several ways.

First, some of the MMO recordings were split into multiple sections (e. g., including just one page of sheet music). To have a backing track of the entire movements, we joined the audio files, which belong to the same movement. The resulting backing tracks are single audio files that serve as a continuous reference timeline for the dataset. Furthermore, we removed audible waveform artifacts at the splitting points. Finally, we removed the silence at the beginning and end of CD audio files. Note that this results in a shorter total duration of the reference timeline than the sum of the MMO tracks. All timings in the provided annotations and documentation refer to the reference timeline of the backing tracks created in this process. We conducted all the modifications with Python scripts, which allows for reproducing our backing tracks from original MMO files.

Second, we removed some clicks in the backing track, provided in MMO in pauses of the orchestral accompaniment where the pianist plays solo. In the rendered excerpts, all clicks are always deactivated. This applies to the backing tracks of `Bach_BWV1056-01`, `Beethoven_Op037-01`, `Beethoven_Op058-02`, `Mendelssohn_Op025-01`, `Mozart_KV414-01`, `Rachmaninoff_Op018-01`, `Rachmaninoff_Op018-02`, `Rachmaninoff_Op018-03`, and `Schumann_Op054-01`.

Third, we finally employed some additional cosmetic pre-processing, including the removal of background noises (using the *iZotope RX8 Audio Editor*) and equalization for more consistent timbral qualities between pieces.

5.4.6 Post-Production

The post-production of the recorded performances was conducted in three steps. First, we edited the recorded takes in *REAPER* to create a coherent rendition of the piano part. The takes were chosen to reduce playing mistakes while still maintaining a consistent musical arc in the performance, similar to post-production in a recording studio. Note that we maintained the timeline of the backing track in the post-production. Only the piano recording was edited to achieve good synchronicity with the MMO orchestral accompaniments.

[7] `https://www.steinberg.net/cubase/`

Second, equalization was applied to the piano recordings to ensure consistent timbral qualities within our dataset without overcompensating the differences between instruments and recording spaces. Some minor noise removal similar to the MMO pre-processing was necessary to remove background noises. Third, to increase the coherence between the piano part and orchestral accompaniment, we applied artificial reverberation to both tracks simultaneously using the *FabFilter Pro-R2* algorithmic reverb software. All tracks are available with and without artificial reverberation to facilitate different use cases (see below). For the dataset, the post-processed excerpts were exported as *WAV* files with 44.1 kHz sampling rate and 16 bits per sample in six different configurations:

- `OP_reverb`: Piano part plus orchestral accompaniment with artificial reverb
- `OP`: Piano part plus orchestral accompaniment without artificial reverb
- `P_reverb`: Piano part only with artificial reverb
- `P`: Piano part only without artificial reverb
- `O_reverb`: Orchestral accompaniment only with artificial reverb
- `O`: Orchestral accompaniment only without artificial reverb

During the recordings of `Beethoven_Op015-01`, and `Mozart_KV467-01`, the orchestral accompaniment was erroneously played back with a rate of 0.995 (`Beethoven_Op015-01`) and 1.005 (`Mozart_KV467-01`), which results in a slightly slower or faster piano part relative to the backing track with the original playback speed. This mistake was corrected in the post-production with *Elastique Pro v3.3.3* by applying time-scale modification (with rates 1.005458 and 0.994616, respectively).

5.5 PCD Interfaces

The main motivation of PCD is to provide a freely available and well-documented multitrack dataset to support MIR research on orchestral music, particularly piano concertos. To this end, the dataset is made publicly accessible through different interfaces in order to support scientific exchange and ensure the reproducibility of scientific results.

Interactive interfaces can lower barriers to access datasets and research results. This can be achieved through features such as playback functionalities [67, 89, 160]. To provide an interactive medium for the researchers, we use an open-source audio player [210] integrated in a web interface, which allows the listener to switch between multiple audio tracks while synchronously indicating the playback position of the audio tracks. As default visualization, the interface offers an overview of the six configurations of stems, as presented in Section 5.4.6. The main page is subdivided into a section called *Excerpts*, which includes links to recorded piano concertos sections with a dedicated sub-page for each excerpt. Figure 5.4

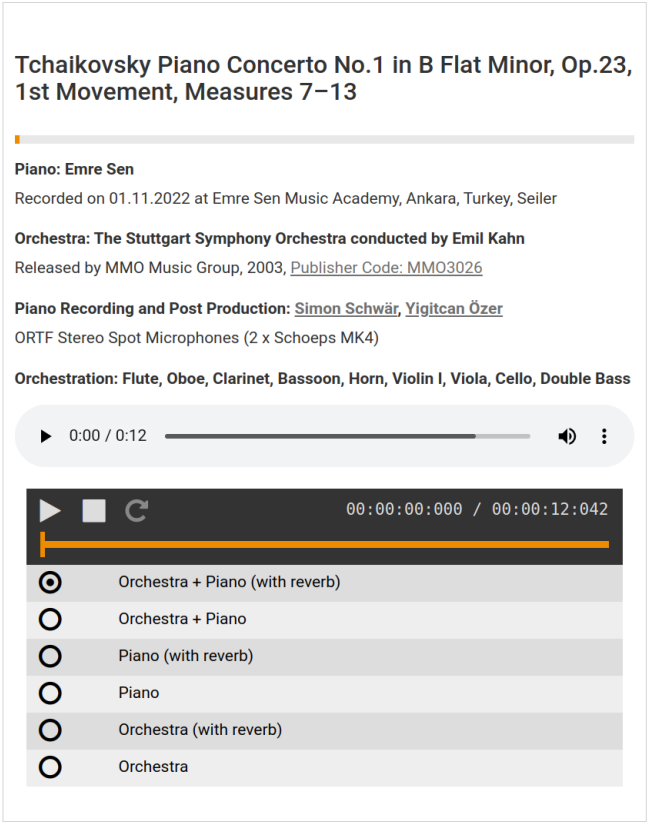

Figure 5.4: Screenshot of our web-based interface with a Track Switcher [210], which comprises six tracks of dry and reverberant recordings of an excerpt from Tchaikovsky's Piano Concerto No. 1 in B Flat Minor, Op. 23, 1st movement.

shows a screenshot of an exemplary sub-page[8], which hosts the multitrack audio files for an excerpt selected from Tchaikovsky's Piano Concerto No. 1 in B Flat Minor, Op. 23, 1st movement.

5.6 Conclusion

In this chapter, we introduced the PCD, which comprises excerpts from piano recordings and orchestral accompaniments of piano concertos ranging from the Baroque to the Post-Romantic era. Using backing tracks from the music publisher MMO, we recorded 15 different piano concertos played by five performers with different instruments under varying acoustic conditions. To address the challenge of precise synchronization with pre-recorded orchestra accompaniments, we created click tracks to guide the pianists during the recording process. As a main contribution of PCD, we provide 81 excerpts of dry and reverberant recordings of piano and orchestra stems and their mixtures. We release the dataset via an interactive web-based interface to provide a convenient access. Diverse musical dimensions of PCD enable various applications for MIR research, particularly for quantitative and subjective evaluation of source separation models.

[8] `https://www.audiolabs-erlangen.de/resources/MIR/PCD/`

6 Source Separation with Musically Motivated Augmentation Techniques

This chapter is based on [132]. The first author Yigitcan Özer is the main contributor to this article. In collaboration with his supervisor Meinard Müller, he devised the ideas, developed the formalization, and wrote the paper. Furthermore, Yigitcan Özer implemented all approaches and conducted the experiments.

Being a genre written for a pianist typically accompanied by an ensemble or orchestra, piano concertos often involve an intricate interplay of the piano and the entire orchestra, leading to high spectro–temporal correlations between the constituent instruments. Moreover, in the case of piano concertos, the lack of multitrack data for training constitutes another challenge in view of data-driven source separation approaches. As a basis for this chapter, we adapt existing DL techniques, mainly used for the separation of popular music recordings. In particular, we investigate spectrogram- and waveform-based approaches as well as hybrid models operating in both spectrogram and waveform domains. As a main contribution, we introduce a musically motivated data augmentation approach for training based on artificially generated samples. Furthermore, we systematically investigate the effects of various augmentation techniques for DL-based models. For our experiments, we use the PCD, which is open-source dataset of multitrack piano concerto recordings, introduced in Chapter 5. Our main findings demonstrate that the best source separation performance is achieved by a hybrid model when combining all augmentation techniques.

The remainder of the chapter is organized as follows. Following the introduction in Section 6.1, in Section 6.2, we discuss the relevant work on source separation. We then revisit in Section 6.3 the architecture and characteristics of four different networks, which we adapt for our application scenario. In Section 6.4, we introduce our musically motivated data augmentation approaches. Then, in Section 6.5, we describe the experimental settings and our design choices and report on the quantitative empirical results, including a subjective evaluation. Finally, in Section 6.6, we conclude with prospects on future work.

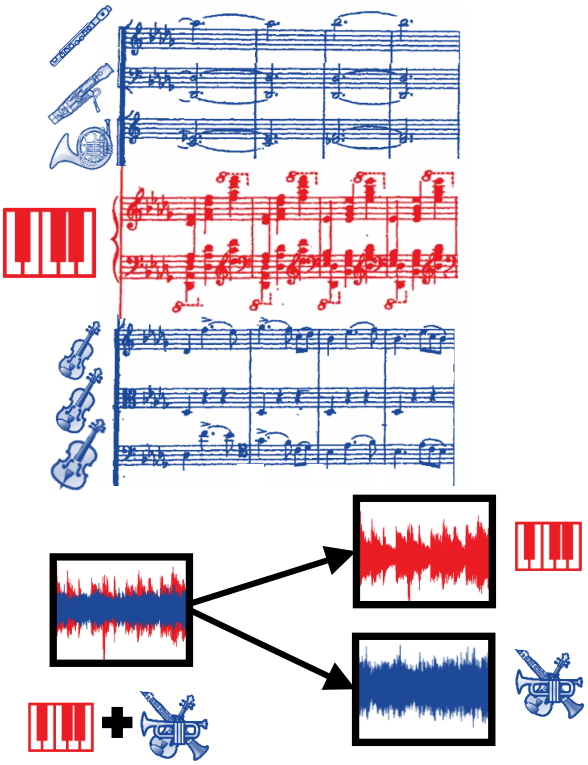

Figure 6.1: An excerpt from Tchaikovsky's Piano Concerto No. 1 in B Flat Minor, Op. 23, 1st Movement. Our goal is to decompose piano concertos into the piano (red) and orchestral (blue) tracks using data-driven MSS techniques.

6.1 Background

Motivated by the need for orchestral accompaniments of amateur or semi-professional pianists, we consider the novel task of separating piano concertos building on Chapter 4, which we substantially extend in this chapter, particularly through the adaptation of four deep learning (DL) models. For an illustration of the task, see Figure 6.1.

MSS aims at separating individual musical sound sources from a recording that contains multiple instruments or voices. Generally, a musical source may refer to singing, an instrument, or an entire group of instruments such as an ensemble or orchestra. The practical importance of separating these individual sources from a sound mixture can be seen in diverse applications, such as creating karaoke systems, aiding in music production, facilitating music transcription, and supporting music analysis. However, MSS poses a significant challenge due to strong spectro–temporal correlations between different sound signals within a music recording [24]. In this context, DNNs have led to substantial improvements in separating and isolating musical sources, see, e. g., [42, 43, 83, 88, 97, 114, 159, 178, 181, 185, 190].

Supervised deep learning models addressing the MSS task typically require a large dataset that consists of multitrack recordings containing the individual stems of the various musical sources. Because

of the availability of such multitrack recordings for popular music, most MSS models focus on the separation of at least four stems including vocals, drums, base, and other [122, 151, 184]. Furthermore, there has been growing interest in the separation of individual sound sources within classical music recordings [29, 75, 121, 143, 164], which is also the main focus of our research. In the case of separating piano concertos, distinct timbral characteristics of the piano (e. g., clear onsets) may help a separation model in distinguishing piano from orchestral instruments such as strings, woodwinds, and brass. However, the source separation algorithms face a challenge when dealing with the strong spectro–temporal correlations among different instruments in piano concertos.

In contrast to popular music production, where individual instruments are often recorded in isolation, the direct interaction between musicians is an essential aspect of performing classical music. As a result, there are hardly multitrack recordings available for classical music [10, 16, 106, 125, 158, 163, 166]. In case multitrack recordings are unavailable, random mixing can be used to artificially generate and augment training data [30, 178]. Following this strategy, we used artificial training material in Chapter 4 by randomly mixing sections selected from the solo piano repertoire (e. g., piano sonatas, etudes, etc.) and orchestral pieces without piano (e. g., symphonies) to train an MSS model based on Spleeter [83]. As a main contribution, we extend the proposed approach in Chapter 4 and adapt four MSS models, each possessing distinct characteristics. As a second main contribution, we propose a musically motivated data augmentation method for training, inspired by the harmonic, rhythmic, and structural elements found in piano concertos.

As another extension of Chapter 4, instead of using artificially generated test data, we evaluate our models using the PCD, which provides a wide range of piano concerto recordings played by five performers in four different acoustic environments. For the evaluation of our models' performance, we use the widely-used SDR [205] and also the 2f-score [92], which is a perceptually motivated quality measure yielding better results in source separation tasks [195]. Finally, we conduct listening tests based on the MUSHRA framework [87] to assess the subjective perceptual separation quality. For the reproducibility of the results, we provide the open-source code and pretrained models as well as all test data used in our experiments and listening test in our GitHub repository[9].

6.2 Related Work in Source Separation

The models used in this chapter build upon DL approaches for general MSS models. Early works on MSS depend on the time–frequency (TF) representations, predicting a spectrogram for each individual musical source of a given recording. Based on the magnitude spectrogram of an input mixture (in our application, an existing piano concerto recording), most spectrogram-based neural network approaches estimate the magnitude spectrogram of the constituent musical sound sources [83, 88, 185]. Binary

[9] `https://github.com/yiitozer/pc-separation`

masking, soft masking, or multichannel Wiener filtering are then typically used to reconstruct the separated audio signals [110]. Besides using the magnitude spectrogram, recent approaches also use the real and imaginary parts or include the phase of the complex-valued spectrogram [31, 109, 196, 207]. For example, Choi et al. [32] report on the enhancement of separation performance with an ablation study conducted with spectrogram-based U-Net models through the usage of the real and imaginary parts. Note that this approach, denoted as *Complex-as-Channel (CaC)*, allows for directly taking the inverse STFT (iSTFT) from the learned representations, eliminating the necessity for further phase estimation methods such as Griffin-Lim [76] or Phase Gradient Heap Integration (PGHI) [148].

A second class of MSS models directly operates in the waveform domain [43, 181]. Waveform-based models receive the raw waveform of an input mixture and then predict the waveforms of the individual separated sources. Generally, these models implicitly perform some kind of TF analysis using convolution in their first layers [113]. Avoiding the computation of an STFT, waveform-based approaches do not require the explicit choice of a window size parameter. Moreover, operating in the waveform domain eliminates the need for an additional phase reconstruction, which is often required in spectrogram-based models.

The third class of MSS models apply hybrid techniques, which intuitively combine the complementary information provided by waveform- and spectrogram-based models [42, 95, 159, 178]. Hybrid approaches incorporate both spectral and temporal branches, merging the latent representations through addition or shared layers to leverage the advantages offered by each domain.

6.3 Adaptation of Source Separation Models

In this section, we first revisit the architecture and characteristics of four different models, which we adapt for our source separation task of piano concertos (see also Figure 6.2). In particular, we first explore the spectrogram-based models Open-Unmix (`UMX`), and Spleeter (`SPL`) in Section 6.3.1 and Section 6.3.2, respectively. Then, we investigate the waveform-based model Demucs (`DMC`) in Section 6.3.3. Finally, we describe in Section 6.3.4 the hybrid model HDemucs (`HDMC`), which operates both in spectrogram and waveform domains.

Under the assumption of an instantaneous linear mixing model [20], we represent the mixture signal $x_{\mathrm{m}} : \mathbb{Z} \to \mathbb{R}$ as a linear combination of waveforms of the estimated source signals $x_{\mathrm{m}} := \sum_{s \in S} x_s$, where S denotes the set of target sources. As we are dealing with the separation of piano concertos into piano and orchestra tracks, we have $S = \{\mathrm{p}, \mathrm{o}\}$, where p denotes the piano and o the orchestra source. It is important to note that all the separation approaches in this chapter are applied to stereo input waveforms or spectrograms, and the resulting output signals also comprise two channels. However, for the sake of simplicity and clarity, we use the monoaural signal model as in previous chapters.

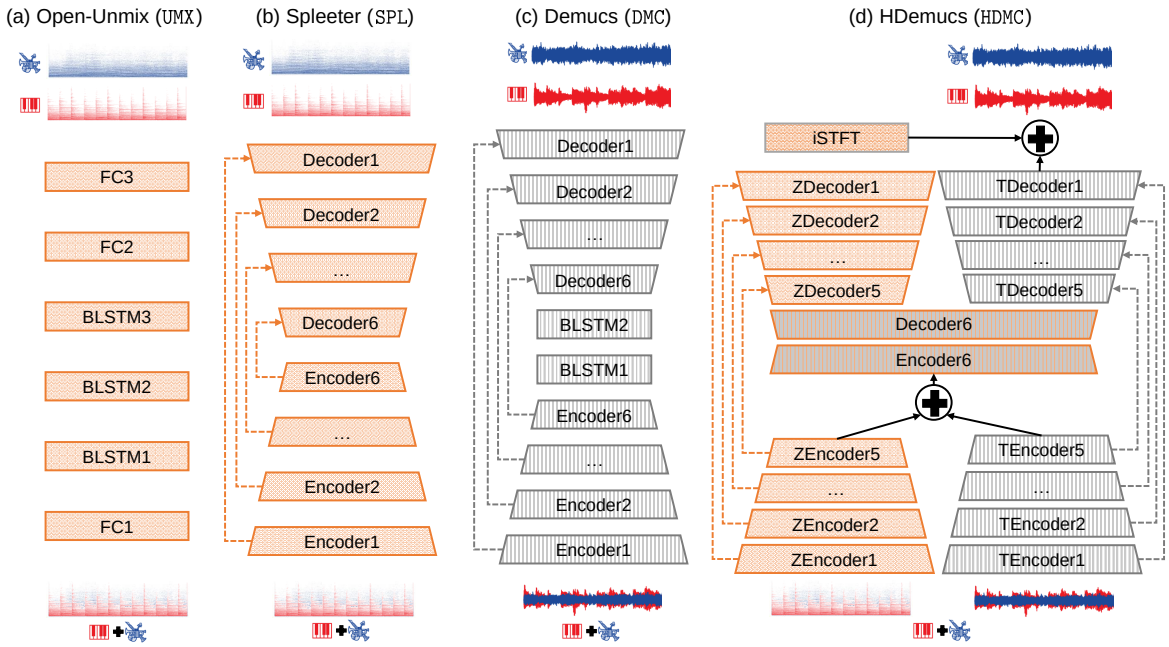

Figure 6.2: An overview of source separate models, which we adapt for separating piano concertos. Note that while only the monaural case is illustrated, all models are designed to work with stereo signals. **(a)** Spectrogram-based Open-Unmix (UMX) [185]. **(b)** Spectrogram-based Spleeter (SPL) [83]. **(c)** Waveform-based Demucs (DMC) [43]. **(d)** Hybrid model HDemucs (HDMC) [42]. Spectral branches are shown in orange and temporal in gray. Dashed lines denote the skip connections of the U-Net-based network architectures.

Table 6.1: List of adapted models.

Model ID	Domain	Size (MB)	#Targets
UMX	Spectrogram	33.93	1
SPL	Spectrogram	74.98	2
DMC	Waveform	510.22	2
HDMC	Hybrid	319.03	2

6.3.1 Open-Unmix (UMX)

Given the magnitude spectrogram $\mathcal{Y}_{\mathrm{m}}$ of an input mixture, UMX [185] learns a soft spectral mask M_s of a target musical source $s \in S$. The estimated magnitude spectrogram of a target source $\hat{\mathcal{Y}}_s$ is computed as:

$$\hat{\mathcal{Y}}_s = \mathcal{Y}_{\mathrm{m}} \odot M_s, \tag{6.1}$$

where $\odot$ denotes the Hadamard product (pointwise multiplication). For the reconstruction of the waveform of the estimated source signals, the input phase is used. In particular Multichannel Wiener Filtering is applied to minimize the total mean squared error (MSE) across all channels [110].

The core architecture of UMX is a three-layer bidirectional long short-term memory (BLSTM) [85] as described in [202] (see Figure 6.2a). Throughout our experiments, we remain consistent with the original

implementation and employ the MSE loss:

$$\mathcal{L}_{\text{MSE}} = \|\mathcal{Y}_s - \hat{\mathcal{Y}}_s\|_2^2, \tag{6.2}$$

where $\mathcal{Y}_s$ denotes the ground-truth magnitude spectrogram of a target source. For an investigation of various loss functions used with the `UMX` network, we refer to [78].

As indicated in Table 6.1, `UMX` is the model with fewest parameters among different approaches. However, in the original `UMX` approach, an independent training run is needed for each target source $s \in S$. This is also the method we follow in our experiments. For a multi-target variant of `UMX`, we refer to [165].

6.3.2 Spleeter (SPL)

Being a spectrogram-based model, `SPL` [83] also aims at approximating the magnitude spectrogram $\mathcal{Y}_s$ of a target source $s \in S$. Its architecture is based on the U-Net [157], which is widely-used model in MIR research to address the MSS task [32, 34, 43, 119, 159, 181]. Following this trend, we adapt the `SPL` implementation to predict the magnitude spectrograms of the constituent piano and orchestral parts in a piano concerto.

In our experiments, we use the same configuration as the U-Net model described in [88], which consists of 12-layer convolutional networks—six layers for encoder and six layers for the decoder (see Figure 6.2b). The skip connections account for the recovery of fine-grained details in the reconstructed representations. Note that `SPL` involves a separate U-Net for each source, which do not share weights. As shown in Table 6.1, the size of the model is 74.98 MB when having two sources. Each additional source adds parameters equivalent to 37.49 MB. The final layer of each U-Net model is a sigmoid activation function, yielding a soft mask M_s for each target source, which contains values between 0 and 1. The estimated magnitude spectrogram $\hat{\mathcal{Y}}_s$ is then computed as in Equation (6.1). Then, the estimated waveform of the target source $\hat{x}_s$ is reconstructed with Wiener Filtering [51].

For the loss function, we use the ℓ^1-norm between the magnitude spectrograms of the masked input mixture $\hat{\mathcal{Y}}_s$ and ground-truth target source $\mathcal{Y}_s$:

$$\mathcal{L}_1^{\text{spec}} = \frac{1}{|S|} \sum_{s \in S} \|\mathcal{Y}_s - \hat{\mathcal{Y}}_s\|_1. \tag{6.3}$$

For further details about the network architecture, we refer to [83, 88].

6.3.3 Demucs (DMC)

`DMC` [43] is a U-Net-based model which operates in the waveform domain. Given the raw waveform of an input mixture, it outputs an estimated waveform for each source without requiring any further

postprocessing step to recover the phase information. Similar to other U-Net-based MSS models in the literature, it contains a convolutional encoder–decoder network with skip connections (see Figure 6.2c). The rationale behind incorporating skip connections in this context is to provide direct access to the phase of the input mixture and transmitting it to the estimated sources. For temporal long-range dependencies, two BLSTM layers are included in the bottleneck. Note that the number of parameters within `DMC`'s encoder and decoder layers is larger than other U-Net-based models used in our experiments. As depicted in Table 6.1, `DMC` has the most parameters among the four models.

`DMC` is trained with an ℓ^1-norm in time domain:

$$\mathcal{L}_1^{\text{time}} = \frac{1}{|S|} \sum_{s \in S} \|x_s - \hat{x}_s\|_1, \tag{6.4}$$

where x_s represents the ground-truth target source in the time domain, and $\hat{x}_s$ the estimated time-domain signal. For a detailed account of the `DMC` model, we refer to [43].

6.3.4 Hybrid Demucs (HDMC)

`HDMC` [42] is an extension of `DMC` with an additional spectral branch. As illustrated in Figure 6.2d, its architecture contains a dual structure composed of U-Net-based networks with shared layers (Encoder6, Decoder6). Here, the spectral layers are denoted with the prefix 'Z' (shown in orange) and the temporal layers with the prefix 'T' (shown in gray), following the original notation in [42].

The spectral input (Figure 6.2d, left) is the complex-valued STFT $\mathcal{X}_\text{m}$ of an input mixture x_m. Following the *CaC* approach by Choi et al. [32], the real part $\text{Re}(\mathcal{X}_\text{m})$ and the imaginary part $\text{Im}(\mathcal{X}_\text{m})$ of the input mixture are encoded by different channels of the spectral branch. The convolutional kernels are applied along the frequency dimension, leading to a one-dimensional representation as the output of the 5^{th} encoder layer (ZEncoder5) of the spectral branch of the network.

The temporal branch (Figure 6.2d, right) receives the raw waveform x_m, similar to `DMC`. The output of the 5^{th} temporal encoder layer (TEncoder5) is of the same size as the output of ZEncoder5. The learned spectral and temporal representations are then summed and used as the input to the 6^{th} encoder layer. The output of the 6^{th} encoder layer serves as an input both for spectral and temporal decoders. To account for the long-range temporal context, the 5^{th} and 6^{th} layers of the encoder involve local attention and BLSTM layers.

As output, the spectral decoder produces a complex-valued spectrogram, which is inverted with iSTFT to generate the waveform $\hat{x}_s^\text{Z}$. Furthermore, the temporal branch directly outputs a waveform $\hat{x}_s^\text{T}$. The outputs from both branches are summed to compute the estimated waveform of the target source:

$$\hat{x}_s = \hat{x}_s^\text{Z} + \hat{x}_s^\text{T}. \tag{6.5}$$

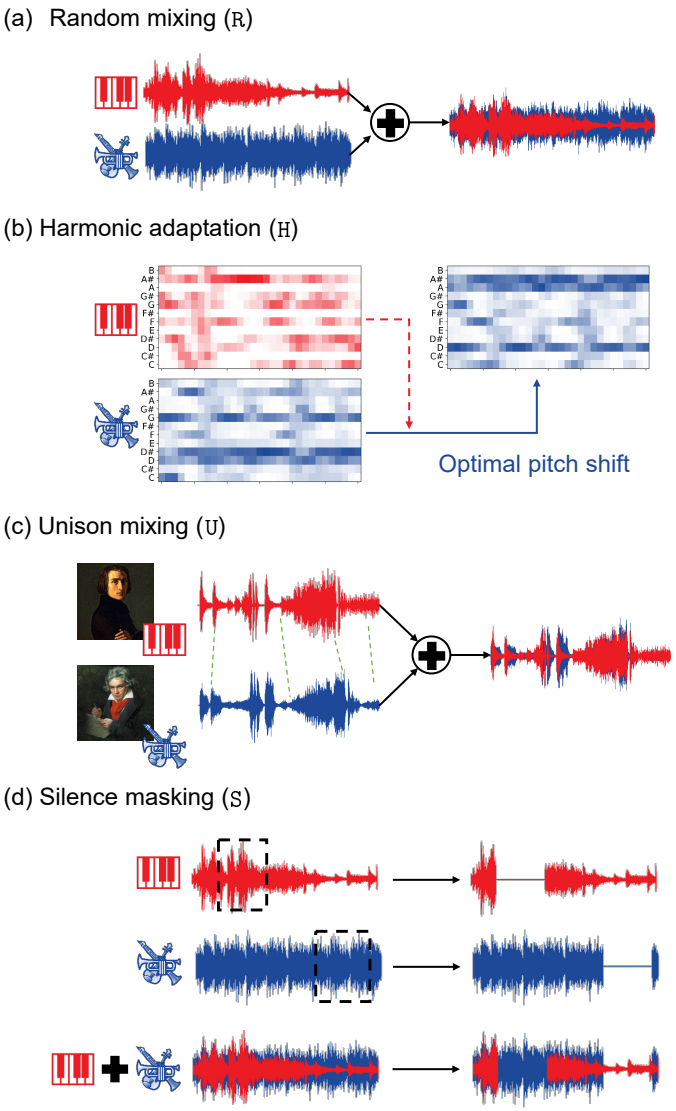

Figure 6.3: Musically motivated data augmentation strategies. **(a)** Random mixing recordings from the solo piano repertoire (e. g., piano sonatas) and orchestral recordings without piano (e. g., symphonies). **(b)** Harmonic adaptation of the orchestral recordings to the piano tracks using optimal pitch shift. **(c)** Creating additional training material by aligning recordings of Beethoven symphonies with their Liszt piano transcriptions. **(d)** Silence masking to replicate the silent passages in the piano or orchestral part.

Similar to DMC, we use the ℓ^1-norm as the loss function of HDMC, as in Equation (6.4). For further details about the network architecture, we refer to [42].

6.4 Musically Motivated Data Augmentation

In this section, we present our strategy to create and augment data for training our MSS models. In particular, we propose four data augmentation techniques as illustrated in Figure 6.3. In the following, we delve deeper into our proposed methods, inspired by the harmonic, rhythmic, and structural elements found in piano concertos.

6.4.1 Random Mixing

Supervised deep learning models designed for MSS typically rely on large datasets containing recordings of isolated stems. Since such multitrack recordings are not available in the case of piano concertos, we create a dataset as in Chapter 4 through random mixes of piano-only recordings (e. g., piano sonatas) and recordings of orchestral music without piano (e. g., symphonies), see Figure 6.3a for an illustration. While this method does not reflect the harmonic and rhythmic interaction among different instruments found in most real recordings, it helps the MSS model identify the timbral characteristics of concurrent musical sources. However, this approach may correspond to passages in piano concertos which are *atonal* and do not follow a *homorhythmic* texture.

Our training data combines open-source datasets and publicly accessible orchestral recordings from the International Music Score Library Project (IMSLP)[10]. As for the piano recordings, we first use MAESTRO [81], which involves 198.7 hours of piano performances recorded on Yamaha Disklaviers. To account for other room acoustic conditions and inclusion of different pianos, we further incorporate the ATEPP [214] dataset, which contains approximately 1000 hours of piano recordings performed by 49 pianists, spanning 1580 movements by 25 composers. Due to their large size, we create subsets randomly selecting piano recordings from the two datasets. The subset derived from the MAESTRO dataset amounts to approximately 6 hours, while we incorporate 24 hours of piano recordings from the ATEPP dataset.

For orchestral recordings, we use symphonies and ensembles selected from four open-source datasets. First, we use the Phenicx Anechoic dataset [166], which consists of clean multitrack recordings of four orchestral excerpts by different composers. Second, we consider Bach10 [50], which comprises multitrack recordings of ten chamber music pieces where each work comprises four parts (SATB) played by violin, clarinet, saxophone, and bassoon. Third, we use the OrchSet dataset [17], which contains 64 audio excerpts from orchestral works interpreted by symphonic orchestras, mostly from the romantic period, as well as classical and 20^{th} century pieces. Fourth, we select a subset of 19 classical music recordings without piano selected from the Real World Computing (RWC) dataset [73]. Furthermore, we also use public-domain symphonies and concertos from IMSLP for training. Given that string instruments usually dominate in orchestral compositions, we also include concertos of woodwind and brass instruments, in particular solo sections of these underrepresented instruments to obtain a more diverse dataset. In summary, this selection helps to balance the training dataset, in particular adding excerpts that involve non-string instruments.

To create our dataset, we first extract 30-second chunks from piano and orchestral recordings. To account for a high variety, we ensure that the chunks selected from a piano recording are mixed with chunks from various orchestral recordings, and vice versa. During the training phase, we also use gains to create a range of volume ratios, which reflects that the piano's sound intensity may substantially change relative to the orchestral track. The total duration of our dataset involving randomly generated mixture recordings is approximately 30 hours.

[10] `https://imslp.org/`

6.4.2 Harmonic Adaptation

Piano concertos are composed specifically to show an interaction between the piano and orchestra. In these compositions, the piano is closely intertwined into the orchestral accompaniment, often sharing melodic, rhythmic, and harmonic elements. Due to the strong interaction of the piano and orchestra, it is not possible to simulate real music recordings simply by superimposing signals extracted from different sources.

While random mixing can help the MSS methods to learn timbral characteristics of the concurrent sources to some extent, it generates harmonically implausible combinations, which may only loosely mimic real music recordings. Given that the majority of piano concertos in the Western classical music repertoire are mostly tonal, the musical elements occurring simultaneously exhibit strong harmonic relationships [20]. In this context, to obtain more realistic mixtures, we incorporate harmonic adaptation into our training process as a further stage of our musically motivated data augmentation procedure.

There are several approaches in the literature, which consider using the chroma features to assess the similarity between different sources in the context of random mixing [30, 93, 213], and apply pitch shifting to create more harmonically plausible mixtures [42]. Inspired by this approach, we first compute the chroma features of the piano and orchestral recordings and apply pitch shifting to the orchestral recordings, taking the corresponding piano track as a reference. Figure 6.3b depicts an example of this strategy, where the harmonics of the orchestral recording are dominated by D♯, whereas the piano recording's harmonic content is primarily in A♯. After optimal pitch shifting, we obtain a more harmonically plausible random mixture.

6.4.3 Unison Mixing

While separating music signals, it is generally assumed that the harmonics and transients of different signals only partially overlap. However, if the constituent sources of a musical mixture play the same notes simultaneously (i.e., in unison), the different sources highly overlap both in time and frequency, leading to a significant challenge for MSS algorithms [182]. This phenomenon can also be understood within the context of multiple-voice *monody* or *monophony*, which represents the most challenging musical textures for separation, given that parallel voices follow the exact same melody [20]. Various piano concertos involve passages, in which piano and orchestra play in unison. For example, this happens in the Bach Piano Concerto in F minor, BWV 1056 and Schumann Piano Concerto in A minor, Op.54 (see, e. g., the excerpts with ID 000, 005, 071, and 073 in the test dataset [136][11]).

To better separate unison mixtures of orchestral instruments, Stöter et al. [183] proposed a method to exploit instrument-specific modulation structures for source separation. It turns out that this approach is particularly suitable for strings and brass instruments. For simulating unison passages in piano concerto

[11] `https://www.audiolabs-erlangen.de/resources/MIR/PCD`

recordings, we consider generating unison data with alignment techniques. To this end, we exploit that many orchestral works were transcribed to piano throughout the music history. An iconic example is the renowned piano transcriptions by Franz Liszt for Beethoven's symphonies. For these piano-reduced versions, one can find multiple recordings by famous pianists such as Glenn Gould. To create highly overlapping piano–orchestra mixtures, we synchronize public-domain recordings of Beethoven symphonies with recordings of their piano-reduced versions (see Figure 6.3c).

For the alignment of orchestra and piano versions, we use DTW, which is a well-known technique for music synchronization [46, 161]. Conventional methods typically use chroma features as the input representation to the alignment algorithm [38, 124]. Despite its robustness for music synchronization in view of harmonic and melodic information, using only chroma features does not ensure a high temporal synchronization accuracy. Since we aim to simulate unison recordings, in which the piano and orchestral tracks play the same notes simultaneously, a high temporal accuracy is required.

To increase the temporal alignment accuracy, Ewert and Müller [57] introduced a combined synchronization approach, which integrates additional onset-related information besides chroma features. The inclusion of onset-based information results in a grid-like structure in the DTW cost matrix, which guides the alignment through activation cues that highlight note onsets. Inspired by this combined synchronization approach, we follow the alignment method introduced in Chapter 3. This method incorporates beat, downbeat, and onset activation functions computed using the open-source *madmom* library [14][12], alongside chroma features, to compute the alignment path. To create a training set of unison recordings, we generate the alignment paths for each pair of the symphony recordings and recordings of their piano transcriptions using the open-source Sync Toolbox [126], which provides an efficient implementation of DTW [146].

To generate orchestral tracks, which are synchronous with the piano recordings, we then employ TSM. Using the alignment path acquired from DTW as an input for the TSM algorithm, we speed up or slow down the orchestral track without affecting the frequency content. For TSM, we use the approach by Driedger et al. [49], which combines harmonic–percussive source separation (HPSS) and classical TSM algorithms, such as phase vocoder [61], and WSOLA [204]. The duration of this additional dataset of unison mixtures is approximately 22 hours.

6.4.4 Silence Masking

Depending on the compositional style, piano concertos may involve long sections where the piano and orchestra do not play together. In particular, in the concertos written in the Classical period, the piano and orchestra often follow a conversational style, such as in Beethoven's Piano Concerto No. 4 in G Major, Op. 58 [26], (see, e. g., the excerpts with the PCD ID 025 and 026 in the test dataset). Moreover, piano concertos often comprise long piano-only (e. g., in the cadenza) and orchestra-only parts (e. g., in the exposition, also called *opening ritornello*). In Chapter 4, we exploit this property of the piano concertos

[12] `https://github.com/CPJKU/madmom`

for further finetuning the MSS model at test time, a strategy called test-time adaptation [188]. Several works in the literature apply activity-based approaches as a prior to enhance audio source separation, e. g., [170, 194]. Inspired by this strategy, we randomly mask out passages either in the piano or in the orchestral track (but never simultaneously), see Figure 6.3d for an illustration.

6.5 Evaluation

In this section, we describe our systematic experiments and report on the separation results acquired by the four MSS models using various musically motivated data augmentation approaches. First, we outline our experimental settings in Section 6.5.1. We then discuss the quantitative empirical results in Section 6.5.2 and present the results of our listening tests in Section 6.5.3. Finally, we elaborate in more detail on the impact of transfer learning and unison mixing in Section 6.5.4.

6.5.1 Experimental Setting

In our experimental setup, we use stereo recordings, which are sampled at 44.1 kHz. For the spectrogram-based and hybrid models, we apply an STFT using a Hanning window of length $N = 4096$ and hop size of $H = 1024$, consistent with the default settings in [42, 43, 83, 185]. For `UMX`, we use two different settings, where we train one model with 6-second random chunks (in [185], default setting) and another model with 20-second random chunks. The random chunks used for training the other models have a duration of 20 seconds, as in the default setting of `SPL`. We use the default learning rates given in the original implementations, ADAM optimizer, and early stopping with patience 20 (indicating the number of epochs with no improvement in the validation loss before terminating the training). All models are trained using a single NVIDIA GeForce RTX 3090 GPU.

We apply a four-stage learning process for each model. Each subsequent stage utilizes transfer learning by initializing the model with weights that were pre-trained during the prior stage, and then proceeds to further train all of these weights. For an in-depth discussion on the effects of this transfer learning approach, please refer to Section 6.5.4. We initially train our models starting with random initialization, using the artificial dataset generated through random mixes with various gains, as detailed in Section 6.4.1. We denote the first training stage as `R`. After reaching convergence in this training stage, we apply pitch shifting with an optimal chroma index to the orchestral recordings (see Section 6.4.2). We call this stage `R_H`. In the third stage, we incorporate the synchronized Beethoven symphony recordings and their transcriptions for solo piano to simulate unison passages within piano concertos (see Section 6.4.3). This stage is denoted as `R_H_HU`. The fourth and final stage called `R_H_HU_HUS` introduces the random silent parts into the two sources (see Section 6.4.4). To account for a fair comparison, we ensure that all DL-based models receive identical training data samples in the same order and using the same randomization parameters (e. g., volume ratio, starting point of a chunk or silence mask).

Given that the first level learns easier aspects of the task and that the difficulty level gradually increases in the subsequent stages due to the rise in overlapping harmonics and onsets, this approach can be thought of as curriculum learning [8], which exploits, particularly in the first three stages, previously learned concepts to ease the learning of new abstractions.

6.5.2 Quantitative Evaluation

To get a first impression of the model performances, we use the SDR [205] as our quantitative evaluation metric for the separation task. Table 6.2 shows the mean SDR values (averaged over all test samples) with corresponding variances of the four models (where `UMX06` denotes the `UMX` model trained on 6-second chunks and `UMX20` denotes the `UMX` model trained on 20-second chunks).

At first, we focus on the SDR results obtained for the separation of the piano. After the first training stage `R`, `HDMC` achieves the highest average SDR value 8.67, followed by the spectrogram-based models `UMX20` yielding 8.45, and `SPL` with a result of 7.93. Among the four models, `DMC` results in the lowest SDR value of 7.47, after the stage `R`.

The SDR results for separating the orchestral track follow a similar trend, although the values, in general, are significantly lower. For the orchestra, `HDMC` yields the highest average SDR value of 3.86 after the first training stage `R`, again followed by the spectrogram-based models `UMX20` yielding 3.65, and `SPL` with a result of 3.32. Among the four models, `DMC` results in the lowest average SDR value after stage `R`, 2.68.

Next, we investigate the effect of different training strategies. In general, the SDR-based results demonstrate that incorporating data augmentation approaches improves the separation performance of the hybrid model `HDMC`. The largest performance boost for `HDMC` occurs after the second stage `R_H` (a rise from 8.67 to 9.30 for the piano, 3.86 to 4.53 for the orchestra), where we apply harmonic adaptation to the orchestral recordings in the training dataset. Similarly, we observe a general improvement by each stage for the models except for `UMX`.

Interestingly, `UMX` model's performance improves with a large margin, when using 20-second chunks instead of 6-second chunks. For example, after the `R` stage, the SDR value of `UMX20` is 8.45 compared to 7.74 for `UMX06`. Whereas the SDR values of `UMX06` are steadily lower than the `SPL` model, employing longer chunks results in significantly higher values, causing the `UMX20` to outperform the other spectrogram-based model `SPL` in our experiments. Furthermore, neither the performance of `UMX06` nor of the `UMX20` model improves with the data augmentation procedures. We hypothesize that the fewer parameters hinder the UMX model from learning more complex tasks (see also Table 6.1).

While SDR is commonly used as a quantitative evaluation metric for MSS, it is widely accepted that SDR is not suitable for determining the perceptual sound quality of separated musical sources [23]. In particular, the analysis conducted by Torcoli et al. [195] for the source separation task reveals that the *2f-score* metric demonstrates the strongest correlation with ground-truth data based on subjective ratings

Model	Piano				Orchestra			
	R	R_H	R_H_HU	R_H_HU_HUS	R	R_H	R_H_HU	R_H_HU_HUS
UMX06	7.74 ± 4.05	7.72 ± 4.13	7.69 ± 3.97	7.72 ± 4.02	3.00 ± 2.22	2.96 ± 2.25	2.94 ± 2.32	2.96 ± 2.30
UMX20	8.45 ± 4.34	8.46 ± 4.33	8.39 ± 4.22	8.38 ± 4.24	3.65 ± 2.14	3.66 ± 2.17	3.61 ± 2.21	3.61 ± 2.19
SPL	7.93 ± 3.99	8.04 ± 3.96	8.15 ± 3.98	8.16 ± 3.99	3.32 ± 2.17	3.45 ± 2.21	3.46 ± 2.26	3.46 ± 2.25
DMC	7.47 ± 4.40	7.58 ± 4.40	7.58 ± 4.37	7.59 ± 4.38	2.68 ± 2.15	2.78 ± 2.16	2.82 ± 2.13	2.82 ± 2.13
HDMC	8.67 ± 4.24	9.30 ± 4.00	9.41 ± 4.18	**9.61 ± 4.42**	3.86 ± 2.34	4.53 ± 2.46	4.61 ± 2.39	**4.75 ± 2.31**

Table 6.2: Mean SDR values and variances of different models trained with various data augmentation methods. The mean values and variances are computed over all test items.

Model	Piano				Orchestra			
	R	R_H	R_H_HU	R_H_HU_HUS	R	R_H	R_H_HU	R_H_HU_HUS
UMX06	32.77 ± 7.54	32.89 ± 8.42	32.65 ± 7.68	32.77 ± 7.90	28.00 ± 7.61	28.27 ± 7.51	28.76 ± 7.51	28.86 ± 7.47
UMX20	34.75 ± 7.94	34.15 ± 8.10	34.01 ± 7.57	33.72 ± 7.24	29.50 ± 7.40	30.14 ± 7.42	30.13 ± 7.54	29.99 ± 7.61
SPL	33.77 ± 9.48	34.50 ± 9.31	34.77 ± 8.95	34.75 ± 8.70	28.58 ± 5.94	28.56 ± 5.94	28.79 ± 5.94	29.01 ± 5.95
DMC	30.45 ± 11.10	30.66 ± 11.18	30.76 ± 11.14	30.80 ± 11.15	25.74 ± 7.74	26.86 ± 7.63	26.86 ± 7.65	26.87 ± 7.66
HDMC	37.66 ± 11.28	39.81 ± 11.22	**40.59 ± 11.00**	40.47 ± 10.89	33.42 ± 6.44	34.76 ± 7.02	**35.40 ± 6.65**	35.01 ± 6.62

Table 6.3: Mean 2f-score values and variances of different models trained with various data augmentation methods. The mean values and variances are computed over all test items.

from MUSHRA listening tests. For a more detailed account on the 2f-score, we refer to [92]. Note that the 2f-score values lie in a range from 0 to 100 following the MUSHRA framework (also see Section 6.5.3). Table 6.3 presents a comparison of the various models trained with different strategies, based on the 2f-score results. In general, one can observe a similar trend as for the SDR. For both, the piano and orchestra, HDMC yields the highest average 2f-score values after each training stage, followed by UMX20, SPL, UMX06, and DMC. Furthermore, we observe a general trend of performance improvement within the first three training stages for SPL, DMC, and HDMC. Interestingly, the 2f-score suggests that the best results are achieved with the HDMC model after the third training stage R_H_HU, which introduces the unison mixing as a data augmentation strategy (see Section 6.4.3). Applying silence masking slightly worsens the resulting 2f-scores for HDMC.

6.5.3 Subjective Evaluation

In this section, we describe the experimental setup for our subjective listening tests to evaluate the perceived quality of separation. For our experiments, we used the MUSHRA framework following the ITU-R BS.1534-3 recommendation [87]. The MUSHRA methodology employs a double-blind multi-stimulus test approach, including a hidden reference and a lower anchor signal. Participants rate the stimuli on a scale of 0 to 100, involving five categories: Bad (0-20), Poor (20-40), Fair (40-60), Good (60-80), and Excellent (80-100).

A total of 26 participants were involved in our listening tests (23 experienced listeners and 3 inexperienced listeners). To ensure the reliability of the results, the MUSHRA methodology recommends a post-screening

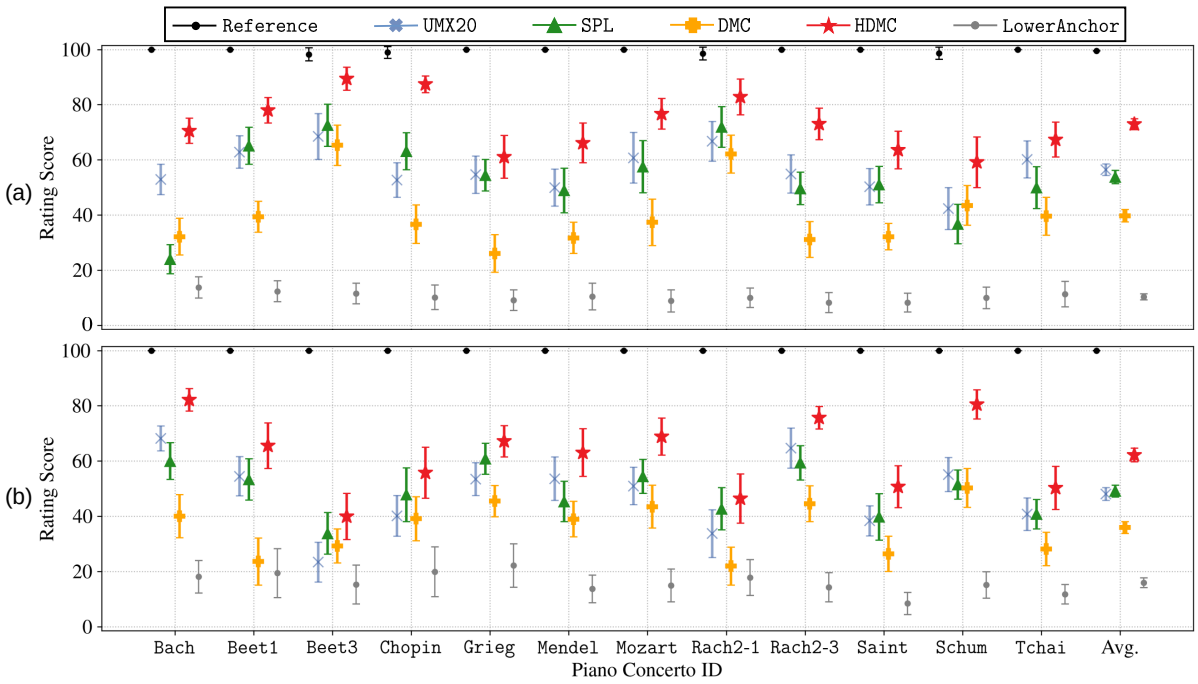

Figure 6.4: Results of our listening tests based on the MUSHRA framework for the **(a)** piano and **(b)** orchestral tracks. The listening test employs models that all incorporate the complete data augmentation approach (`R_H_HU_HUS`). The colored markers indicate the average rating scores enclosed by 95% confidence intervals (shown as the vertical lines).

of the participants stating that participants should be excluded from the listening test if they assign the hidden reference to a score lower than 90 for more than 15% of the test items. Following these criteria, none of the participants was excluded after post-screening.

To assess the subjective quality of separated source signals, we conducted two listening tests. In our first listening test, we asked the participants to rate the overall audio quality of waveforms of separated piano source obtained by the four MSS models (`UMX20`, `SPL`, `DMC`, `HDMC`). The participants gave their ratings with respect to a reference signal, which is a clean piano-only excerpt. Similarly, our second listening test evaluated the overall quality of the separated orchestral tracks following the same procedure as in the first listening test. Each of the two listening tests contains 12 test items selected from PCD. With these test items, we cover excerpts of piano concertos composed by 10 composers, spanning from the Baroque to the Post-Romantic era, played by different performers in different acoustic environments. This selection introduces a multitude of challenges for the MSS algorithms, due to the variations in orchestration, compositional style, performance technique, and acoustical characteristics of the recording environments.

For the subjective evaluation of each test item, we generated six signals (also called *conditions*). The first signal is the hidden reference, i.e., a replication of the ground-truth source signal. The second condition is a lower anchor. As in [52], we created this lower anchor by low-pass filtering the test mixtures with a

Model	Piano			Orchestra		
	R	H	R_H	R	H	R_H
UMX06	7.74 ± 4.05	7.89 ± 4.16	7.72 ± 4.13	3.00 ± 2.22	3.12 ± 2.15	2.96 ± 2.25
UMX20	8.45 ± 4.34	8.71 ± 4.21	8.46 ± 4.33	3.65 ± 2.14	3.89 ± 2.25	3.66 ± 2.17
SPL	7.93 ± 3.99	7.70 ± 3.74	8.04 ± 3.96	3.32 ± 2.17	3.23 ± 2.31	3.45 ± 2.21
DMC	7.47 ± 4.40	6.19 ± 4.68	7.58 ± 4.40	2.68 ± 2.15	1.42 ± 2.12	2.78 ± 2.16
HDMC	8.67 ± 4.24	9.00 ± 4.47	**9.30** ± **4.00**	3.86 ± 2.34	4.17 ± 2.13	**4.53** ± **2.46**

Table 6.4: Mean SDR values and variances of different models trained with various data augmentation methods. The mean values and variances are computed over all test items.

3.5kHz cut-off frequency and by adding musical noise. The other four signals involve estimated piano or orchestral sources separated by UMX20, SPL, DMC, and HDMC. For our listening tests, we used the models trained with the learning strategy R_H_HU_HUS, which involves all the data augmentation approaches described in Section 6.4. For an overview of the test items used for the listening test, please refer to our demo webpage[13].

Figure 6.4 provides an overview of the results from our listening tests. First, one can observe that the participants rated the reference signal with an average MUSHRA rating score of 100, the lower anchor was rated significantly below the other conditions. The general trend of the performances by UMX20, SPL, DMC, and HDMC support our quantitative analysis results, inferring that the hybrid model HDMC outperforms other models by a large margin. Spectrogram-based models UMX20 and SPL yield similar scores, whereas the waveform-based DMC has the lowest ratings among the four MSS models. In general, the piano separation is rated better than the orchestral part, which is consistent with the quantitative results based on SDR and 2f-score.

Upon observing the rating scores of the piano concertos individually, it is noticeable that there are substantial differences in the ratings across the various test items (most of the participants also noted the variation in perceived separation quality between different works). This trend in separation performance remains consistent across different test items, with the hybrid model HDMC consistently achieving the highest scores. It is important to remark that the test items are diverse regarding several aspects. For example, Bach and Schum involve unison passages, yielding a high overlap both in time and frequency domains. In particular, unison passages constitute a big challenge for the spectrogram-domain approaches (see Bach). Furthermore, the excerpts Rach and Tchai involve loud piano passages and a complex orchestration consisting of a diverse and high number of instruments (see the orchestrations in PCD).

6.5.4 Further Experiments

In this section, we investigate the effect of transfer learning and unison mixing in more detail to gain a deeper understanding how different training methodologies influence the MSS models' performance.

[13] https://www.audiolabs-erlangen.de/resources/MIR/2023-PianoConcertoSeparation/

Model	Piano			Orchestra		
	RR⋆	HU	R_H_HU	RR⋆	HU	R_H_HU
UMX06	8.70 ± 3.97	7.96 ± 3.68	7.69 ± 3.97	3.93 ± 2.42	3.23 ± 2.50	2.94 ± 2.32
UMX20	8.81 ± 4.25	8.50 ± 3.86	8.39 ± 4.22	4.02 ± 2.25	3.74 ± 2.40	3.61 ± 2.21
SPL	8.31 ± 4.19	8.11 ± 3.60	8.15 ± 3.98	3.83 ± 2.22	3.30 ± 2.03	3.46 ± 2.26
DMC	6.15 ± 4.09	6.79 ± 4.25	7.58 ± 4.37	1.44 ± 2.32	2.05 ± 1.96	2.82 ± 2.13
HDMC	8.99 ± 4.32	9.14 ± 4.38	**9.41 ± 4.18**	4.16 ± 2.46	4.33 ± 2.22	**4.61 ± 2.39**

Table 6.5: Mean SDR values and variances of different models trained with various data augmentation methods. The mean values and variances are computed over all test items.

Instead of training with random mixes (R) and then continuing with harmonic adaptation (R_H), we now train all models from scratch using only the harmonically adapted training dataset, a process referred to as H in the following.
Table 6.4 presents the mean SDR values with corresponding variances of the different models for the three training strategies, R, H, and R_H. The results indicate that for the simpler models, UMX06 and UMX20, using H directly yields a minor improvement compared to R. For SPL, using H even slightly worsens the separation performance, and, for DMC, it surprisingly results in a decay of SDR scores of more than 1 dB for both piano and orchestra. Furthermore, in case of R_H, we observe a positive impact of the transfer-learning-based strategy for SPL, DMC, and HDMC, compared to training with harmonically adapted dataset from scratch (H).

Next, we explore the effect of unison mixing as a data augmentation strategy. In particular, we investigate whether the improvements through unison mixing reported in Section 6.5.2 can be attributed to the mixing process itself or the inclusion of additional training material involving Beethoven symphony recordings and their piano transcriptions underlying the mixing process. To this end, we generate a new dataset, called R⋆, by randomly mixing excerpts from the original orchestral versions with completely unrelated (in particular unaligned) excerpts from piano transcriptions. We combine R⋆ with the random mixes from R, yielding the dataset RR⋆, which is then employed to train different models from scratch. Additionally, we also train different models using the training material created with unison mixing (i.e., synchronized Beethoven symphony recordings and their solo piano transcriptions), merged with the mixes from H – harmonically-adapted random mixes from R – from scratch. We refer to this training procedure as HU. Note that this training dataset is identical to the one used in the last training stage of R_H_HU, which employs transfer learning by initializing the model weights from its prior stage R_H, as described in Section 6.5.1.

Mean SDR scores and their variances for the various models, evaluated across the three training strategies RR⋆, HU, and R_H_HU, are presented in Table 6.5. For piano separation, HU results in lower SDR scores for the spectrogram-based models UMX06, UMX20 and SPL compared to RR⋆. This observation can be attributed to the difficulty in distinguishing unison sound sources when using only magnitude spectrograms for the separation task. In contrast, waveform-based DMC and HDMC, which also considers audio waveforms as input, benefit from unison mixing. For orchestra, when comparing RR⋆ and HU, similar observations

can also be made. Confirming the results in Table 6.2, the training procedure based on transfer learning, R_H_HU yields a better separation performance for DMC, and HDMC, compared to HU. Notably, for HDMC, HU results in a mean SDR score of 9.14 and with R_H_HU, it improves to 9.41 for piano separation. Similarly, for separating orchestra, it improves from 4.33 to 4.61 with transfer learning.

In summary, these final experiments show that our data augmentations including unison mixing in combination with transfer learning are beneficial for our best-performing model HDMC. However, this approach does not appear to yield similar improvements for smaller models, e. g., UMX06 and UMX20.

6.6 Conclusion

In this chapter, we addressed the rarely-considered task of decomposing piano concerto recordings into separate piano and orchestral tracks. We identified the challenges associated with this task, including the intricate interplay and high spectro–temporal correlations between the constituent instruments, as well as the lack of multitrack training data for piano concertos. To address the challenge, we adapted four DL-based methods of different characteristics and conducted systematic experiments to explore spectrogram-, waveform-based as well as hybrid source separation models. We introduced a musically motivated data augmentation approach, inspired by the harmonic, rhythmic, and structural elements found in piano concertos. The key finding is that the best source separation performance was accomplished by the hybrid model trained with a full suite of augmentation techniques. In future work, we would like to investigate and improve the interpretability of the hybrid models by analyzing the outputs of the individual time and spectral branches. Furthermore, we aim at incorporating score information to further enhance the separation performance.

7 Notewise Evaluation of Source Separation

This chapter is based on a publication in progress. The first author Yigitcan Özer will be the main contributor to this article. In collaboration with Hans-Ulrich Berendes, Vlora Arifi-Müller, Fabian-Robert Stöter, and his supervisor Meinard Müller, he devised the ideas, developed the formalization, and wrote the paper. Furthermore, Yigitcan Özer implemented all approaches and conducted the experiments.

Deep learning has significantly advanced MSS, aiming to decompose music recordings into individual tracks corresponding to singing or specific instruments. Typically, results are evaluated using quantitative measures like SDR computed for entire excerpts or songs. As the main contribution of this chapter, 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 NMF approach to decompose the reference and separated 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.

The remainder of the chapter is organized as follows. Following the introduction in Section 7.1, in Section 7.2, we review relevant literature on source separation and introduce the MSS models used for separating piano concertos. Subsequently, in Section 7.3, we elaborate on the score-based extension of PCD (see Chapter 5) and outline our evaluation approach, covering NMF-based audio decomposition and notewise SDR-based metrics. In Section 7.4, we provide details on the experimental settings and report our empirical findings. Finally, in Section 7.5, we conclude and discuss potential directions for future work.

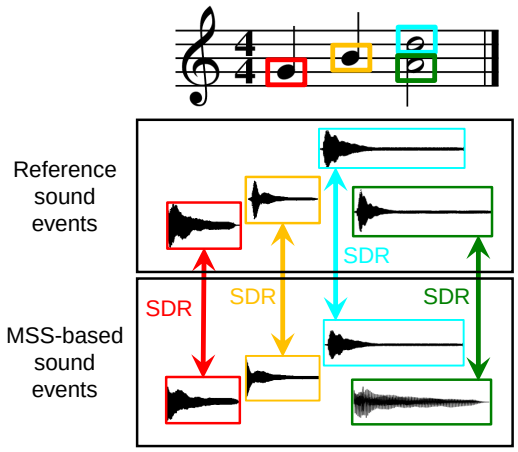

Figure 7.1: Illustration of the proposed evaluation method for MSS, considering SDR values based on notewise sound events rather than entire recordings.

7.1 Background

MSS is a key task in MIR, involving the separation of a musical mixture into individual components like vocals, instruments, and other sound elements [24]. 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 [42, 83, 114, 185]. In scenarios with limited training data, MSS systems are often trained using artificially generated mixes through synthesis techniques [163] or data augmentation approaches [93]. An example of such a scenario, also addressed in this chapter, is presented in Chapter 4, where the goal is to separate piano concertos into piano and orchestra tracks.

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) [184] and the Music Demixing Challenge [122] 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 SDR [205]. 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 [23, 195]. On the other hand, designing perceptually or musically more relevant measures is challenging, and performing listening tests is often cumbersome and infeasible.

In this chapter, 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 NMF [54] to decompose signals into notewise sound events.

Table 7.1: MSS models considered in our experiments. TS denotes the size (in hours) of the training set used.

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+

Then, we calculate SDR values for individual units before aggregating this information in various ways (see Figure 7.1).

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 PCD, 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 [57, 126] to align these scores with all recorded excerpts. As an additional contribution of this chapter, 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.

7.2 Music Source Separation (MSS)

In this chapter, we again 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 approaches in Chapters 4 and 6 propose 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 Chapter 6, including the use of further data augmentation techniques. In our experiments, we employ four pre-trained models introduced in the Chapter 6, shown in Table 7.1. Note that we only use the UMX model, trained with 20-second chunks in this chapter (UMX20). 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

Table 7.2: 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).

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

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 chapter. 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 7.4.

7.3 Evaluation Approach

We now introduce our novel evaluation approach, which we will apply to compare reference piano recordings and separated piano tracks. In Section 7.3.1, we briefly revisit the PCD collection, which will serve as test dataset, and present our score-based extensions. Then, in Section 7.3.2, we revisit the score-informed NMF approach for audio decomposition. Finally, in Section 7.3.3, we define the SDR-based evaluation metrics, which we use to gain a deeper understanding of the source separation results.

7.3.1 Piano Concerto Dataset and its Extension

The PCD collection, introduced in Chapter 5, 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*[14]. 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 7.2, the PCD comprises excerpts recorded in four distinct acoustic settings with different grand piano models. The total duration of excerpts amounts to 972 seconds, with the shortest total duration of excerpts recorded in any single room being at least three minutes.

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[15] for the piano tracks (and piano-reduced versions of the orchestra tracks,

[14] `https://www.halleonard.com/series/MMONE`
[15] `http://www.sibelius.com/`

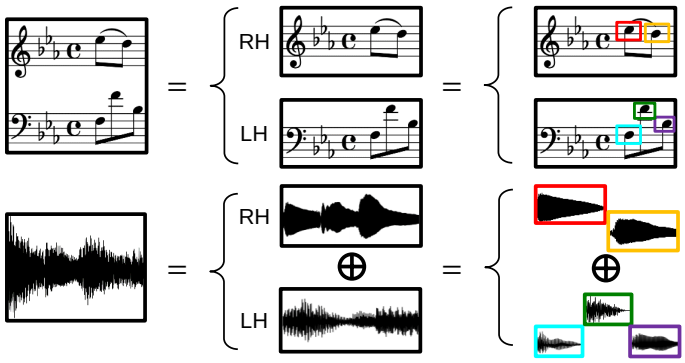

Figure 7.2: Illustration of the decomposition of the piano track into left-hand (LH), right-hand (RH), and individual note events as indicated by the rectangular windows.

which are not utilized in this chapter). We employed the *Sync Toolbox* [126][16] 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* [21] 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 sound events. 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 chapter.

7.3.2 NMF-Based Audio Decomposition

NMF is an algorithm for approximating a nonnegative matrix as the product of two low-ranked nonnegative matrices [101]. In the context of music processing, NMF has been widely applied to decompose a magnitude spectrogram into the product of two nonnegative matrices [175], 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 [149] or use aligned score information to constrain the activation matrix [54]. In addition to stabilizing the convergence of the NMF algorithm, such constraints also guide the factorization process to yield decompositions of musical relevance [58]. For a more detailed account of score-informed NMF with multiplicative update rules, see Section 8.2.

[16] `https://github.com/meinardmueller/synctoolbox`

Following the approach in [48], 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

$$x = \sum_{m=1}^{M} x^m + r. \tag{7.1}$$

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 7.2.

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 decomposition 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 7.2). 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. 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 7.4.

7.3.3 SDR-Based Metrics

The 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 [205]. 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 Sound Demixing (SDX) Challenge [60], also denoted as SDR:

$$\mathrm{SDR}(x, \hat{x}) := 10 \log_{10} \frac{\|x\|^2}{\|\hat{x} - x\|^2}. \tag{7.2}$$

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:

$$\mathrm{SDR}_{\mathrm{local}} := \frac{1}{K} \sum_{k=1}^{K} \mathrm{SDR}(x_k, \hat{x}_k) \tag{7.3}$$

Table 7.3: $\mathrm{SDR}_{\mathrm{local}}$ values (mean and standard deviation) averaged over all PCD excerpts for different MSS systems (see Table 7.1).

Model	Piano	Orchestra
UMX20	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**

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 (7.1) and consider notewise SDR values:

$$\mathrm{SDR}_{\mathrm{note}} := \mathrm{SDR}(x^m, \hat{x}^m), \tag{7.4}$$

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 $m \in [1:M]$.

7.4 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 7.2, 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 on the source separation quality of the MSS systems (Section 7.4.1). Subsequently, we adopt a more fine-grained perspective, delving into the separation quality depending on the musical pitch (Section 7.4.2). Finally, we assume an excerptwise view and discuss specific examples to illustrate how separation errors may occur in musically complex situations (Section 7.4.3). 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.

7.4.1 Global Perspective

To gain an initial understanding of the overall performance of the five MSS systems, Table 7.3 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 Chapter 6.

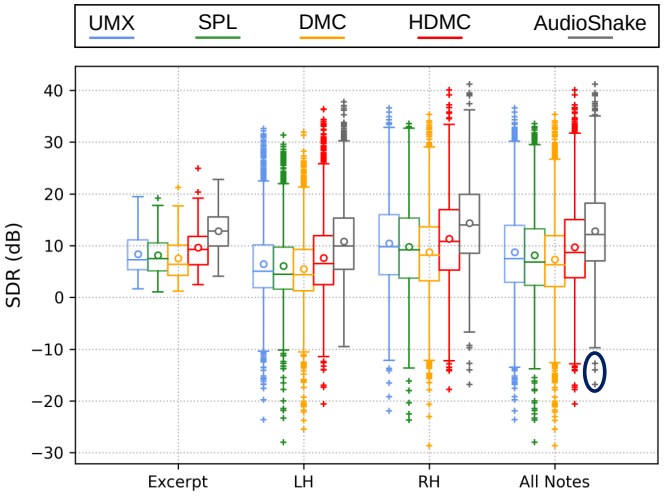

Figure 7.3: Comparison of different evaluation methodologies for the piano case using boxplots. The three outliers for `AudioShake`, indicated by the black oval, are shown in Figure 7.8.

In the subsequent finer-grained evaluation, we employ notewise evaluation metrics. Since we have the required symbolic score information for the score-based NMF decomposition exclusively for the piano tracks, we confine our analysis to the piano case. 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 chapter. Extending the evaluation methodology for the five MSS systems, Figure 7.3 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 $\text{SDR}_{\text{local}}$ values computed as in Table 7.3. The second (LH) and third (RH) groups show the SDR_{note} values for the left-hand and right-hand notes, respectively, and the last group (All Notes) shows the SDR_{note} values for all individual notes.

While the general trends for the five MSS systems are similar to those shown in Table 7.3, the different evaluation methodologies provide additional information. Firstly, being based on notewise aggregation, outliers in the SDR_{note}-based boxplots offer explicit cues worth further investigation. For instance, outliers such as the three indicated by the black oval in Figure 7.3 yield interesting examples for musically complex passages as further explored in Section 7.4.3. The boxplots in Figure 7.3 also facilitate a comparison of SDR_{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 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.

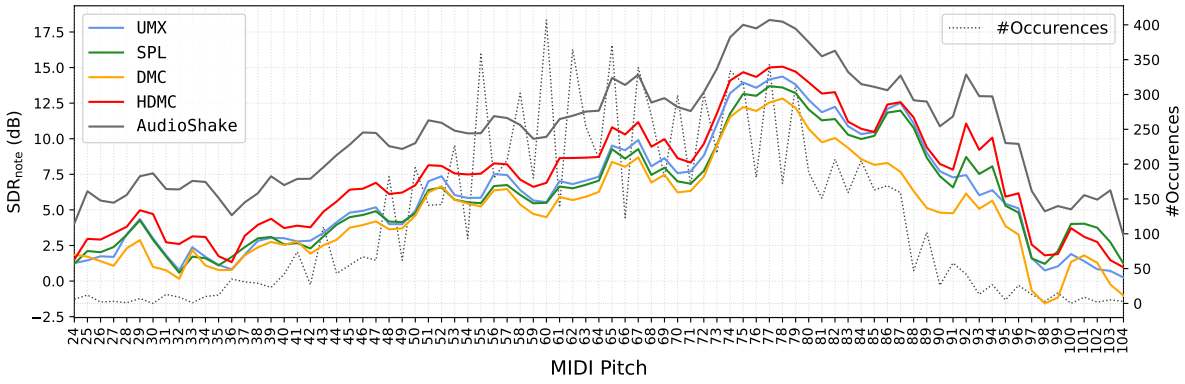

Figure 7.4: SDR_{note} values aggregated by pitch (specified by MIDI note number) shown for five MSS systems.

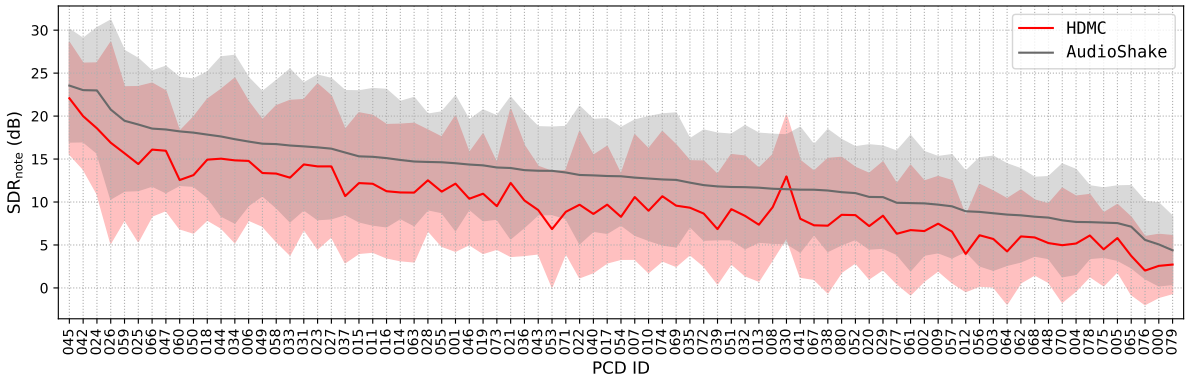

Figure 7.5: SDR_{note} values aggregated by excerpt (specified by PCD ID) shown for the two best-performing MSS systems, `HDMC` and `AudioShake`. The mean (solid line) and standard deviations (filled regions) are indicated. The excerpts are sorted based on decreasing mean values with regard to `AudioShake`.

7.4.2 Pitchwise Evaluation

Considering that RH typically contains higher notes than LH, one may conjecture that source separation quality depends on the pitch of the notes. To test this hypothesis, Figure 7.4 provides an overview of the SDR_{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 SDR_{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 7.4 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.

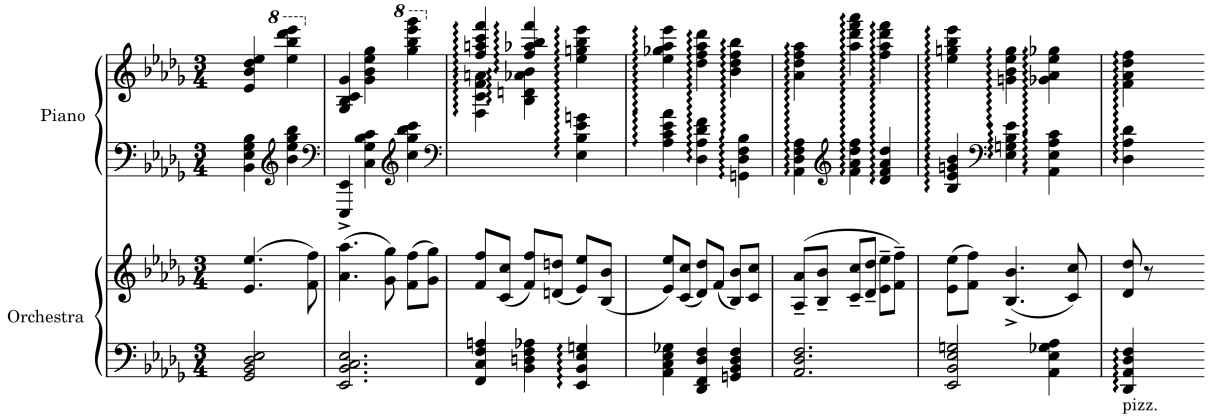

Figure 7.6: Excerpt with PCD ID 079: Tchaikovsky's Piano Concerto Op. 23, measures 18–24 of the first movement.

7.4.3 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 7.5 $\mathrm{SDR}_{\mathrm{note}}$ 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. For a more detailed account of the excerptwise results, see also Table A.1.

Guided by the plot in Figure 7.5, 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, we 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 7.6. Evidently, this passage exhibits a high musical complexity, with both piano and orchestra playing numerous 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 $\mathrm{SDR}_{\mathrm{note}}$ value. This excerpt corresponds to the first measures of Bach's Piano Concerto BWV 1056 (see Figure 7.7), where the piano and orchestra play many notes in

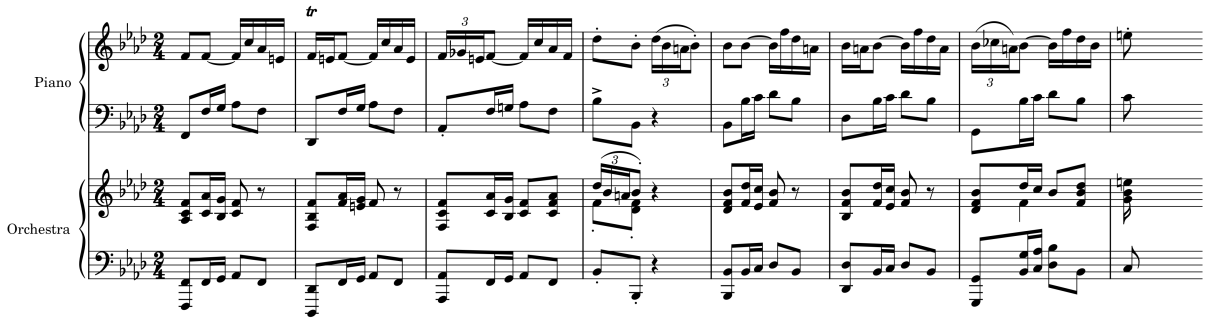

Figure 7.7: Excerpt with PCD ID 000: Bach's Piano Concerto BWV 1056, measures 1–8 of the first movement.

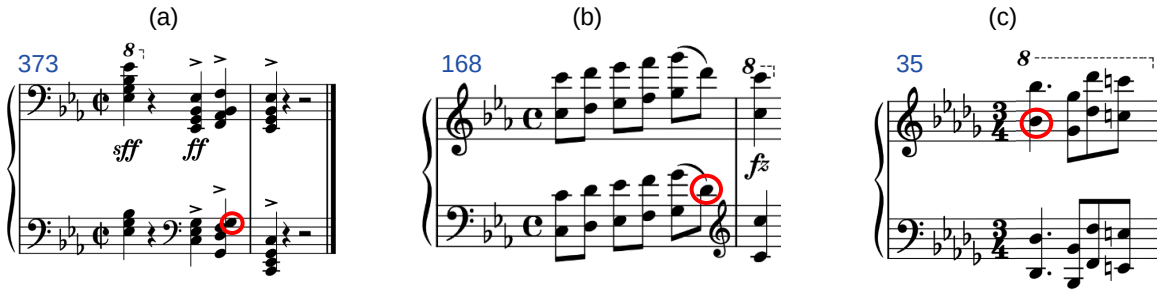

Figure 7.8: Musical context within the piano scores for the three notewise outliers marked in Figure 7.3 (here indicated by the red circles). **(a)** PCD ID: 052. **(b)** PCD ID: 061. **(c)** PCD ID: 077.

unison. This scenario represents one of the most challenging situations for source separation models to deal with [20].

Finally, we revisit the boxplots shown in Figure 7.3, where we marked three outliers indicating problematic notewise sound events with low SDR values, poorly separated by `AudioShake`. Figure 7.8 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 to handle.

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

7.5 Conclusion

In this chapter, 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. 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 an interdisciplinary dialogue between engineering and music experts.

8 Nonnegative Autoencoders for Efficient Audio Decomposition

This chapter is based on [134]. The first author Yigitcan Özer is the main contributor to this article. In collaboration with Jonathan Hansen, Tim Zunner, and his supervisor Meinard Müller, he devised the ideas, developed the formalization, and wrote the paper. Furthermore, Yigitcan Özer implemented all approaches and conducted the experiments.

NMF is a powerful technique for decomposing a music recording's magnitude spectrogram into musically meaningful spectral and activation patterns. In recent years, musically informed NMF-based audio decomposition has been simulated using neural networks, which opens up new paths of exploiting recent deep learning frameworks, including libraries for efficient gradient computations. In this chapter, we continue this strand of research by considering NAEs in combination with gradient projection and structured dropout techniques. Conducting experiments based on piano recordings, we compare the decomposition results of NAE-based approaches with those obtained from a score-informed NMF variant. In this context, we examine various gradient descent methods using fixed and adaptive learning rates for deriving the NAE encoder and decoder parameters. Among others, we show how the famous multiplicative update rules for NMF can be transferred to the case of NAEs. The overall goal of our contribution is to illustrate the benefits and limitations of the various techniques concerning implementation issues, convergence speed, and overall runtime.

The remainder of the chapter is organized as follows. Following the introduction in Section 8.1, we provide an overview of the score-informed NMF approach that serves as our reference in Section 8.2. In Section 8.3, we investigate the simulation of NMF through NAE and introduce the multiplicative update rules for NAE. In Section 8.4, we report on our systematic experiments and conclude in Section 8.5 with prospects on future work.

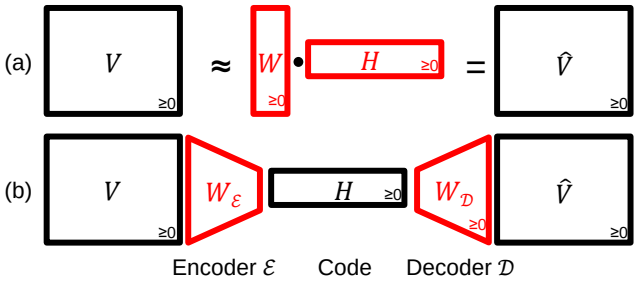

Figure 8.1: (a) NMF used for decomposing a nonnegative matrix V into the product of a nonnegative template matrix W and nonnegative activation matrix H. **(b)** Simulation of the decomposition using NAE (see the text for details). The learned components are shown in red.

8.1 Background

NMF is a prominent low-rank factorization method that imposes nonnegativity constraints in all matrices involved. Notably, its effectiveness and ability to yield *interpretable* results have attracted great attention in various research fields [68, 101]. In the context of music processing, NMF has been widely applied for the decomposition of complex sound mixtures, using the magnitude spectrogram of music signals as input representation [9, 44, 58, 63, 82, 102, 175, 177]. As a result of the decomposition, NMF approximates the magnitude spectrogram by the product of two nonnegative matrices, 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*).

Motivated by recent advances in designing and training neural networks, Smaragdis and Venkataramani [176] introduced a NAE architecture as a neural network alternative for NMF-based audio decomposition. Figure 8.1 gives an overview of the simulation of NMF through a shallow NAE architecture, which comprises a single-layer encoder and a single-layer decoder. The NAE decoder directly corresponds to the NAE template matrix. However, rather than learning an activation matrix as in NMF, the NAE learns an encoder which yields an activation matrix as output (also called *code*). To ensure the nonnegativity constraints of templates and activations, one can combine NAEs with gradient projection and structured dropout techniques [56]. The simulation of NMF through NAE makes it possible to exploit recent deep learning frameworks including libraries for automatic and GPU-accelerated gradient computations. This may also open up new paths for tackling the audio decomposition problem with deeper and more complex models.

As starting point of this chapter, we consider the work by Ewert and Müller [54], which uses a score-informed NMF variant for decomposing the magnitude spectrograms of piano recordings. As the main contribution of this chapter, we simulate this original approach by considering different NAE variants (inspired by [56, 176]) and conduct systematic experiments to compare the resulting decompositions with the NMF-based approach used as a reference. In particular, we show how one can adapt the famous multiplicative update rules of NMF [101] to the case of NAEs. Furthermore, we investigate projected versions of additive gradient descent methods such as stochastic gradient descent (SGD), root mean square propagation (RMSprop) [193], and ADAM [96]. Our systematic experiments highlight the benefits and

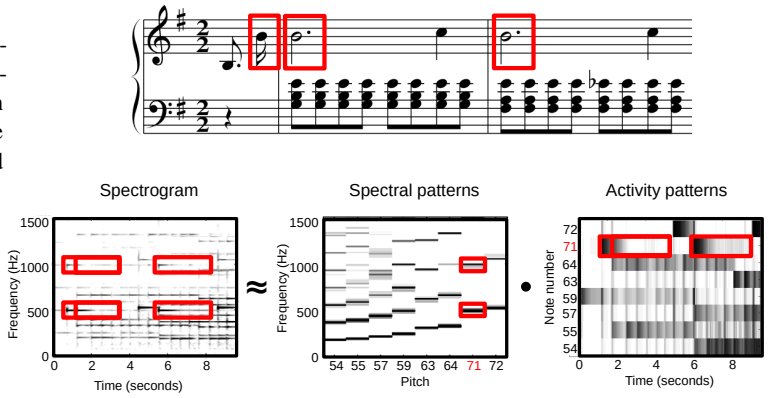

Figure 8.2: Decomposing the magnitude spectrogram of an audio excerpt of Chopin's Prelude Op .28 No. 4 into template and activation matrices. The information related to the note number $p = 71$ (B4) is indicated by the red rectangular frames.

limitations of different techniques in terms of implementation issues, convergence speed, and overall runtime.

8.2 Score-Informed NMF for Audio Decomposition

NMF is a nonnegative factorization algorithm that accounts for an additive, part-based representation of a nonnegative input matrix. Nonnegative matrix entries prevent undesired effects such as destructive inferences, where a positive component might be canceled out by adding a kind of inverse (negative) component.

Given the magnitude spectrogram of a music recording $V \in \mathbb{R}_{\geq 0}^{K \times N}$ and a *target rank* $R \in \mathbb{N}$ that is much smaller than both $K \in \mathbb{N}$ and $N \in \mathbb{N}$, NMF seeks an optimal approximation $V \approx WH$ enforcing both learned matrices $W \in \mathbb{R}^{K \times R}$ and $H \in \mathbb{R}^{R \times N}$ to be nonnegative. As shown in Figure 8.2, W indicates the *template matrix* and H the *activation matrix*, where K and N, respectively, denote the number of frequency bins and time frames in the input spectrogram. In this example, the target rank R corresponds to the number of distinct pitches played in the input music recording. The loss function of the least-square optimization problem (with additional nonnegativity constraints for W and H) can be written as

$$\varphi(W, H) = \|V - WH\|_F^2, \tag{8.1}$$

where $\|\cdot\|_F$ is the Frobenius norm.

A common method when using the NMF algorithm is the *alternating least squares (ALS)*, an optimization technique, where the first matrix W is updated with fixed H, and then H is updated with fixed W, in iterative cycles until a stopping criteria is fulfilled. In particular, ALS is enhanced by the multiplicative update rules [101], which offers a straightforward and efficient implementation. The crucial idea is to use an adaptive learning rate, which transforms the additive update rules of the usual gradient descent to

multiplicative ones, resulting in

$$\begin{aligned} H &\leftarrow H \odot (W^\top V) \oslash (W^\top W H + \varepsilon), \\ W &\leftarrow W \odot (V H^\top) \oslash (W H H^\top + \varepsilon), \end{aligned} \tag{8.2}$$

for the case of the Euclidean loss. Here, $\odot$ and $\oslash$ denote pointwise multiplication and division, respectively. The parameter ε denotes the machine epsilon, which is used to avoid division by 0.

Besides nonnegativity constraints, prior musical knowledge, e. g., coming from a musical score, can also be easily integrated into the learning process of NMF to guide the decomposition [58, 63, 91, 149, 156]. Multiplicative update rules in Equation (8.2) ensure that the zero-valued matrix entries in the template and activation matrices remain zero during the entire learning process. Therefore, one can avoid undesired template and activation values by initiating the corresponding positions in the matrices with zero. In [54], the templates are initialized using a sparse, binary matrix $W^{\mathrm{C}} \in \{0, 1\}^{K \times R}$ to constrain frequencies and enforce an overtone model. Similarly, using the score information, the activation matrix can be constrained through a sparse, binary matrix $H^{\mathrm{C}} \in \{0, 1\}^{R \times N}$. As an example, the red boxes in Figure 8.2 indicate spectral and activation constraints (initialized with one-values inside and with zero-values outside the red boxes) corresponding to the note number $p = 71$ (B4).

This chapter uses the multiplicative NMF with Euclidean loss as the reference model. We apply the same score-informed initialization procedure as described in [54].

8.3 Simulation via Constrained NAEs

Following [56, 176], we now show how one can simulate constrained NMF via an NAE model in combination with projected gradient descent methods and rectifier activation functions.

The NMF model can be reformulated through a simple linear autoencoder [84, 176] as

$$\begin{aligned} H &= W_{\mathcal{E}} V, \\ \hat{V} &= W_{\mathcal{D}} H. \end{aligned} \tag{8.3}$$

The matrix $W_{\mathcal{E}} \in \mathbb{R}^{R \times K}$ denotes the *encoder*, which yields the activation matrix H as output. The *decoder* $W_{\mathcal{D}} \in \mathbb{R}^{K \times R}$ can be thought of as the equivalent to the template matrix W in the NMF decomposition. To ensure the nonnegativity of the activation output matrix H and the template weight matrix $W_{\mathcal{D}}$ in NAE, one has to introduce further constraints.

Our proposed NAE model applies a ReLU after the encoder layer as in [176] to ensure the nonnegativity of the activation matrix H. For the nonnegativity of the decoder matrix $W_{\mathcal{D}}$, we use a projected gradient descent method as in [108], setting the negative values in $W_{\mathcal{D}}$ to zero during training. Chorowski and Zurada [33] state that constraining the weight matrices to be nonnegative improves the interpretability of an

autoencoder's operation, whereas it does not lower the network's capability. In contrast, our experiments showed that applying a simple ReLU after the encoder layer resulted in a better convergence, rather than using projected gradients for the encoder layer as well.

As in the NMF case, prior music knowledge can also be integrated into the NAE model to guide the learning process. Ewert and Sandler introduced the *structured dropout* for activation constraints in [56]. Dropout layers typically regularize networks to avoid overfitting by randomly setting neurons to zero during the training process [179]. In contrast, structured dropout imposes prior musical knowledge and selectively removes undesired activations by setting

$$H' = H^{\mathrm{C}} \odot H. \tag{8.4}$$

To enforce structured dropout, one can adapt the loss function in Equation (8.1) to the constrained NAE case as follows:

$$\begin{aligned} \varphi(W_{\mathcal{E}}, W_{\mathcal{D}}) &= \|V - W_{\mathcal{D}} H'\|_F^2 \\ &= \|V - W_{\mathcal{D}}(\sigma(W_{\mathcal{E}} V) \odot H^{\mathrm{C}})\|_F^2, \end{aligned} \tag{8.5}$$

where σ denotes the ReLU activation function. Computing the gradients with respect to the encoder and decoder matrices, one can derive multiplicative update rules for this NAE model similar to the NMF case:

$$\begin{aligned} W_{\mathcal{E}} &\leftarrow W_{\mathcal{E}} \odot \Bigg(\Big(\big((W_{\mathcal{D}}^{\top} V) \odot H^{\mathrm{C}} \big) V^{\top} \Big) \oslash \\ &\qquad \Big(\big((W_{\mathcal{D}}^{\top} W_{\mathcal{D}} H') \odot H^{\mathrm{C}} \big) V^{\top} + \varepsilon \Big) \Bigg), \\ W_{\mathcal{D}} &\leftarrow W_{\mathcal{D}} \odot \big((V H'^{\top}) \oslash (W_{\mathcal{D}} H' H'^{\top} + \varepsilon) \big). \end{aligned} \tag{8.6}$$

For a derivation of the multiplicative update rules for NAE, we refer to [215]. We call this model as *multiplicative NAE*.

To train an NAE, additive methods like SGD, in which the learning rate remains constant during training, can also be used. Another alternative is the integration of other adaptive strategies for optimizers such as RMSprop [193] and ADAM [96], which adjust the learning rate during training.

As for multiplicative NMF, the multiplicative NAE has the property that zero-valued entries remain zero. To enforce this property for template weights $W_{\mathcal{D}}$ also in the case of using additive update rules, we add further projection by applying binary masking on $W_{\mathcal{D}}$ using the constraint matrix W^{C}:

$$W_{\mathcal{D}} \leftarrow W_{\mathcal{D}} \odot W^{\mathrm{C}}. \tag{8.7}$$

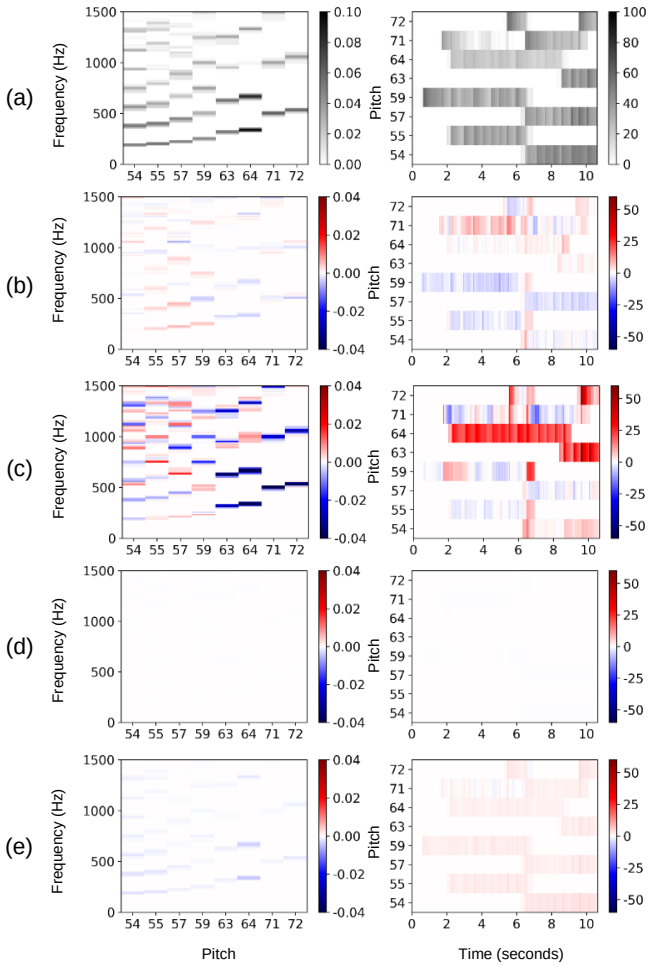

Figure 8.3: Continuation of our Chopin example from Figure 8.2. **(a)** Template matrix W (left) and activation matrix H (right) learned by the score-informed NMF model. **(b)-(e)** Difference between template (left) and activation (right) matrices obtained from NMF (used as reference) and NAE-based approaches. The columns of W and $W_{\mathcal{D}}$ are ℓ^1-normalized. **(b)** NAE trained with multiplicative update rules. **(c)** NAE trained with SGD with a fixed learning rate $\gamma = 0.1$. **(d)** NAE trained with ADAM. **(e)** NAE trained with RMSprop.

8.4 Experiments

This section reports on our experiments comparing various NAE-based approaches with the score-informed NMF model used as reference. To this end, we decompose the magnitude spectrograms of piano recordings into musically meaningful spectral vectors and their activations. In our experiments, we use eight publicly-available, nonsynthetic piano recordings using the same experimental setting as in [54][17]. These pieces are listed in Table 8.1. The music recordings are in mono format, sampled at 22.05 kHz, with durations ranging from approximately 100 seconds to 9 minutes.

During the preprocessing phase, we compute the magnitude spectrograms of each recording, using a Hann window of size 2048 and a hop size of 1024. For the reference NMF model, we employ the same initialization procedure as described in [54]. Similarly, we initialize the decoder matrix $W_{\mathcal{D}}$ of NAEs using the binary constrained matrix W^{C}, while we initialize and the encoder matrix $W_{\mathcal{E}}$ randomly. At the end of the training of each model, we ℓ^1-normalize the columns of the learned matrices W and $W_{\mathcal{D}}$,

[17] `http://resources.mpi-inf.mpg.de/MIR/ICASSP2012-ScoreInformedNMF/`

	Model				
File ID	NMF	$\text{NMF}_{\text{Mult.}}$	NMF_{SGD}	NMF_{ADAM}	$\text{NMF}_{\text{RMSprop.}}$
Chopin_Op028-04_SMD	46.4	49.0	62.4	57.6	48.1
Chopin_Op028-15_SMD	48.5	53.2	67.3	66.2	50.9
Chopin_Op066_SMD	79.5	87.2	139.0	101.1	85.7
Beethoven_Op031No2-01_SMD	90.7	99.2	105.1	104.4	94.9
Chopin_Op028-01_SMD	94.8	103.6	299.9	122.8	97.4
Bach_BWV875-01_SMD	97.5	107.3	219.9	129.2	104.3
Beethoven_Op111-01_EA	103.7	129.4	328.5	148.4	113.0
Chopin_Op064No1_EA	131.9	145.9	383.6	161.6	137.2

Table 8.1: Approximation error between V and $\hat{V}$ (columnwise average) of NMF and NAE-Based Approaches

and accordingly scale the columns of the activation matrix H. This normalization and rescaling accounts for the scale ambiguity inherent in NMF decomposition and makes the decomposition results better comparable.

In the following, we regard an iteration to be the update of both the matrices W and H in the NMF case, and similarly $W_{\mathcal{E}}$ and $W_{\mathcal{D}}$ in the NAE case. In our experiments, we performed $10,000$ iterations in the training phase. During training, we used a learning rate of $\gamma = 0.1$ for the SGD, and the recommended default values for the RMSprop [193] and ADAM [96] optimizers.

Implementing the multiplicative NMF and NAE is straightforward, following the derived update rules in Equation (8.2) and Equation (8.6), respectively. We implemented the multiplicative models with NumPy using matrix operations. Furthermore, we used the Tensorflow library to exploit the automatic gradient computation and GPU acceleration to train the NAE models that use additive gradient descent techniques. For the GPU-based computations we used a single NVIDIA GTX 1080 Ti GPU.

To get a first impression of the approximation behavior of the various decomposition approaches, Figure 8.3 shows a comparison of learned template and activation matrices learned by the reference NMF model and ones learned by the various NAE-based approaches. First, note that Multiplicative NMF and NAE yield similar template and activation matrices. Furthermore, among the additive NAE-based approaches, the NAE variant trained with SGD leads to the worst results compared to the NMF reference. Our comparison also indicates that NAEs trained with adaptive gradient descent methods lead to template and activation matrices close to the NMF case when using a huge number of iterations (up to $10,000$ in our experiments).

Next, we compare the approximation quality of the various decompositions in a quantitative fashion. Table 8.1 shows a comparison of approximation errors between V and $\hat{V}$ yielded by the NMF reference and NAE-based approaches based on the entire dataset. Here, each entry indicates the average columnwise ℓ^1-error between the approximation matrix $\hat{V}$ and the input spectrogram V. For example, the first row shows the results obtained from the spectrogram decomposition of the entire recording of Chopin's Prelude Op.28 No.4. The multiplicative NMF results in the approximation error of 46.4, and the multiplicative NAE in a similar value of 49.0. Among the NAE variants trained with additive gradient descent techniques, RMSprop reaches the smallest approximation error of 48.1, whereas SGD results in

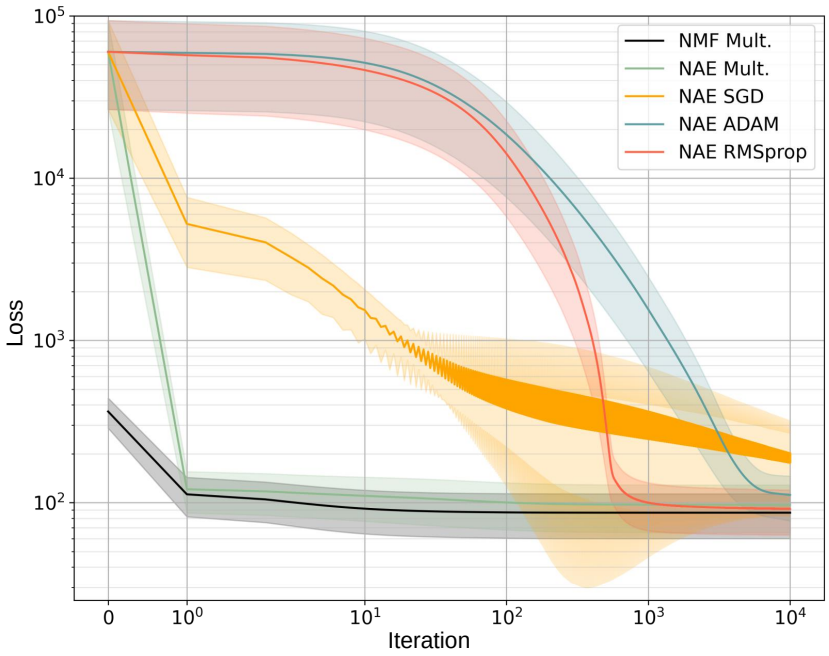

Figure 8.4: Average column-wise absolute approximation loss between $\hat{V}$ and V per iteration, evaluated on the entire dataset. All the NAE variants use the same weight initialization procedure.

the highest approximation error of 62.4. Moreover, we can infer that the reference NMF model and NAE variants perform similarly over the entire dataset: the multiplicative NMF leads to the best approximation, while NAE with RMSprop results in the smallest approximation error among the NAE variants.

In our next experiment, we analyze the convergence behavior of all approaches over the number of iterations. Figure 8.4 illustrates the mean and standard deviations per iteration over the columnwise Euclidean error, evaluated using all eight recordings in the dataset. The rapid decay in error after the first iteration of the multiplicative models is remarkable, whereas additive NAE variants need more iterations until they reach a steeper decline in the error. NAE with SGD shows a slow and unstable convergence behavior, resulting in a poor approximation even after $10,000$ iterations. We also have tried using other learning rates for the SGD case; however, it is unclear how to choose an optimal learning rate that guarantees convergence. $\gamma = 0.1$ has shown the best performance among various learning rates. In contrast, NAE with ADAM converges after $10,000$ iterations to a similar decomposition as NMF. Similarly, NAE with RMSprop converges to this result after only $1,000$ iterations. It is also worthwhile to note that the both adaptive NAE variants reach a decomposition result as NMF, although NAEs learn fewer parameters than the NMF reference. (The encoder $W_{\mathcal{E}} \in \mathbb{R}^{R \times K}$ has usually much fewer parameters than the activation matrix $H \in \mathbb{R}^{R \times N}$.)

Finally, we compare in Figure 8.5 the training runtime of the multiplicative NAE and NAE with RMSprop. Although the multiplicative NAE shows a very steep decay within the first second, NAE with RMSprop trained on CPU and GPU both outperform the multiplicative model after around 100 seconds. Additionally,

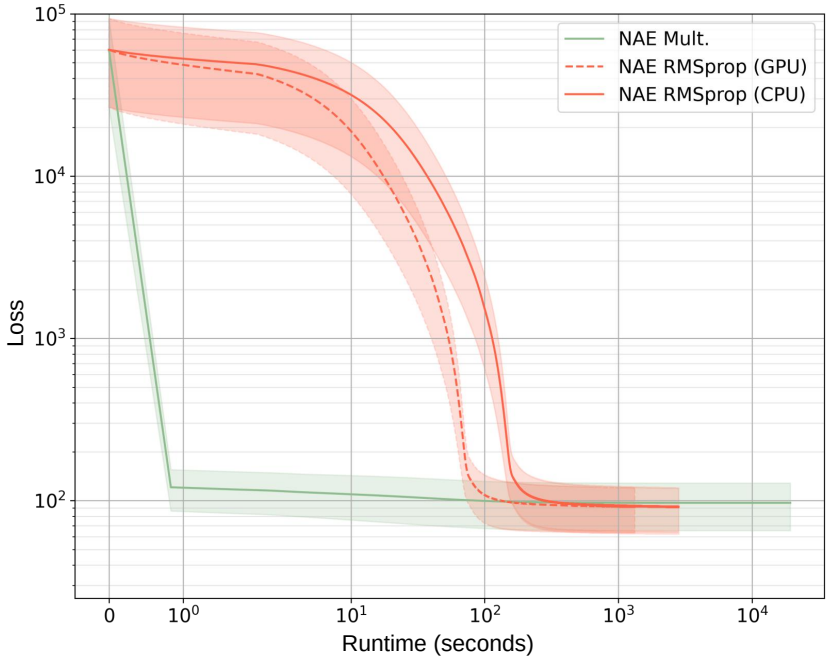

Figure 8.5: Runtime comparison of the multiplicative NAE and NAE variants trained with RMSprop. The NAE variant trained on GPU with RMSprop is shown with dashed lines.

the gradient computation of the multiplicative update rules for NAEs becomes challenging for deeper networks. The implementation of NAE with RMSprop, on the other hand, exploit the automatic gradient computation. We also see that the GPU-accelerated model converges twice as fast as the NAE with RMSprop trained on central processing unit (CPU). The hardware acceleration becomes more evident in the case of deeper and more complex networks, which involve more matrix multiplications.

8.5 Conclusion

In this chapter, we investigated different NAE-based approaches for decomposing piano recordings into musically meaningful spectral vectors. We simulated the reference score-informed NMF model with various NAE-based methods. We showed that NAEs acquire higher efficiency through hardware-accelerated frameworks while yielding similar results as the reference NMF model. We also explored different adaptive gradient technique methods, including multiplicative rules for NAEs. We showed that the GPU-accelerated, adaptive RMSprop method outperformed other NAE variants in terms of the approximation quality and efficiency, while the learned templates and activations remain similar to those of the NMF reference. In the future, we aim to develop deeper and more complex models, which result in faster and better convergence and preserve interpretability. This will enable the design of explainable deep learning models, as our constrained NAE, while improving the performance of the network.

9 Summary and Future Work

This thesis investigated computational techniques and application scenarios for source separation in music, with a particular focus on separating piano concerto recordings into piano and orchestral tracks. To tackle this task, we adapted and evaluated a multitude of DL-based models, which were originally designed for separating speech signals and popular music recordings. A key challenge we encountered has been the need for a large training dataset, which, in the case of MSS, involves multitrack recordings with (isolated) individual sources or stems. Since most of the open-source datasets containing isolated stems are limited to popular music, we generated artificial training examples through random mixing and introduced musically motivated data augmentation approaches to enhance the separation performance. Our findings indicate that the hybrid model, trained with a full suite of augmentation techniques, yields the best source separation performance. To further enable the quantitative and qualitative evaluation of piano concerto separation, we created the multitrack dataset PCD, incorporating music synchronization and beat tracking. Additionally, we introduced a novel quantitative evaluation approach for MSS systems, which uses SDR results based on the decomposed note events to gain a deeper understanding of source separation artifacts. This approach provided better insights into both the common failure modes of the source separation and the musical complexities inherent in the excerpts.

Our quantitative and subjective evaluation results suggest that there is still room for improvement, which could be achieved through more complex model architectures and a larger training dataset comprising multitrack recordings. While random mixing can be a part of a feasible training strategy, it introduces additional challenges to MSS models, especially in terms of room acoustics in different recordings. This approach tends to misattribute background noise and reverberation to the timbre of musical sources. In order to overcome this issue, accurate estimation of acoustic parameters is crucial, as it can provide insightful information about the acoustic environment and the sound quality within recording settings [79]. Such information can guide the MSS models and pave the way for developing a more coherent and effective training dataset [111].

Many source separation approaches are based on the assumption of a linear mixture model (see Section 6.3), which defines the input signal to an MSS model as the superposition of separated musical sources. This may not necessarily be desired from an application perspective, since, for example recording artifacts in the input mixture should ideally be discarded. Generative approaches, on the other hand, offer greater freedom and creativity, moving beyond the limitations imposed by the acoustic properties of an input mixture. For example, rather than relying on Wiener Filtering for estimated magnitude spectrograms, or

directly applying the iSTFT to complex spectrograms of separated sources, generative enhancements can be applied in a post-processing step, e. g., via denoising autoencoders [120], flow-based models [187], generative adversarial networks (GANs) [139], or diffusion-based strategies [104]. Alternatively, one may draw inspiration from an analysis-by-synthesis approach to resynthesize the constituent sources in a music recording [173]. In particular, techniques like pitch estimation [94, 211] or automatic transcription [80, 115, 212] can be utilized to resynthesize the sources with a differentiable parametric source model, e. g., with Differential Digital Signal Processing (DDSP) [53, 155, 171]. Furthermore, given the synchronized MIDI of an input mixture, one can employ diffusion-based approaches for multi-instrument synthesis, incorporating *performance conditioning*. This involves synthesizing music with the style and timbre of specific instruments derived from other performances [116].

Our vision for this work was that, through source separation, pianists would have the ability to choose a piano concerto recording and extract the orchestra track to play along with. However, for a truly engaging user experience, this alone may not suffice due to the varied tempo and dynamics inherent in classical music performances. The tempo decisions made by performers at both a global and local level make their interpretations unique and enrich the way they perform. In a real-life recording process, the auditory and visual interaction between the pianist and other musicians plays a crucial role in achieving optimal synchronization and cohesion between the piano and orchestra [40, 70]. To create their own mix in an *offline* fashion, the pianists can first freely perform and record their part of the chosen piano concerto. This recorded piano track can then be synchronized with a separated or reconstructed orchestral track using music synchronization techniques and TSM [49, 200].

The final step towards our envisioned scenario is the development of an interactive, real-time accompaniment system, which could be utilized not only for artistic performances but also for practice sessions. While we focused on adapting the orchestral track to the piano performer similar to [154], there has been a growing interest in real-time accompaniment systems that adapt to the soloist, for example, based on the rendition of a score with a Disklavier in an expressive manner [28]. Performing with an interactive accompaniment system not only enhances the way pianists can interact with classical music performances, but also combines a variety of challenges in MIR, such as real-time score following [1, 153], beat tracking [41], music synchronization [46], and expressive music performance [27] among others.

Appendix

A Excerptwise Evaluation of Source Separation

Table A.1 presents a comprehensive overview of the PCD, including the PCD IDs and names of excerpts following the naming conventions outlined in Section 5.4.2. Additionally, the table details the excerptwise SDR results for the two best-performing models `HDMC` and `AudioShake` (see also Section 7.4.3).

PCD ID	Excerpt Name	HDMC			AudioShake		
		Piano		Orchestra	Piano		Orchestra
		SDR_{note}	SDR_{local}	SDR_{local}	SDR_{note}	SDR_{local}	SDR_{local}
000	Bach_BWV1056-01-mm001-008_YO-V1	2.57 ± 3.66	2.44	5.66	5.07 ± 4.82	4.38	7.60
001	Bach_BWV1056-01-mm021-028_YO-V1	12.14 ± 7.87	8.81	7.50	14.51 ± 7.82	10.49	9.18
002	Bach_BWV1056-01-mm047-054_YO-V1	6.62 ± 5.82	5.20	2.69	9.84 ± 6.02	7.87	5.36
003	Bach_BWV1056-01-mm079-086_YO-V1	5.70 ± 5.57	4.59	5.37	8.69 ± 6.63	7.15	7.93
004	Bach_BWV1056-01-mm111-116_YO-V1	5.17 ± 5.51	2.92	6.16	7.68 ± 6.10	4.11	7.34
005	Bach_BWV1056-01-mm001-008_YO-V2	5.81 ± 3.90	6.16	5.06	7.55 ± 4.32	7.25	6.15
006	Bach_BWV1056-01-mm021-028_YO-V2	14.78 ± 6.89	11.26	5.57	17.04 ± 7.43	12.93	7.24
007	Bach_BWV1056-01-mm047-054_YO-V2	10.58 ± 7.23	9.36	0.85	12.84 ± 6.69	11.97	3.46
008	Bach_BWV1056-01-mm079-086_YO-V2	9.40 ± 6.10	7.62	4.27	11.54 ± 6.34	9.79	6.44
009	Bach_BWV1056-01-mm111-116_YO-V2	7.48 ± 5.48	5.22	4.76	9.69 ± 5.59	7.11	6.65
010	Beethoven_Op015-01-mm118-125_MM	8.99 ± 7.19	6.89	6.41	12.74 ± 7.19	9.92	9.44
011	Beethoven_Op015-01-mm167-174_MM	12.12 ± 7.95	10.56	6.12	15.26 ± 7.94	13.06	8.63
012	Beethoven_Op015-01-mm295-302_MM	3.95 ± 4.37	5.21	8.55	8.92 ± 4.66	9.20	12.54
013	Beethoven_Op015-01-mm306-313_MM	7.36 ± 6.56	7.66	7.32	11.67 ± 6.33	11.35	11.03
014	Beethoven_Op015-01-mm363-370_MM	11.11 ± 7.93	5.24	7.16	14.89 ± 6.80	13.39	14.10
015	Beethoven_Op015-01-mm382-389_MM	12.20 ± 8.17	11.10	5.31	15.32 ± 7.61	14.12	8.33
016	Beethoven_Op019-01-mm095-102_ES-V1	11.26 ± 7.75	11.99	1.64	15.11 ± 7.99	17.37	7.02
017	Beethoven_Op019-01-mm117-124_ES-V1	9.69 ± 6.78	9.29	6.98	13.03 ± 6.66	12.45	10.14
018	Beethoven_Op019-01-mm095-102_ES-V2	14.93 ± 7.02	14.57	2.60	17.85 ± 7.28	18.45	6.46
019	Beethoven_Op019-01-mm117-124_ES-V2	10.97 ± 6.95	9.33	7.33	14.27 ± 6.44	11.86	9.86
020	Beethoven_Op037-01-mm124-130_ES-V1	7.20 ± 6.13	5.73	5.51	10.59 ± 6.06	9.80	9.58
021	Beethoven_Op037-01-mm148-154_ES-V1	12.22 ± 8.55	11.41	3.49	13.96 ± 8.23	13.10	5.17
022	Beethoven_Op037-01-mm124-130_ES-V2	9.69 ± 8.49	9.00	4.50	13.15 ± 7.99	12.79	8.30
023	Beethoven_Op037-01-mm148-154_ES-V2	14.14 ± 9.59	13.35	2.88	16.36 ± 8.40	16.09	5.60
024	Beethoven_Op037-02-mm030-032_LR	18.58 ± 7.60	17.32	4.73	22.99 ± 7.29	21.91	9.25
025	Beethoven_Op058-02-mm031-035_ES-V1	14.41 ± 9.01	17.14	4.29	19.02 ± 7.68	20.55	7.32
026	Beethoven_Op058-02-mm031-035_ES-V2	16.90 ± 11.67	18.88	4.92	20.76 ±10.40	21.68	7.29
027	Chopin_Op021-03-mm003-016_ES	14.15 ± 8.19	13.61	3.15	16.21 ± 8.15	15.78	5.30
028	Chopin_Op021-03-mm145-157_ES	12.53 ± 5.84	14.50	3.44	14.66 ± 5.55	16.45	5.39
029	Chopin_Op021-03-mm215-229_ES	8.41 ± 6.26	8.48	4.49	10.57 ± 5.93	10.58	6.59
030	Chopin_Op021-03-mm249-260_ES	12.98 ± 7.13	9.51	2.34	11.49 ± 6.31	12.60	5.42
031	Chopin_Op021-03-mm411-424_ES	14.37 ± 7.54	15.10	2.92	16.48 ± 7.37	17.08	4.90
032	Grieg_Op016-01-mm027-030_ES	8.38 ± 6.96	11.07	6.51	11.73 ± 7.15	13.30	8.65
033	Mendelssohn_MWVO007-01-mm023-029_ES-V1	12.83 ± 8.99	15.59	2.17	16.58 ± 8.87	19.24	6.10
034	Mendelssohn_MWVO007-01-mm023-029_ES-V2	14.85 ± 9.55	18.30	3.57	17.31 ± 9.75	19.30	4.84
035	Mozart_KV414-01-mm180-186_YO	9.34 ± 5.45	8.41	5.02	12.25 ± 5.09	11.16	7.77
036	Mozart_KV414-01-mm192-196_YO	10.18 ± 6.39	11.67	4.92	13.71 ± 6.64	14.47	7.57
037	Mozart_KV467-01-mm145-151_YO	10.68 ± 7.75	9.96	9.30	15.76 ± 7.20	15.61	14.95
038	Mozart_KV467-01-mm186-192_YO	7.24 ± 7.81	7.88	6.06	11.34 ± 7.08	11.72	9.89
039	Mozart_KV467-01-mm231-237_YO	6.84 ± 6.36	5.34	5.39	11.82 ± 6.19	10.52	10.58
040	Mozart_KV467-01-mm253-259_YO	8.61 ± 6.82	7.47	6.42	13.10 ± 6.50	12.94	11.89
041	Mozart_KV467-01-mm369-375_YO	8.05 ± 6.79	8.61	5.85	11.44 ± 7.25	11.27	8.51
042	Mozart_KV467-02-mm024-026_YO-V1	19.99 ± 6.16	19.16	3.66	23.02 ± 6.00	21.92	6.40
043	Mozart_KV467-02-mm030-032_YO-V1	9.06 ± 5.09	8.29	4.20	13.64 ± 5.11	12.90	8.81
044	Mozart_KV467-02-mm037-039_YO-V1	15.04 ± 8.06	13.46	6.59	17.63 ± 9.25	17.01	10.13
045	Mozart_KV467-02-mm024-026_YO-V2	22.07 ± 6.48	20.35	3.01	23.55 ± 6.57	22.77	5.41
046	Mozart_KV467-02-mm030-032_YO-V2	10.38 ± 5.33	9.26	5.34	14.36 ± 5.18	13.54	9.62
047	Mozart_KV467-02-mm037-039_YO-V2	15.97 ± 6.95	13.13	5.58	18.44 ± 7.35	16.75	9.20
048	Rachmaninoff_Op018-01-mm012-018_JL	5.22 ± 4.56	6.40	1.77	8.19 ± 4.76	9.12	4.50
049	Rachmaninoff_Op018-01-mm114-120_JL	13.37 ± 6.18	14.27	-0.15	16.79 ± 6.05	16.70	2.29

PCD ID	Excerpt Name	HDMC			AudioShake		
		Piano		Orchestra	Piano		Orchestra
		SDR_{note}	SDR_{local}	SDR_{local}	SDR_{note}	SDR_{local}	SDR_{local}
050	Rachmaninoff_Op018-01-mm182-190_JL	13.13 ± 6.74	13.01	0.65	18.07 ± 6.23	16.65	4.29
051	Rachmaninoff_Op018-01-mm245-252_JL	9.17 ± 6.35	9.44	2.76	11.75 ± 6.13	12.00	5.32
052	Rachmaninoff_Op018-01-mm366-374_JL	8.48 ± 5.55	10.76	2.20	11.04 ± 5.39	12.68	4.13
053	Rachmaninoff_Op018-02-mm013-015_JL	6.84 ± 6.71	5.77	4.46	13.63 ± 5.08	13.50	12.19
054	Rachmaninoff_Op018-02-mm034-036_JL	8.28 ± 4.94	4.37	6.66	13.00 ± 5.66	10.41	12.69
055	Rachmaninoff_Op018-02-mm055-059_JL	11.21 ± 6.34	9.29	3.00	14.63 ± 5.85	13.40	7.11
056	Rachmaninoff_Op018-02-mm089-093_JL	6.13 ± 5.92	4.37	4.57	8.85 ± 6.27	7.08	7.29
057	Rachmaninoff_Op018-02-mm116-121_JL	6.56 ± 5.93	7.60	3.72	9.50 ± 6.00	9.98	6.09
058	Rachmaninoff_Op018-03-mm063-073_JL	13.31 ± 7.89	11.11	2.11	16.74 ± 7.39	14.01	5.01
059	Rachmaninoff_Op018-03-mm122-128_JL	15.67 ± 7.72	24.93	0.19	19.45 ± 8.17	22.74	1.27
060	Rachmaninoff_Op018-03-mm150-154_JL	12.56 ± 5.67	9.88	4.19	18.22 ± 6.27	17.05	11.34
061	Rachmaninoff_Op018-03-mm167-177_JL	6.73 ± 7.53	5.06	7.78	9.87 ± 7.86	10.95	12.73
062	Rachmaninoff_Op018-03-mm431-436_JL	6.00 ± 5.40	6.36	3.26	8.45 ± 5.41	8.58	5.47
063	Rachmaninoff_Op030-01-mm003-008_ES-V1	11.09 ± 8.06	9.93	8.92	14.71 ± 7.45	14.85	13.84
064	Rachmaninoff_Op030-01-mm021-027_ES-V1	4.25 ± 6.13	3.93	8.55	8.54 ± 5.90	9.94	14.55
065	Rachmaninoff_Op030-01-mm033-038_ES-V1	3.73 ± 4.49	4.26	5.69	7.13 ± 4.77	7.58	9.00
066	Rachmaninoff_Op030-01-mm003-008_ES-V2	16.10 ± 7.73	11.76	7.88	18.54 ± 6.69	15.55	11.67
067	Rachmaninoff_Op030-01-mm021-027_ES-V2	7.29 ± 6.29	6.36	7.13	11.43 ± 5.21	11.83	12.59
068	Rachmaninoff_Op030-01-mm033-038_ES-V2	5.88 ± 4.35	5.73	3.50	8.29 ± 4.57	8.42	6.18
069	Saint_Op022-01-mm053-055_ES	9.58 ± 7.06	10.61	2.18	12.57 ± 7.79	15.11	6.69
070	Saint_Op022-01-mm061-062_ES	4.97 ± 6.63	6.32	5.08	7.87 ± 6.56	9.45	8.20
071	Schumann_Op054-01-mm019-024_ES-V1	8.86 ± 4.80	10.48	-0.20	13.44 ± 5.35	14.38	3.69
072	Schumann_Op054-01-mm034-040_ES-V1	8.65 ± 6.10	12.14	2.68	11.97 ± 6.40	14.75	6.43
073	Schumann_Op054-01-mm019-024_ES-V2	9.51 ± 5.00	10.52	0.89	14.02 ± 6.00	13.67	4.04
074	Schumann_Op054-01-mm034-040_ES-V2	10.68 ± 7.52	13.13	4.09	12.63 ± 7.62	13.31	5.74
075	Tschaikovsky_Op023-01-mm007-013_ES-V1	4.49 ± 4.24	6.10	5.12	7.60 ± 4.03	8.34	7.37
076	Tschaikovsky_Op023-01-mm018-024_ES-V1	2.02 ± 3.95	2.82	10.89	5.59 ± 4.50	6.02	14.10
077	Tschaikovsky_Op023-01-mm030-036_ES-V1	6.30 ± 5.80	6.41	7.94	9.91 ± 6.00	9.95	11.47
078	Tschaikovsky_Op023-01-mm007-013_ES-V2	6.10 ± 4.76	7.49	4.95	7.66 ± 4.22	8.88	6.34
079	Tschaikovsky_Op023-01-mm018-024_ES-V2	2.72 ± 3.37	2.92	8.94	4.38 ± 3.97	4.35	10.37
080	Tschaikovsky_Op023-01-mm030-036_ES-V2	8.51 ± 6.68	7.52	6.10	11.15 ± 6.12	10.41	8.99

Table A.1: An excerptwise overview of the source separation results, including SDR_{local} and SDR_{note} scores by the best-performing models HDMC and AudioShake.

Abbreviations

ADAM adaptive moment optimization
ALS alternating least squares
BLSTM bidirectional long short-term memory
CaC Complex-as-Channel
CPU central processing unit
CNN convolutional neural network
DBN dynamic Bayesian network
DL deep learning
DNN deep neural network
DFT discrete Fourier transform
DLNCO decaying locally adaptive chroma onset
DTW dynamic time warping
GAN generative adversarial network
GPU graphics processing unit
HPSS harmonic–percussive source separation
MSE mean squared error
MIR Music Information Retrieval
MIDI musical instrument digital interface
MMO Music Minus One
MSS music source separation
MUSHRA multiple stimulus with hidden reference and anchors
NMF nonnegative matrix factorization
NAE nonnegative autoencoder
PCD Piano Concerto Dataset
RMSprop root mean square propagation
ReLU rectified linear unit
RNN recurrent neural network
SDR signal-to-distortion ratio
SF spectral flux
SGD stochastic gradient descent
STFT short-time Fourier transform
TSM time-scale modification
TTA test-time adaptation

Bibliography

[1] Andreas Arzt and Gerhard Widmer. Real-time music tracking using multiple performances as a reference. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 357–363, Málaga, Spain, 2015.

[2] Andreas Arzt, Gerhard Widmer, and Simon Dixon. Automatic page turning for musicians via real-time machine listening. In *Proceedings of the European Conference on Artificial Intelligence (ECAI)*, pages 241–245, Amsterdam, The Netherlands, 2008. IOS Press.

[3] Andreas Arzt, Gerhard Widmer, and Simon Dixon. Adaptive distance normalization for real-time music tracking. In *Proceedings of the European Signal Processing Conference (EUSIPCO)*, pages 2689–2693, Bucharest, Romania, 2012.

[4] Mert Bay, Andreas F. Ehmann, and J. Stephen Downie. Evaluation of multiple-f0 estimation and tracking systems. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 315–320, Kobe, Japan, 2009.

[5] Juan Pablo Bello and Jeremy Pickens. A robust mid-level representation for harmonic content in music signals. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 304–311, London, UK, 2005.

[6] Juan Pablo Bello, Laurent Daudet, Samer Abdallah, Chris Duxbury, Mike Davies, and Mark B. Sandler. A tutorial on onset detection in music signals. *IEEE Transactions on Speech and Audio Processing*, 13(5): 1035–1047, 2005.

[7] Emmanouil Benetos, Simon Dixon, Zhiyao Duan, and Sebastian Ewert. Automatic music transcription: An overview. *IEEE Signal Processing Magazine*, 36(1):20–30, 2019. doi: 10.1109/MSP.2018.2869928.

[8] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 41–48, New York, NY, USA, 2009. Association for Computing Machinery. doi: 10.1145/1553374.1553380. URL `https://doi.org/10.1145/1553374.1553380`.

[9] Nancy Bertin, Roland Badeau, and Emmanuel Vincent. Enforcing harmonicity and smoothness in bayesian non-negative matrix factorization applied to polyphonic music transcription. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):538–549, 2010.

[10] Rachel M. Bittner, Justin Salamon, Mike Tierney, Matthias Mauch, Chris Cannam, and Juan Pablo Bello. MedleyDB: A multitrack dataset for annotation-intensive MIR research. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 155–160, Taipei, Taiwan, 2014. doi: 10.5281/zenodo.1417889.

[11] Rachel M. Bittner, Magdalena Fuentes, David Rubinstein, Andreas Jansson, Keunwoo Choi, and Thor Kell. mirdata: Software for reproducible usage of datasets. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 99–106, Delft, The Netherlands, 2019. URL `http://archives.ismir.net/ismir2019/paper/000009.pdf`.

[12] Sebastian Böck and Markus Schedl. Enhanced beat tracking with context-aware neural networks. In *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, pages 135–139, Paris, France, 2011.

[13] Sebastian Böck and Gerhard Widmer. Maximum filter vibrato suppression for onset detections. In *Proceedings of the International Confenference on Digital Audio Effects (DAFx)*, Maynooth, Ireland, 2013.

[14] Sebastian Böck, Filip Korzeniowski, Jan Schlüter, Florian Krebs, and Gerhard Widmer. madmom: A new Python audio and music signal processing library. In *Proceedings of the ACM International Conference on Multimedia (ACM-MM)*, pages 1174–1178, Amsterdam, The Netherlands, 2016. doi: 10.1145/2964284.2973795.

[15] Sebastian Böck, Florian Krebs, and Gerhard Widmer. Joint beat and downbeat tracking with recurrent neural networks. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 255–261, New York City, New York, USA, 2016. doi: 10.5281/zenodo.1415835.

[16] Christoph Böhm, David Ackermann, and Stefan Weinzierl. A multi-channel anechoic orchestra recording of Beethoven's Symphony no. 8 op. 93. *Journal of the Audio Engineering Society*, 68(12):977–984, 2021. doi: 10.17743/jaes.2020.0056.

[17] Juan J. Bosch, Ricard Marxer, and Emilia Gómez. Evaluation and combination of pitch estimation methods for melody extraction in symphonic classical music. *Journal of New Music Research*, 45(2):101–117, 2016. doi: 10.1080/09298215.2016.1182191.

[18] Albert S. Bregman. *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press, 1990.

[19] Andreas Bugler, Bryan Pardo, and Prem Seetharaman. A study of transfer learning in music source separation. 2020.

[20] Juan José Burred. *From Sparse Models to Timbre Learning: New Methods for Musical Source Separation*. PhD thesis, Technische Universität Berlin, Berlin, Germany, 2009.

[21] Chris Cannam, Christian Landone, and Mark B. Sandler. Sonic Visualiser: An open source application for viewing, analysing, and annotating music audio files. In *Proceedings of the International Conference on Multimedia*, pages 1467–1468, Florence, Italy, 2010.

[22] Estefanía Cano, Christian Dittmar, and Gerald Schuller. Efficient implementation of a system for solo and accompaniment separation in polyphonic music. In *Proceedings of the European Signal Processing Conference (EUSIPCO)*, pages 285–289, Bucharest, Romania, 2012.

[23] Estefanía Cano, Derry FitzGerald, and Karlheinz Brandenburg. Evaluation of quality of sound source separation algorithms: Human perception vs quantitative metrics. In *Proceedings of the European Signal Processing Conference (EUSIPCO)*, pages 1758–1762, 2016. doi: 10.1109/EUSIPCO.2016.7760550.

[24] Estefanía Cano, Derry FitzGerald, Antoine Liutkus, Mark D. Plumbley, and Fabian-Robert Stöter. Musical source separation: An introduction. *IEEE Signal Processing Magazine*, 36(1):31–40, 2019. doi: 10.1109/MSP.2018.2874719.

[25] Giorgia Cantisani, Alexey Ozerov, Slim Essid, and Gaël Richard. User-guided one-shot deep model adaptation for music source separation. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 111–115, 2021. doi: 10.1109/WASPAA52581.2021.9632717.

[26] William E. Caplin. *Classical Form: A Theory of Formal Functions for the Instrumental Music of Haydn, Mozart, and Beethoven*. Oxford University Press, USA, 1998. ISBN 9780195143997.

[27] Carlos Eduardo Cancino Chacón, Maarten Grachten, Werner Goebl, and Gerhard Widmer. Computational models of expressive music performance: A comprehensive and critical review. *Frontiers Digit. Humanit.*, 5: 25, 2018. doi: 10.3389/fdigh.2018.00025. URL `https://doi.org/10.3389/fdigh.2018.00025`.

[28] Carlos Eduardo Cancino Chacón, Silvan Peter, Patricia Hu, Emmanouil Karystinaios, Florian Henkel, Francesco Foscarin, Nimrod Varga, and Gerhard Widmer. The ACCompanion: Combining reactivity, robustness, and musical expressivity in an automatic piano accompanist. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 5779–5787, Macao, China, 2023.

[29] Ke Chen, Hao-Wen Dong, Yi Luo, Julian McAuley, Taylor Berg-Kirkpatrick, Miller Puckette, and Shlomo Dubnov. Improving choral music separation through expressive synthesized data from sampled instruments. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 726–732, Bengaluru, India, 2022.

[30] Ching-Yu Chiu, Wen-Yi Hsiao, Yin-Cheng Yeh, Yi-Hsuan Yang, and Alvin Wen-Yu Su. Mixing-specific data augmentation techniques for improved blind violin/piano source separation. In *Proceedings of the Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6, 2020.

[31] Hyeong-Seok Choi, Jang-Hyun Kim, Jaesung Huh, Adrian Kim, Jung-Woo Ha, and Kyogu Lee. Phase-aware speech enhancement with deep complex U-net. In *Proceedings of the International Conference on Learning Representations (ICLR)*, Vancouver, Canada, 2018.

[32] Woosung Choi, Minseok Kim, Jaehwa Chung, Daewon Lee, and Soonyoung Jung. Investigating U-Nets with various intermediate blocks for spectrogram-based singing voice separation. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, online, 2020.

[33] Jan Chorowski and Jacek M. Zurada. Learning understandable neural networks with nonnegative weight constraints. *IEEE Transactions on Neural Networks and Learning Systems*, 26(1):62–69, 2015. doi: 10.1109/TNNLS.2014.2310059.

[34] Alice Cohen-Hadria, Axel Roebel, and Geoffroy Peeters. Improving singing voice separation using deep U-net and wave-U-net with data augmentation. In *Proceedings of the European Signal Processing Conference (EUSIPCO)*, A Coruña, Spain, 2019.

[35] William Cole. *The Form of Music*. The Associated Board of the Royal Schools of Music (ABRSM), London, UK, 1997.

[36] Helena Cuesta, Emilia Gómez, Agustín Martorell, and Felipe Loáiciga. Analysis of intonation in unison choir singing. In *Proceedings of the International Conference of Music Perception and Cognition (ICMPC)*, pages 125–130, Graz, Austria, 2018.

[37] Roger B. Dannenberg. An on-line algorithm for real-time accompaniment. In *Proceedings of the International Computer Music Conference (ICMC)*, pages 193–198, Paris, France, 1984.

[38] Roger B. Dannenberg and Ning Hu. Polyphonic audio matching for score following and intelligent audio editors. In *Proceedings of the International Computer Music Conference (ICMC)*, pages 27–34, San Francisco, USA, 2003.

[39] Roger B. Dannenberg and Christopher Raphael. Music score alignment and computer accompaniment. *Communications of the ACM, Special Issue: Music Information Retrieval*, 49(8):38–43, 2006.

[40] Jane W. Davidson. Visual perception of performance manner in the movements of solo musicians. *Psychology of Music*, 21(2):103–113, 1993. doi: 10.1177/030573569302100201. URL https://doi.org/10.1177/030573569302100201.

[41] Matthew E. P. Davies, Paul M. Brossier, and Mark D. Plumbley. Beat tracking towards automatic musical accompaniment. In *Proceedings of the AES International Conference on Semantic Audio*, Barcelona, Spain, 2005. URL http://www.aes.org/e-lib/browse.cfm?elib=13124.

[42] Alexandre Défossez. Hybrid spectrogram and waveform source separation. In *Proceedings of the ISMIR 2021 Workshop on Music Source Separation*, Online, 2021.

[43] Alexandre Défossez, Nicolas Usunier, Léon Bottou, and Francis R. Bach. Music source separation in the waveform domain, 2019. URL http://arxiv.org/abs/1911.13254.

[44] Christian Dittmar and Meinard Müller. Reverse engineering the Amen break – score-informed separation and restoration applied to drum recordings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(9):1531–1543, 2016. doi: 10.1109/TASLP.2016.2567645.

[45] Christian Dittmar, Bernhard Lehner, Thomas Prätzlich, Meinard Müller, and Gerhard Widmer. Cross-version singing voice detection in classical opera recordings. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 618–624, Málaga, Spain, October 2015. doi: 10.5281/zenodo.1416958.

[46] Simon Dixon and Gerhard Widmer. MATCH: A music alignment tool chest. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 492–497, London, UK, 2005. doi: 10.5281/zenodo.1416952.

[47] Jonathan Driedger and Meinard Müller. Extracting singing voice from music recordings by cascading audio decomposition techniques. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 126–130, Brisbane, Australia, 2015.

[48] Jonathan Driedger, Harald Grohganz, Thomas Prätzlich, Sebastian Ewert, and Meinard Müller. Score-informed audio decomposition and applications. In *Proceedings of the ACM International Conference on Multimedia (ACM-MM)*, pages 541–544, Barcelona, Spain, 2013.

[49] Jonathan Driedger, Meinard Müller, and Sebastian Ewert. Improving time-scale modification of music signals using harmonic–percussive separation. *IEEE Signal Processing Letters*, 21(1):105–109, 2014.

[50] Zhiyao Duan and Bryan Pardo. Soundprism: An online system for score-informed source separation of music audio. *IEEE Journal of Selected Topics in Signal Processing*, 5(6):1205–1215, 2011.

[51] Ngoc Q. K. Duong, Emmanuel Vincent, and Rémi Gribonval. Under-determined reverberant audio source separation using a full-rank spatial covariance model. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 18(7):1830–1840, 2010.

[52] Valentin Emiya, Emmanuel Vincent, Niklas Harlander, and Volker Hohmann. Subjective and objective quality assessment of audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 19 (7):2046–2057, 2011.

[53] Jesse Engel, Lamtharn Hantrakul, Chenjie Gu, and Adam Roberts. DDSP: Differentiable digital signal processing. In *Proceedings of the International Conference on Learning Representations (ICLR)*, Virtual, 2020. URL `https://openreview.net/forum?id=B1x1ma4tDr`.

[54] Sebastian Ewert and Meinard Müller. Using score-informed constraints for NMF-based source separation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 129–132, Kyoto, Japan, 2012. doi: 10.1109/ICASSP.2012.6287834.

[55] Sebastian Ewert and Mark B. Sandler. Piano transcription in the studio using an extensible alternating directions framework. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(11): 1983–1997, 2016.

[56] Sebastian Ewert and Mark B. Sandler. Structured dropout for weak label and multi-instance learning and its application to score-informed source separation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 2277–2281, New Orleans, Louisiana, USA, 2017. doi: 10.1109/ICASSP.2017.7952562.

[57] Sebastian Ewert, Meinard Müller, and Peter Grosche. High resolution audio synchronization using chroma onset features. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1869–1872, Taipei, Taiwan, 2009. doi: 10.1109/ICASSP.2009.4959972.

[58] Sebastian Ewert, Bryan Pardo, Meinard Müller, and Mark Plumbley. Score-informed source separation for musical audio recordings: An overview. *IEEE Signal Processing Magazine*, 31(3):116–124, April 2014.

[59] Florian Eyben, Sebastian Böck, Björn Schuller, and Alex Graves. Universal onset detection with bidirectional long short-term memory neural networks. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 589–594, Utrecht, The Netherlands, August 2010.

[60] Giorgio Fabbro, Stefan Uhlich, Chieh-Hsin Lai, Woosung Choi, Marco Martínez-Ramírez, Weihsiang Liao, Igor Gadelha, Geraldo Ramos, Eddie Hsu, Hugo Rodrigues, Fabian-Robert Stöter, Alexandre Défossez, Yi Luo, Jianwei Yu, Dipam Chakraborty, Sharada Mohanty, Roman Solovyev, Alexander Stempkovskiy, Tatiana Habruseva, Nabarun Goswami, Tatsuya Harada, Minseok Kim, Jun Hyung Lee, Yuanliang Dong, Xinran Zhang, Jiafeng Liu, and Yuki Mitsufuji. The sound demixing challenge 2023 – music demixing track. *Transactions of the International Society for Music Information Retrieval (TISMIR)*, 7(1):63–84, 2024. doi: 10.5334/tismir.171.

[61] J. L. Flanagan and R. M. Golden. Phase vocoder. *Bell System Technical Journal*, 45:1493–1509, 1966.

[62] Jonathan T. Foote. Content-based retrieval of music and audio. In *Multimedia Storage and Archiving Systems II*, volume 3229, pages 138–147. International Society for Optics and Photonics, SPIE, 1997.

[63] Joachim Fritsch and Mark D. Plumbley. Score informed audio source separation using constrained nonnegative matrix factorization and score synthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 888–891, Vancouver, Canada, May 2013. doi: 10.1109/ICASSP.2013.6637776.

[64] Hiromasa Fujihara, Masataka Goto, Jun Ogata, and Hiroshi G. Okuno. LyricSynchronizer: Automatic synchronization system between musical audio signals and lyrics. *IEEE Journal of Selected Topics in Signal Processing*, 5(6):1252–1261, 2011.

[65] Takuya Fujishima. Realtime chord recognition of musical sound: A system using common lisp music. In *Proceedings of the International Computer Music Conference (ICMC)*, pages 464–467, Beijing, China, 1999.

[66] Dennis Gabor. Theory of communication. *Journal of the Institution of Electrical Engineers (IEE)*, 93(26): 429–457, 1946.

[67] Martin Gasser, Andreas Arzt, Thassilo Gadermaier, Maarten Grachten, and Gerhard Widmer. Classical music on the web – user interfaces and data representations. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 571–577, Málaga, Spain, 2015.

[68] Nicolas Gillis. *Nonnegative Matrix Factorization*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2020. doi: 10.1137/1.9781611976410. URL `https://epubs.siam.org/doi/abs/10.1137/1.9781611976410`.

[69] Cuthbert Morton Girdlestone. *Mozart & His Piano Concertos*. Cassell & Company Ltd., London, UK, 1948.

[70] Werner Goebl and Caroline Palmer. Synchronization of timing and motion among performing musicians. *Music Perception: An Interdisciplinary Journal*, 26(5):427–438, 2009. URL `http://www.jstor.org/stable/10.1525/mp.2009.26.5.427`.

[71] Emilia Gómez. *Tonal Description of Music Audio Signals*. PhD thesis, Universitat Pompeu Fabra, Barcelona, Spain, 2006.

[72] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka. RWC music database: Popular, classical and jazz music databases. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 287–288, Paris, France, 2002.

[73] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka. RWC music database: Music genre database and musical instrument sound database. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 229–230, Baltimore, Maryland, USA, 2003.

[74] Prachi Govalkar, Ahmed Mustafa, Nicola Pia, Judith Bauer, Metehan Yurt, Yigitcan Özer, and Christian Dittmar. A lightweight neural TTS system for high-quality German speech synthesis. In *Proceedings of the ITG Conference on Speech Communication*, pages 39–43, 2021.

[75] Matan Gover and Philippe Depalle. Score-informed source separation of choral music. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2020.

[76] Daniel W. Griffin and Jae S. Lim. Signal estimation from modified short-time Fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(2):236–243, 1984.

[77] Peter Grosche, Meinard Müller, and Sebastian Ewert. Combination of onset-features with applications to high-resolution music synchronization. In *Proceedings of the International Conference on Acoustics (NAG/DAGA)*, pages 357–360, 2009.

[78] Enric Gusó, Jordi Pons, Santiago Pascual, and Joan Serrà. On loss functions and evaluation metrics for music source separation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 306–310, 2022. doi: 10.1109/ICASSP43922.2022.9746530.

[79] Emanuël A. P. Habets. *Single- and Multi-Microphone Speech Dereverberation using Spectral Enhancement*. PhD thesis, Technische Universiteit Eindhoven, Eindhoven, The Netherlands, 2017.

[80] Curtis Hawthorne, Erich Elsen, Jialin Song, Adam Roberts, Ian Simon, Colin Raffel, Jesse H. Engel, Sageev Oore, and Douglas Eck. Onsets and frames: Dual-objective piano transcription. In *Proceedings of the International Society for Music Information Retrieval Conference, (ISMIR)*, pages 50–57, Paris, France, 2018. doi: 10.5281/zenodo.1492341.

[81] Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, Sander Dieleman, Erich Elsen, Jesse H. Engel, and Douglas Eck. Enabling factorized piano music modeling and generation with the MAESTRO dataset. In *Proceedings of the International Conference on Learning Representations (ICLR)*, New Orleans, Louisiana, USA, 2019. URL `https://openreview.net/forum?id=r1lYRjC9F7`.

[82] Romain Hennequin, Roland Badeau, and Bertrand David. NMF with time–frequency activations to model nonstationary audio events. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):744–753, 2011.

[83] Romain Hennequin, Anis Khlif, Felix Voituret, and Manuel Moussallam. Spleeter: a fast and efficient music source separation tool with pre-trained models. *Journal of Open Source Software*, 5(50):2154, 2020. doi: 10.21105/joss.02154. URL `https://doi.org/10.21105/joss.02154`. Deezer Research.

[84] Geoffry Hinton and Ruslan Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313:504–507, 2006. doi: 10.1126/science.1127647.

[85] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, November 1997. doi: 10.1162/neco.1997.9.8.1735.

[86] Yukara Ikemiya, Kazuyoshi Yoshii, and Katsutoshi Itoyama. Singing voice analysis and editing based on mutually dependent f0 estimation and source separation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 574–578, Brisbane, Australia, 2015.

[87] International Telecommunications Union. ITU-R Rec. BS.1534-3: Method for the subjective assessment of intermediate quality levels of coding systems, 2015.

[88] Andreas Jansson, Eric J. Humphrey, Nicola Montecchio, Rachel M. Bittner, Aparna Kumar, and Tillman Weyde. Singing voice separation with deep U-net convolutional networks. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 745–751, Suzhou, China, 2017.

[89] Dasaem Jeong, Taegyun Kwon, Chaelin Park, and Juhan Nam. PerformScore: Toward performance visualization with the score on the web browser. In *Demos and Late Breaking News of the International Society for Music Information Retrieval Conference (ISMIR)*, Suzhou, China, 2017.

[90] Cyril Joder and Björn W. Schuller. Score-informed leading voice separation from monaural audio. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 277–282, Porto, Portugal, 2012.

[91] Hirokazu Kameoka, Takuya Nishimoto, and Shigeki Sagayama. A multipitch analyzer based on harmonic temporal structured clustering. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3): 982–994, 2007.

[92] Thorsten Kastner and Jürgen Herre. An efficient model for estimating subjective quality of separated audio source signals. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 95–99, New Paltz, New York, USA, 2019. doi: 10.1109/WASPAA.2019.8937179.

[93] Hyemi Kim, Jiyun Park, Taegyun Kwon, Dasaem Jeong, and Juhan 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. doi: 10.1109/ICASSP49357.2023.10095061.

[94] Jong Wook Kim, Justin Salamon, Peter Li, and Juan Pablo Bello. CREPE: A convolutional representation for pitch estimation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 161–165, Calgary, Canada, 2018. doi: 10.1109/ICASSP.2018.8461329.

[95] Minseok Kim, Woosung Choi, Jaehwa Chung, Daewon Lee, and Soonyoung 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.

[96] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference for Learning Representations (ICLR)*, San Diego, California, USA, 2015. URL `http://arxiv.org/abs/1412.6980`.

[97] Qiuqiang Kong, Yin Cao, Haohe Liu, Keunwoo Choi, and Yuxuan Wang. Decoupling magnitude and phase estimation with deep resunet for music source separation. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 342–349, Online, 2021.

[98] Verena Konz, Meinard Müller, and Rainer Kleinertz. A cross-version chord labelling approach for exploring harmonic structures—a case study on Beethoven's Appassionata. *Journal of New Music Research*, 42(1): 61–77, 2013. doi: 10.1080/09298215.2012.750369.

[99] Filip Korzeniowski and Gerhard Widmer. Feature learning for chord recognition: The deep chroma extractor. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 37–43, New York City, New York, USA, 2016. doi: 10.5281/zenodo.1416314.

[100] Taegyun Kwon, Dasaem Jeong, and Juhan Nam. Polyphonic piano transcription using autoregressive multi-state note model. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Montreal, Canada, 2020.

[101] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In *Proceedings of the Neural Information Processing Systems (NIPS)*, pages 556–562, Denver, Colorado, USA, November 2000.

[102] Augustin Lefevre, Francis Bach, and Cédric Févotte. Semi-supervised NMF with time–frequency annotations for single-channel source separation. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 115–120, Porto, Portugal, 2012.

[103] Bernhard Lehner, Reinhard Sonnleitner, and Gerhard Widmer. Towards light-weight, real-time-capable singing voice detection. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 53–58, Curitiba, Brazil, 2013.

[104] Jean-Marie Lemercier, Julius Richter, Simon Welker, and Timo Gerkmann. StoRM: A diffusion-based stochastic regeneration model for speech enhancement and dereverberation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 31:2724–2737, 2023. doi: 10.1109/TASLP.2023.3294692.

[105] Alexander Lerch, Claire Arthur, Ashis Pati, and Siddharth Gururani. Music performance analysis: A survey. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 33–43, Delft, The Netherlands, 2019. doi: 10.5281/zenodo.3527735.

[106] Bochen Li, Xinzhao Liu, Karthik Dinesh, Zhiyao Duan, and Gaurav Sharma. Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications. *IEEE Transactions on Multimedia*, 21(2):522–535, 2019.

[107] Jian Liang, Ran He, and Tieniu Tan. A comprehensive survey on test-time adaptation under distribution shifts. 2023. URL `https://arxiv.org/abs/2303.15361`.

[108] Chih-Jen Lin. Projected gradient methods for nonnegative matrix factorization. *Neural Computation*, 19(10): 2756–2779, 2007. URL `https://doi.org/10.1162/neco.2007.19.10.2756`.

[109] Haohe Liu, Qiuqiang Kong, and Jiafeng Liu. CWS-PResUNet: Music source separation with channel-wise subband phase-aware ResUNet. In *Proceedings of the ISMIR 2021 Workshop on Music Source Separation*, Online, 2021.

[110] Antoine Liutkus and Roland Badeau. Generalized Wiener filtering with fractional power spectrograms. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 266–270, Brisbane, Australia, April 2015.

[111] Paula Sánchez López, Paul Callens, and Milos Cernak. A universal deep room acoustics estimator. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 356–360, 2021. doi: 10.1109/WASPAA52581.2021.9632738.

[112] Patricio López-Serrano, Christian Dittmar, Yigitcan Özer, and Meinard Müller. NMF toolbox: Music processing applications of nonnegative matrix factorization. In *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, Birmingham, UK, 2019.

[113] Yi Luo and Nima Mesgarani. Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(8):1256–1266, 2019. doi: 10.1109/TASLP.2019.2915167.

[114] Yi Luo and Jianwei Yu. Music source separation with Band-Split RNN. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1893–1901, 2023. doi: 10.1109/TASLP.2023.3271145.

[115] Ben Maman and Amit H. Bermano. Unaligned supervision for automatic music transcription in the wild. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 14918–14934, Baltimore, Maryland, USA, 2022.

[116] Ben Maman, Johannes Zeitler, Meinard Müller, and Amit H. Bermano. Performance conditioning for diffusion-based multi-instrument music synthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 5045–5049, Seoul, South Korea, 2024.

[117] A. Marco Martínez-Ramírez, , Wei-Hsiang Liao, Giorgio Fabbro, Stefan Uhlich, Chihiro Nagashima, and Yuki Mitsufuji. Automatic music mixing with deep learning and out-of-domain data. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 411–418, Bengaluru, India, 2022.

[118] Andrew McLeod, Rodrigo Schramm, Mark Steedman, and Emmanouil Benetos. Automatic transcription of polyphonic vocal music. *Applied Sciences*, 7(12), 2017.

[119] Gabriel Meseguer-Brocal and Geoffroy Peeters. Conditioned-U-net: Introducing a control mechanism in the U-net for multiple source separations. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 159–165, Delft, The Netherlands, 2019.

[120] Stylianos Ioannis Mimilakis, Konstantinos Drossos, Estefanía Cano, and Gerald Schuller. Examining the mapping functions of denoising autoencoders in singing voice separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:266–278, 2019. doi: 10.1109/TASLP.2019.2952013.

[121] Marius Miron, Jordi Janer, and Emilia Gómez. Monaural score-informed source separation for classical music using convolutional neural networks. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 55–62, Suzhou, China, 2017.

[122] Yuki Mitsufuji, Giorgio Fabbro, Stefan Uhlich, Fabian-Robert Stöter, Alexandre Défossez, Minseok Kim, Woosung Choi, Chin-Yun Yu, and Kin-Wai Cheuk. Music demixing challenge 2021. *Frontiers in Signal Processing*, 1, 2022. doi: 10.3389/frsip.2021.808395.

[123] Meinard Müller. *Fundamentals of Music Processing – Audio, Analysis, Algorithms, Applications*. Springer Verlag, 2015. ISBN 978-3-319-21944-8. doi: 10.1007/978-3-319-21945-5.

[124] Meinard Müller. *Fundamentals of Music Processing – Using Python and Jupyter Notebooks*. Springer Verlag, 2nd edition, 2021. ISBN 978-3-030-69807-2. doi: 10.1007/978-3-030-69808-9.

[125] Meinard Müller, Thomas Prätzlich, and Christian Dittmar. Freischütz Digital – When computer science meets musicology. In Kristina Richts and Peter Stadler, editors, *Festschrift für Joachim Veit zum 60. Geburtstag*, pages 551–573, München, Germany, 2016. Allitera.

[126] Meinard Müller, Yigitcan Özer, Michael Krause, Thomas Prätzlich, and Jonathan Driedger. Sync Toolbox: A Python package for efficient, robust, and accurate music synchronization. *Journal of Open Source Software (JOSS)*, 6(64):3434:1–4, 2021. doi: 10.21105/joss.03434.

[127] Juhan Nam, Jiquan Ngiam, Honglak Lee, and Malcolm Slaney. A classification-based polyphonic piano transcription approach using learned feature representations. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 175–180, Miami, Florida, 2011.

[128] Bernhard Niedermayer and Gerhard Widmer. A multi-pass algorithm for accurate audio-to-score alignment. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 417–422, Utrecht, The Netherlands, 2010.

[129] Alan V. Oppenheim, Alan S. Willsky, and Hamid Nawab. *Signals and Systems*. Prentice Hall, 1996.

[130] Yigitcan Özer and Müller. A computational approach for creating orchestral accompaniments from piano concerto recordings. In *Proceedings of the Deutsche Jahrestagung für Akustik (DAGA)*, pages 1370–1373, Hamburg, Germany, 2023.

[131] Yigitcan Özer and Meinard Müller. Source separation of piano concertos with test-time adaptation. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 493–500, Bengaluru, India, 2022.

[132] Yigitcan Özer and Meinard Müller. Source separation of piano concertos using musically-motivated augmentation techniques. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 32:1214–1225, 2024. doi: 10.1109/TASLP.2024.3356980.

[133] Yigitcan Özer, Michael Krause, and Meinard Müller. Using the sync toolbox for an experiment on high-resolution music alignment. In *Demos and Late Breaking News of the International Society for Music Information Retrieval Conference (ISMIR)*, Online, 2021. URL `https://archives.ismir.net/ismir2021/latebreaking/000025.pdf`.

[134] Yigitcan Özer, Jonathan Hansen, Tim Zunner, and Meinard Müller. Investigating nonnegative autoencoders for efficient audio decomposition. In *Proceedings of the European Signal Processing Conference (EUSIPCO)*, pages 254–258, Belgrade, Serbia, 2022. doi: 10.23919/EUSIPCO55093.2022.9909787.

[135] Yigitcan Özer, Matěj Ištvánek, Vlora Arifi-Müller, and Meinard Müller. Using activation functions for improving measure-level audio synchronization. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 749–756, Bengaluru, India, 2022.

[136] Yigitcan Özer, Simon Schwär, Vlora Arifi-Müller, Jeremy Lawrence, Emre Sen, and Meinard Müller. Piano Concerto Dataset (PCD): A multitrack dataset of piano concertos. *Transactions of the International Society for Music Information Retrieval (TISMIR)*, 6(1):75–88, 2023. doi: 10.5334/tismir.160.

[137] Yigitcan Özer, Hans-Ulrich Berendes, Vlora Arifi-Müller, Fabian-Robert Stöter, and Meinard Müller. Notewise evaluation for music source separation: A case study for separated piano tracks. *Submitted for publication*, 2024.

[138] Yigitcan Özer, Leo Brütting, Simon Schwär, and Meinard Müller. libsoni: A Python toolbox for sonifying music annotations and feature representations. *Journal of Open Source Software (JOSS)*, 9(96):1–6, 2024. doi: 10.21105/joss.06524.

[139] Santiago Pascual, Antonio Bonafonte, and Joan Serrà. SEGAN: Speech enhancement generative adversarial network. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 3642–3646, 2017. doi: 10.21437/Interspeech.2017-1428.

[140] Jouni Paulus, Meinard Müller, and Anssi Klapuri. Audio-based music structure analysis. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 625–636, Utrecht, The Netherlands, 2010.

[141] Geoffroy Peeters. Deriving musical structure from signal analysis for music audio summary generation: "sequence" and "state" approach. In *Computer Music Modeling and Retrieval*, volume 2771 of *Lecture Notes in Computer Science*, pages 143–166. Springer Berlin / Heidelberg, 2004.

[142] Igor Pereira, Felipe Araújo, Filip Korzeniowski, and Richard Vogl. MoisesDB: A dataset for source separation beyond 4-stems. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 619–626, Milan, Italy, 2023.

[143] Darius Petermann, Pritish Chandna, Helena Cuesta, Jordi Bonada, and Emilia Gómez. Deep learning based source separation applied to choir ensembles. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 733–739, Montreal, Canada, 2020.

[144] Alastair Porter. Evaluating musical fingerprinting systems. Master's thesis, McGill University, Montreal, Canada, 2012.

[145] Thomas Prätzlich and Meinard Müller. Triple-based analysis of music alignments without the need of ground-truth annotations. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 266–270, Shanghai, China, March 2016.

[146] Thomas Prätzlich, Jonathan Driedger, and Meinard Müller. Memory-restricted multiscale dynamic time warping. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 569–573, Shanghai, China, March 2016. doi: 10.1109/ICASSP.2016.7471739.

[147] John G. Proakis and Dimitris G. Manolakis. *Digital Signal Processing*. Prentice Hall, 1996.

[148] Zdeněk Pruša and Peter L. Søndergaard. Real-time spectrogram inversion using phase gradient heap integration. In *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, pages 17–21, Brno, Czech Republic, September 2016.

[149] Stanislaw Andrzej Raczynski, Nobutaka Ono, and Shigeki Sagayama. Multipitch analysis with harmonic nonnegative matrix approximation. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 381–386, Vienna, Austria, September 2007.

[150] Zafar Rafii and Bryan Pardo. Repeating pattern extraction technique (REPET): A simple method for music/voice separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(1):71–82, 2013.

[151] Zafar Rafii, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, and Rachel Bittner. The MUSDB18 corpus for music separation, 2017. URL `https://doi.org/10.5281/zenodo.1117372`.

[152] Zafar Rafii, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, Derry FitzGerald, and Bryan Pardo. An overview of lead and accompaniment separation in music. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(8):1307–1335, 2018. doi: 10.1109/TASLP.2018.2825440.

[153] Christopher Raphael. Music plus one and machine learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 21–28, Haifa, Israel, 2010.

[154] Christopher Raphael and Yupeng Gu. Orchestral accompaniment for a reproducing piano. In *Proceedings of the International Computer Music Conference (ICMC)*, pages 501–504, Montreal, Quebec, Canada, 2009.

[155] Lenny Renault, Rémi Mignot, and Axel Roebel. Differentiable piano model for midi-to-audio performance synthesis. In *Proceedings of the 25th International Conference on Digital Audio Effects*, pages 232–239, Vienna, Austria, 2022.

[156] Francisco J. Rodriguez-Serrano, Zhiyao Duan, Pedro Vera-Candeas, Bryan Pardo, and Julio J. Carabias-Orti. Online score-informed source separation with adaptive instrument models. *Journal of New Music Research*, 44(2):83–96, 2015. doi: 10.1080/09298215.2014.989174.

[157] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention MICCAI*, pages 234–241, Munich, Germany, 2015. doi: 10.1007/978-3-319-24574-4_28.

[158] Sebastian Rosenzweig, Helena Cuesta, Christof Weiß, Frank Scherbaum, Emilia Gómez, and Meinard Müller. Dagstuhl ChoirSet: A multitrack dataset for MIR research on choral singing. *Transactions of the International Society for Music Information Retrieval (TISMIR)*, 3(1):98–110, 2020. doi: 10.5334/tismir.48.

[159] Simon Rouard, Francisco Massa, and Alexandre 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. doi: 10.1109/ICASSP49357.2023.10096956.

[160] Daniel Röwenstrunk, Thomas Prätzlich, Thomas Betzwieser, Meinard Müller, Gerd Szwillus, and Joachim Veit. Das Gesamtkunstwerk Oper aus Datensicht – Aspekte des Umgangs mit einer heterogenen Datenlage im BMBF-Projekt "Freischütz Digital". *Datenbank-Spektrum*, 15(1):65–72, 2015. doi: 10.1007/s13222-015-0179-0.

[161] Stan Salvador and Philip Chan. FastDTW: Toward accurate dynamic time warping in linear time and space. In *Proceedings of the KDD Workshop on Mining Temporal and Sequential Data*, 2004.

[162] Craig Stuart Sapp. Comparative analysis of multiple musical performances. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 497–500, Vienna, Austria, 2007. doi: 10.5281/zenodo.1417693.

[163] Saurjya Sarkar, Emmanouil Benetos, and Mark Sandler. EnsembleSet: A new high quality dataset for chamber ensemble separation. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 625–632, Bengaluru, India, 2022.

[164] Saurjya Sarkar, Louise Thorpe, Emmanouil Benetos, and Mark 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)*, pages 1–5, 2023. doi: 10.1109/WASPAA58266.2023.10248118.

[165] Ryosuke Sawata, Stefan Uhlich, Shusuke Takahashi, and Yuki Mitsufuji. All for one and one for all: Improving music separation by bridging networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 51–55, Toronto, ON, Canada, 2021. doi: 10.1109/ICASSP39728.2021.9414044.

[166] Markus Schedl, David Hauger, Marko Tkalčič, Mark Melenhorst, and Cynthia C. S. Liem. A dataset of multimedia material about classical music: PHENICX-SMM. In *Proceedings of International Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 1–4, 2016. doi: 10.1109/CBMI.2016.7500240.

[167] Arnold Schering. *Geschichte des Instrumentalkonzerts*. Breitkopf & Härtel, Leipzig, Germany, 1905.

[168] Jan Schlüter and Sebastian Böck. Improved musical onset detection with convolutional neural networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 6979–6983, Florence, Italy, May 2014. doi: 10.1109/ICASSP.2014.6854953.

[169] Rodrigo Schramm and Emmanouil Benetos. Automatic transcription of a cappella recordings from multiple singers. In *Proceedings of the AES International Conference on Semantic Audio*, pages 108–115, Erlangen, Germany, 2017.

[170] Kilian Schulze-Forster, Clement S. J. Doire, Gaël Richard, and Roland Badeau. Weakly informed audio source separation. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 273–277, New Paltz, USA, 2019. doi: 10.1109/WASPAA.2019.8937266.

[171] Kilian Schulze-Forster, Gaël Richard, Liam Kelley, Clement S. J. Doire, and Roland Badeau. Unsupervised music source separation using differentiable parametric source models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 31:1276–1289, 2023. doi: 10.1109/TASLP.2023.3252272.

[172] Diemo Schwarz, Nicola Orio, and Norbert Schnell. Robust polyphonic midi score following with hidden Markov models. In *International Computer Music Conference (ICMC)*, Miami, Florida, USA, 2004.

[173] Xavier Serra and Julius Smith III. Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition. *Computer Music Journal*, 14(4):12–24, 1990.

[174] Alexander Sheh and Daniel P. W. Ellis. Chord segmentation and recognition using EM-trained hidden Markov models. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 185–191, Baltimore, Maryland, USA, 2003.

[175] Paris Smaragdis and Judith 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)*, pages 177–180, 2003.

[176] Paris Smaragdis and Shrikant Venkataramani. A neural network alternative to non-negative audio models. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 86–90, New Orleans, Louisiana, USA, 2017. doi: 10.1109/ICASSP.2017.7952123.

[177] Paris Smaragdis, Cédric Févotte, Gautham J. Mysore, Nasser Mohammadiha, and Matthew D. Hoffman. Static and dynamic source separation using nonnegative factorizations: A unified view. *IEEE Signal Processing Magazine*, 31(3):66–75, 2014. doi: 10.1109/MSP.2013.2297715.

[178] Xuchen Song, Qiuqiang Kong, Xingjian Du, and Yuxuan Wang. Catnet: music source separation system with mix-audio augmentation. *CoRR*, abs/2102.09966, 2021.

[179] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56): 1929–1958, June 2014. URL `https://jmlr.org/papers/v15/srivastava14a.html`.

[180] Christian J. Steinmetz, Jordi Pons, Santiago Pascual, and Joan Serrà. Automatic multitrack mixing with a differentiable mixing console of neural audio effects. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 71–75, Barcelona, Spain, 2020. doi: 10.1109/ICASSP39728.2021.9414364.

[181] Daniel Stoller, Sebastian Ewert, and Simon 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)*, pages 334–340, Paris, France, 2018.

[182] Fabian-Robert Stöter, Stefan Bayer, and Bernd Edler. Unison source separation. In *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, pages 235–241, Erlangen, Germany, 2014.

[183] Fabian-Robert Stöter, Antoine Liutkus, Roland Badeau, Bernd Edler, and Paul Magron. Common fate model for unison source separation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 126–130, Shanghai, China, 2016.

[184] Fabian-Robert Stöter, Antoine Liutkus, and Nobutaka Ito. The 2018 Signal Separation Evaluation Campaign. In *Proceedings of the International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, volume 10891 of *Lecture Notes in Computer Science*, pages 293–305. Springer, 2018. doi: 10.1007/978-3-319-93764-9_28.

[185] Fabian-Robert Stöter, Stefan Uhlich, Antoine Liutkus, and Yuki Mitsufuji. Open-Unmix – A reference implementation for music source separation. *Journal of Open Source Software*, 4(41), 2019. doi: 10.21105/joss.01667. URL `https://doi.org/10.21105/joss.01667`.

[186] Sebastian Strahl, Yigitcan Özer, Hans-Ulrich Berendes, and Meinard Müller. Hearing your way through music recordings: A text alignment and synthesis approach. *Submitted for publication*, 2024.

[187] Martin Strauss and Bernd Edler. A flow-based neural network for time domain speech enhancement. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 5754–5758, Toronto, ON, Canada, 2021. doi: 10.1109/ICASSP39728.2021.9413999.

[188] Yu Sun, Xiaolong Wang, Liu Zhang, John Miller, Moritz Hardt, and Alexei A. Efros. Test-time training with self-supervision for generalization under distribution shifts. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020.

[189] Naoya Takahashi and Yuki Mitsufuji. Multi-scale multi-band densenets for audio source separation. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 21–25, 2017. doi: 10.1109/WASPAA.2017.8169987.

[190] Naoya Takahashi, Nabarun Goswami, and Yuki Mitsufuji. MMDenseLSTM: An efficient combination of convolutional and recurrent neural networks for audio source separation. In *International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 106–110, Tokyo, Japan, 2018. doi: 10.1109/IWAENC.2018.8521383.

[191] Nazif Can Tamer, Yigitcan Özer, Meinard Müller, and Xavier Serra. High-resolution violin transcription using weak labels. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 223–230, Milan, Italy, 2023.

[192] Nazif Can Tamer, Yigitcan Özer, Meinard Müller, and Xavier Serra. TAPE: An end-to-end timbre-aware pitch estimator. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Rhodes Island, Greece, 2023. doi: 10.1109/ICASSP49357.2023.10096762.

[193] Tijmen Tieleman and Geoffrey Hinton. RmsProp: Divide the gradient by a running average of its recent magnitude, October 2012.

[194] Matteo Torcoli and Emanuël A. P. Habets. Better together: Dialogue separation and voice activity detection for audio personalization in TV. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Rhodes Island, Greece, 2023. doi: 10.1109/ICASSP49357.2023.10095153.

[195] Matteo Torcoli, Thorsten Kastner, and Jürgen Herre. Objective measures of perceptual audio quality reviewed: An evaluation of their application domain dependence. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1530–1541, 2021. doi: 10.1109/TASLP.2021.3069302.

[196] Chiheb Trabelsi, Olexa Bilaniuk, Ying Zhang, Dmitriy Serdyuk, Sandeep Subramanian, João Felipe Santos, Soroush Mehri, Negar Rostamzadeh, Yoshua Bengio, and Christopher J Pal. Deep complex networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, Vancouver, Canada, 2018.

[197] Christopher Tralie and Elizabeth Dempsey. Exact, parallelizable dynamic time warping alignment with linear memory. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 462–469, Montréal, Canada, 2020.

[198] T. J. Tsai, Kavi Dey, Yigitcan Özer, and Meinard Müller. Customizing piano concerto accompaniments using hybrid dense-sparse dynamic time warping. *Submitted for publication*, 2024.

[199] TJ Tsai. Segmental DTW: A parallelizable alternative to dynamic time warping. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 106–110, Toronto, ON, Canada, 2021. doi: 10.1109/ICASSP39728.2021.9413827.

[200] TJ Tsai, Steven K. Tjoa, and Meinard Müller. Make your own accompaniment: Adapting full-mix recordings to match solo-only user recordings. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 79–86, Suzhou, China, 2017.

[201] George Tzanetakis and Perry Cook. Multifeature audio segmentation for browsing and annotation. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 103–106, New Paltz, New York, USA, 1999.

[202] Stefan Uhlich, Marcello Porcu, Franck Giron, Michael Enenkl, Thomas Kemp, Naoya Takahashi, and Yuki 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)*, pages 261–265, New Orleans, Louisiana, USA, March 2017.

[203] Barry Vercoe. The synthetic performer in the context of live performance. In *Proceedings of the International Computer Music Conference (ICMC)*, pages 199–200, 1984.

[204] Werner Verhelst and Marc Roelands. 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.

[205] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1462–1469, 2006.

[206] Yu Wang, Juan Pablo Bello, Daniel Stoller, and Rachel Bittner. Few-shot musical source separation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 121–125, Singapore, Singapore, 2022. doi: 10.1109/ICASSP43922.2022.9747536.

[207] Zhong-Qiu Wang, Samuele Cornell, Shukjae Choi, Younglo Lee, Byeong-Yeol Kim, and Shinji Watanabe. TF-GRIDNET: Making time-frequency domain models great again for monaural speaker separation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Rhodes Island, Greece, 2023. doi: 10.1109/ICASSP49357.2023.10094992.

[208] Christof Weiß, Vlora Arifi-Müller, Thomas Prätzlich, Rainer Kleinertz, and Meinard Müller. Analyzing measure annotations for Western classical music recordings. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 517–523, New York, USA, 2016. doi: 10.5281/zenodo.1417449.

[209] Christof Weiß, Hendrik Schreiber, and Meinard Müller. Local key estimation in music recordings: A case study across songs, versions, and annotators. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2919–2932, 2020. doi: 10.1109/TASLP.2020.3030485.

[210] Nils Werner, Stefan Balke, Fabian-Robert Stöter, Meinard Müller, and Bernd Edler. trackswitch.js: A versatile web-based audio player for presenting scientific results. In *Proceedings of the Web Audio Conference (WAC)*, London, UK, 2017.

[211] Jun Wu, Emmanuel Vincent, Stanislaw Andrzej Raczynski, Takuya Nishimoto, Nobutaka Ono, and Shigeki Sagayama. Multipitch estimation by joint modeling of harmonic and transient sounds. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 25–28, Prague, Czech Republic, 2011.

[212] Yu-Te Wu, Berlin Chen, and Li Su. Multi-instrument automatic music transcription with self-attention-based instance segmentation. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 28:2796–2809, 2020. doi: 10.1109/TASLP.2020.3030482.

[213] Siyuan Yuan, Zhepei Wang, Umut Isik, Ritwik Giri, Jean-Marc Valin, Michael M. Goodwin, and Arvindh Krishnaswamy. Improved singing voice separation with chromagram-based pitch-aware remixing. In

Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages 111–115, Virtual and Singapore, 2022. doi: 10.1109/ICASSP43922.2022.9747612.

[214] Huan Zhang, Jingjing Tang, Syed Rifat Mahmud Rafee, Simon Dixon, and György Fazekas. ATEPP: A dataset of automatically transcribed expressive piano performance. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 446–453, Bengaluru, India, 2022. doi: 10.5281/zenodo.7342764.

[215] Tim Zunner. Neural networks with nonnegativity constraints for decomposing music recordings. Master thesis, Friedrich-Alexander-University of Erlangen-Nuremberg, Erlangen, Germany, 2021.