Source Separation of Piano Concertos Using Musically Motivated Augmentation Techniques

Yigitcan Özer and Meinard Müller, *Fellow, IEEE*

***Abstract*—In this work, 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. Moreover, in the case of piano concertos, the lack of multi-track data for training constitutes another challenge in view of data-driven source separation approaches. As a basis for our work, we adapt existing deep learning (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 a recently published, open-source dataset of multi-track piano concerto recordings. Our main findings demonstrate that the best source separation performance is achieved by a hybrid model when combining all augmentation techniques.**

***Index Terms*—Audio source separation, piano concerto, orchestral music, music processing, music information retrieval.**

I. Introduction

THE piano concerto is a genre of great importance in Western classical music. This genre is generally composed for pianists, accompanied by an ensemble or orchestra, to demonstrate their virtuosity. A piano concerto typically consists of multiple movements, with the piano playing the primary role and the orchestra taking over the accompaniment [1]. Piano concertos have been written by numerous composers spanning various periods, starting from the Baroque era and persisting until today. This enduring and widely embraced form of classical music continues to fascinate audiences worldwide.

Although practicing and playing piano concertos is a main activity of pianists in their career, only first-class pianists get the opportunity to perform alongside an orchestra. Motivated by the need for orchestral accompaniments of amateur or semi-professional pianists, we consider the novel task of separating piano concertos building on our previous work [2], which we substantially extend in this paper, particularly through the adaptation of four deep learning (DL) models. For an illustration of the task, see Fig. 1.

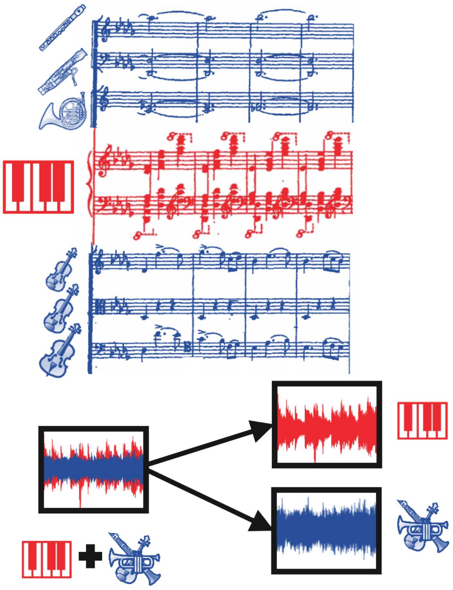

Fig. 1. Excerpt from Tchaikovsky's Piano Concerto No. 1 in B Flat Minor, Op. 23, 1st Movement. Our goal is to decompose piano concertos into the piano (red) and orchestral (blue) tracks using data-driven music source separation (MSS) techniques.

Music source separation (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 [3]. In this context, deep neural networks (DNNs) have led to substantial improvements in separating and isolating musical sources, see, e.g., [4], [5], [6], [7], [8], [9], [10], [11], [12].

Manuscript received 21 July 2023; revised 23 November 2023 and 9 January 2024; accepted 14 January 2024. Date of publication 24 January 2024; date of current version 3 February 2024. This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Grants 328416299 and DFG MU 2686/10-2. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Juhan Nam. *(Corresponding author: Yigitcan Özer.)*

The authors are with International Audio Laboratories Erlangen, Friedrich-Alexander-Universitat Erlangen-Nurnberg, 91058 Erlangen, Germany (e-mail: yigitcan.oezer@audiolabs-erlangen.de; meinard.mueller@audiolabs-erlangen.de).

Digital Object Identifier 10.1109/TASLP.2024.3356980

Supervised deep learning models addressing the MSS task typically require a large dataset that consists of multi-track recordings containing the individual stems of the various musical sources. Because of the availability of such multi-track recordings for popular music, most MSS models focus on the separation of at least four stems including vocals, drums, base, and other [13], [14], [15]. Furthermore, there has been growing interest in the separation of individual sound sources within classical music recordings [16], [17], [18], [19], 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 multi-track recordings available for classical music [20], [21], [22], [23], [24], [25], [26]. In case multi-track recordings are unavailable, random mixing can be used to artificially generate and augment training data [10], [27]. Following this strategy, we used artificial training material in a previous work [2] 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 [5]. As a main contribution of this paper, we extend our previous work 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 [2], instead of using artificially generated test data, we evaluate our models using the Piano Concerto Dataset (PCD) [28], 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 Signal to Distortion Ratio (SDR) [29] and also the 2f-score [30], which is a perceptually motivated quality measure yielding better results in source separation tasks [31]. Finally, we conduct listening tests based on the Multiple Stimulus with Hidden Reference and Anchors (MUSHRA) framework [32] 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.[1]

The remainder of our article is organized as follows. Section II discusses the relevant work on source separation. We then revisit in Section III the architecture and characteristics of four different networks, which we adapt for our application scenario. In Section IV, we introduce our musically motivated data augmentation approaches. Then, in Section V, we describe the experimental settings and our design choices and report on the quantitative empirical results, including a subjective evaluation. Finally, in Section VI, we conclude with prospects on future work.

II. Related Work

The models used in this paper 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 [4], [5], [6]. Binary masking, soft masking, or multichannel Wiener filtering are then typically used to reconstruct the separated audio signals [33]. Besides using the magnitude spectrogram, recent approaches also use the real and imaginary parts or include the phase of the complex-valued spectrogram [34], [35], [36], [37]. For example, Choi et al. [38] 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 Channels (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 [39] or Phase Gradient Heap Integration (PGHI) [40].

A second class of MSS models directly operates in the waveform domain [7], [8]. 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 [41]. 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 [9], [10], [11], [42]. 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.

III. Adaptation of Source Separation Models

In this section, we first introduce the basic notation in Section III-A, which we use throughout this article. Then, we revisit the architecture and characteristics of four different models, which we adapt for our source separation task of piano concertos (see also Fig. 2). In particular, we first explore the spectrogram-based models Open-Unmix (`UMX`), and Spleeter (`SPL`) in Sections III-B and III-C, respectively. Then, we investigate the waveform-based model Demucs (`DMC`) in Section III-D. Finally, we describe in Section III-E the hybrid model HDemucs (`HDMC`), which operates both in spectrogram and waveform domains.

It is important to note that all the separation approaches are applied to stereo input waveforms or spectrograms, and the resulting output signals also comprise two channels. However,

[1] [Online]. Available: https://github.com/yiitozer/pc-separation

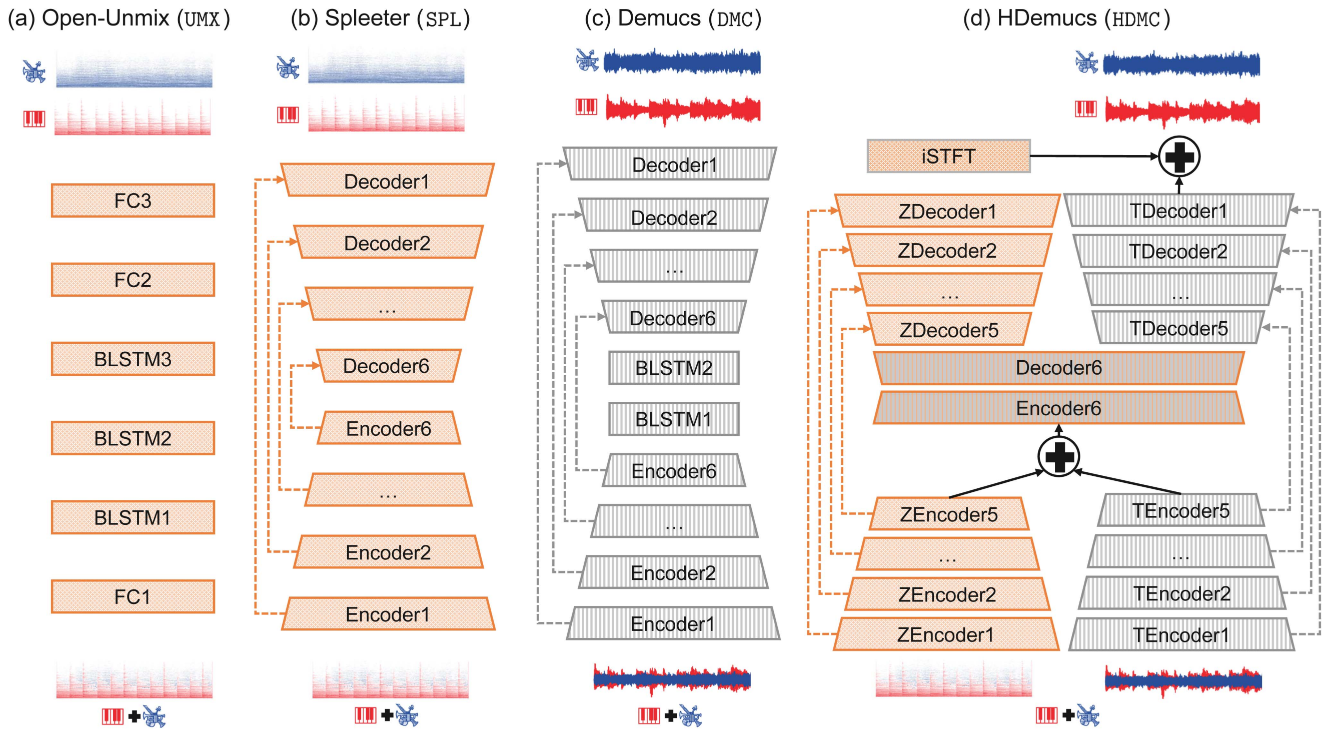

Fig. 2. Overview of source separate models, which we adapt for separating piano concertos. Note that while only the monaural case is illustrated, all models are designed to work with stereo signals. (a) Spectrogram-based Open-Unmix (UMX) [4]. (b) Spectrogram-based Spleeter (SPL) [5]. (c) Waveform-based Demucs (DMC) [8]. (d) Hybrid model HDemucs (HDMC) [9]. Spectral branches are shown in orange and temporal in gray. Dashed lines denote the skip connections of the U-Net-based network architectures.

for the sake of simplicity and clarity, we chose to formulate the signal model for the monaural case.

A. Basic Notation

Given a real-valued, discrete, time-domain signal $x : \mathbb{Z} \to \mathbb{R}$, we employ the Short-Time Fourier Transform (STFT) as follows: At time frame $m \in [0 : M-1]$ and spectral bin $k \in [0 : K]$, we compute the complex-valued STFT coefficient $\mathcal{X}(m,k)$ using a suitable window function $w : [0 : N-1] \to \mathbb{R}$ of even length $N \in \mathbb{N}$ as

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

where $H \in \mathbb{N}$ denotes the hop size. The number of frequency bins[2] is the frequency index corresponding to the Nyquist frequency $K = N/2$. The number of spectral frames $M \in \mathbb{N}$ is determined by the number of discrete signal samples. From the complex-valued spectrogram $\mathcal{X} \in \mathbb{C}^{M\times K}$, we derive the magnitude spectrogram $\mathcal{Y} \in \mathbb{R}_{\geq 0}^{M\times K}$ by $\mathcal{Y}(m,k) = |\mathcal{X}(m,k)|$.

In our source separation approaches, under the assumption of an instantaneous linear mixing model [43], 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. In our setting, we have $S = \{\mathrm{p}, \mathrm{o}\}$, where p denotes the piano and o the orchestra source.

B. Open-Unmix (UMX)

Given the magnitude spectrogram $\mathcal{Y}_\mathrm{m}$ of an input mixture, UMX [4] 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:

$$\hat{\mathcal{Y}}_s = \mathcal{Y}_\mathrm{m} \odot M_s, \quad (2)$$

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 [33].

The core architecture of UMX is a three-layer bidirectional long short-term memory (BLSTM) [44] as described in [45] (see Fig. 2(a)). Throughout our experiments, we remain consistent with the original implementation and employ the MSE loss:

$$\mathcal{L}_\mathrm{MSE} = ||\mathcal{Y}_s - \hat{\mathcal{Y}}_s||_2^2, \quad (3)$$

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 [46].

As indicated in Table I, 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 [47].

[2] Due to the real-valued nature of signal x, its spectrum exhibits Hermitian symmetry. Therefore, we eliminate the frequency bins in the upper half of the frequency spectrum.

TABLE I
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

C. Spleeter (SPL)

Being a spectrogram-based model, SPL [5] 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 [48], which is widely-used model in MIR research to address the MSS task [7], [8], [11], [38], [49], [50]. 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 [6], which consists of 12-layer convolutional networks—six layers for encoder and six layers for the decoder (see Fig. 2(b)). 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 I, 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 (2). 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 $\mathcal{Y}_s$:

$$\mathcal{L}_1^{\text{spec}} = \frac{1}{|S|} \sum_{s \in S} ||\mathcal{Y}_s - \hat{\mathcal{Y}}_s||_1. \tag{4}$$

For further details about the network architecture, we refer to [5], [6].

D. Demucs (DMC)

DMC [8] 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 Fig. 2(c)). 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 I, DMC has the most parameters among the four models.

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

$$\mathcal{L}_1^{\text{time}} = \frac{1}{|S|} \sum_{s \in S} ||x_s - \hat{x}_s||_1, \tag{5}$$

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 [8].

E. Hybrid Demucs (HDMC)

HDMC [9] is an extension of DMC with an additional spectral branch. As illustrated in Fig. 2(d), 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 [9].

The spectral input (Fig. 2(d), left) is the complex-valued STFT $\mathcal{X}_{\mathrm{m}}$ of an input mixture x_{m}. Following the *CaC* approach by Choi et al. [38], the real part $\mathrm{Re}(\mathcal{X}_{\mathrm{m}})$ and the imaginary part $\mathrm{Im}(\mathcal{X}_{\mathrm{m}})$ 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 5th encoder layer (ZEncoder5) of the spectral branch of the network.

The temporal branch (Fig. 2(d), right) receives the raw waveform x_{m}, similar to DMC. The output of the 5th 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 6th encoder layer. The output of the 6th encoder layer serves as an input both for spectral and temporal decoders. To account for the long-range temporal context, the 5th and 6th 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 $\hat{x}_s^{\mathrm{Z}}$. Furthermore, the temporal branch directly outputs a waveform $\hat{x}_s^{\mathrm{T}}$. The outputs from both branches are summed to compute the estimated waveform of the target source:

$$\hat{x}_s = \hat{x}_s^{\mathrm{Z}} + \hat{x}_s^{\mathrm{T}}. \tag{6}$$

Similar to DMC, we use the ℓ^1-norm as the loss function of HDMC, as in (5). For further details about the network architecture, we refer to [9].

IV. 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 Fig. 3. In the following, we delve deeper into our proposed methods, inspired by the harmonic, rhythmic, and structural elements found in piano concertos.

A. Random Mixing

Supervised deep learning models designed for MSS typically rely on large datasets containing recordings of isolated stems.

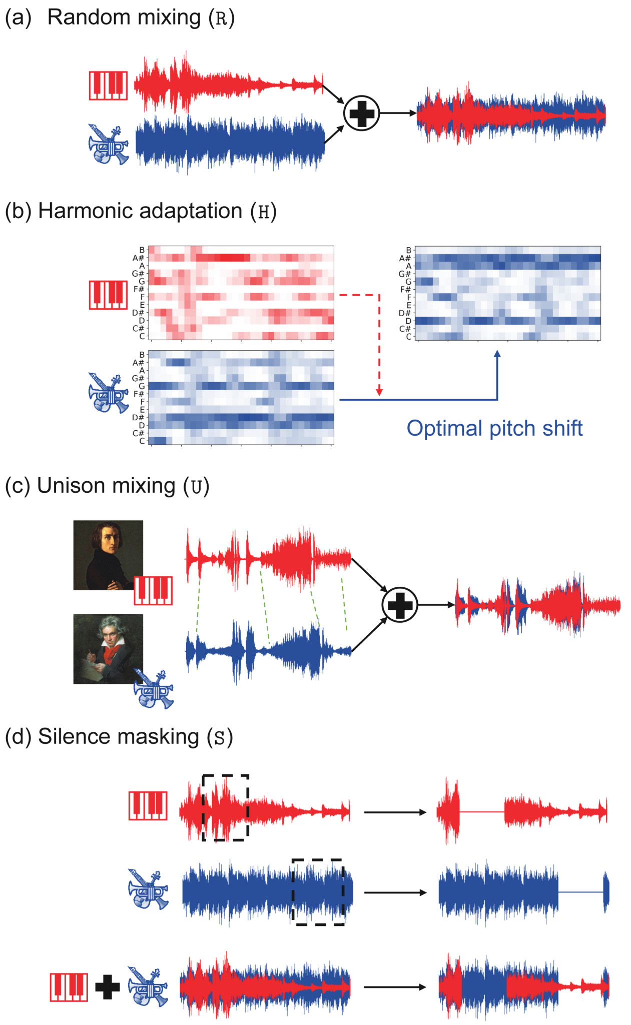

Fig. 3. Musically-motivated data augmentation strategies. (a) Random mixing recordings from the solo piano repertoire (e.g., piano sonatas) and orchestral recordings without piano (e.g., symphonies). (b) Harmonic adaption of the orchestral recordings to the piano tracks using optimal pitch shift. (c) 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.

Since such multi-track recordings are not available in the case of piano concertos, we create a dataset as in our previous work [2] through random mixes of piano-only recordings (e.g., piano sonatas) and recordings of orchestral music without piano (e.g., symphonies), see Fig. 3(a) 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).[3] As for the piano recordings, we first use MAESTRO [52], 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 [53] 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 [22], which consists of clean multi-track recordings of four orchestral excerpts by different composers. Second, we consider Bach10 [54], which comprises multi-track 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 [55], which contains 64 audio excerpts from orchestral works interpreted by symphonic orchestras, mostly from the romantic period, as well as classical and 20th century pieces. Fourth, we select a subset of 19 classical music recordings without piano selected from the Real World Computing (RWC) dataset [56]. 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.

B. 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 intricate interaction between 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 [43]. In this context, to

[3] [Online]. Available: https://imslp.org/

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 [27], [57], [58], and apply pitch shifting to create more harmonically plausible mixtures [9]. 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. Fig. 3(b) 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.

C. 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 [59]. 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 [43]. 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 PCD ID 000, 005, 071, and 073 in the test dataset [28][4]).

To better separate unison mixtures of orchestral instruments, Stöter et al. [60] 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 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 Fig. 3(c)).

For the alignment of orchestra and piano versions, we use Dynamic Time Warping (DTW), which is a well-known technique for music synchronization [61], [62]. Conventional methods typically use chroma features as the input representation to the alignment algorithm [63], [64]. 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 [65] 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 in [66]. This method incorporates beat, downbeat, and onset activation functions computed using the open-source *madmom* library [67][5], 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 [68], which provides an efficient implementation of DTW [69].

To generate orchestral tracks, which are synchronous with the piano recordings, we then employ Time-Scale Modification (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. [70], which combines harmonic–percussive source separation (HPSS) and classical TSM algorithms, such as phase vocoder [71], and WSOLA [72]. The duration of this additional dataset of unison mixtures is approximately 22 hours.

D. 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 [73], (see, e.g., the excerpts with the PCD ID 025 and 026 in the test dataset [28]). 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*). Our previous work [2] exploits this property of the piano concertos for further finetuning the MSS model at test time, a strategy called test-time adaptation [74]. Several works in the literature apply activity-based approaches as a prior to enhance audio source separation, e.g., [75], [76]. Inspired by this strategy, we randomly mask out passages either in the piano or in the orchestral track (but never simultaneously), see Fig. 3(d) for an illustration.

V. 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

[4] [Online]. Available: https://www.audiolabs-erlangen.de/resources/MIR/PCD

[5] [Online]. Available: https://github.com/CPJKU/madmom

approaches. First, we outline our experimental settings in Section V-A. Then, in Section V-B, we provide a brief description of our test dataset [28]. We discuss the quantitative empirical results in Section V-C and present the results of our listening tests in Section V-D. Finally, we elaborate in more detail on the impact of transfer learning and unison mixing in Section V-E.

A. 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 [4], [5], [8], [9]. For UMX, we use two different settings, where we train one model with 6-second random chunks (in [4], 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 V-E. We initially train our models starting with random initialization, using the artificial dataset generated through random mixes with various gains, as detailed in Section IV-A. 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 IV-B). 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 IV-C). 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 IV-D). 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 [77], which exploits, particularly in the first three stages, previously learned concepts to ease the learning of new abstractions.

B. Piano Concerto Dataset (PCD)

For assessing the quantitative and subjective evaluation of our experiments, we use the dry recordings without artificial reverberation from PCD [28] as our test dataset, which contains 81 excerpts with separate piano and orchestral tracks, performed by five pianists. These excerpts are carefully selected from piano concertos written by 10 different composers, spanning from the Baroque to the Post-Romantic era. The excerpts represent a variety of harmonic and structural characteristics of piano concertos from different periods. Additionally, the dataset embraces a wide range of acoustic characteristics ranging from a small and relatively dry domestic room, small recital halls, to a spacious concert hall environment. Moreover, each excerpt has a duration of 12 seconds, which is recommended as the maximum duration for MUSHRA listening tests [32].

C. Quantitative Evaluation

To get a first impression of the model performances, we use the SDR [29] as our quantitative evaluation metric for the separation task. Table II 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 I).

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

TABLE II
MEAN SDR VALUES AND VARIANCES OF DIFFERENT MODELS TRAINED WITH VARIOUS DATA AUGMENTATION METHODS

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**

The mean values and variances are computed over all test items.

TABLE III
MEAN 2F-SCORE VALUES AND VARIANCES OF DIFFERENT MODELS TRAINED WITH VARIOUS DATA AUGMENTATION METHODS

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

The mean values and variances are computed over all test items.

separated musical sources [78]. In particular, the analysis conducted by Torcoli et al. [31] for the source separation task reveals that the *2f-score* metric demonstrates the strongest correlation with ground-truth data based on subjective ratings from MUSHRA listening tests. For a more detailed account on the 2f-score, we refer to [30]. Note that the 2f-score values lie in a range from 0 to 100 following the MUSHRA framework (also see Section V-D). Table III 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 IV-C). Applying silence masking slightly worsens the resulting 2f-scores for HDMC.

D. 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 [32]. 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 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 [79], we created this lower anchor by low-pass filtering the test mixtures with a 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 IV. For an overview of the test items used for the listening test, please refer to our demo webpage.[6]

Fig. 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

[6] [Online]. Available: https://www.audiolabs-erlangen.de/resources/MIR/2024-TASLP-PianoConcertoSeparation/

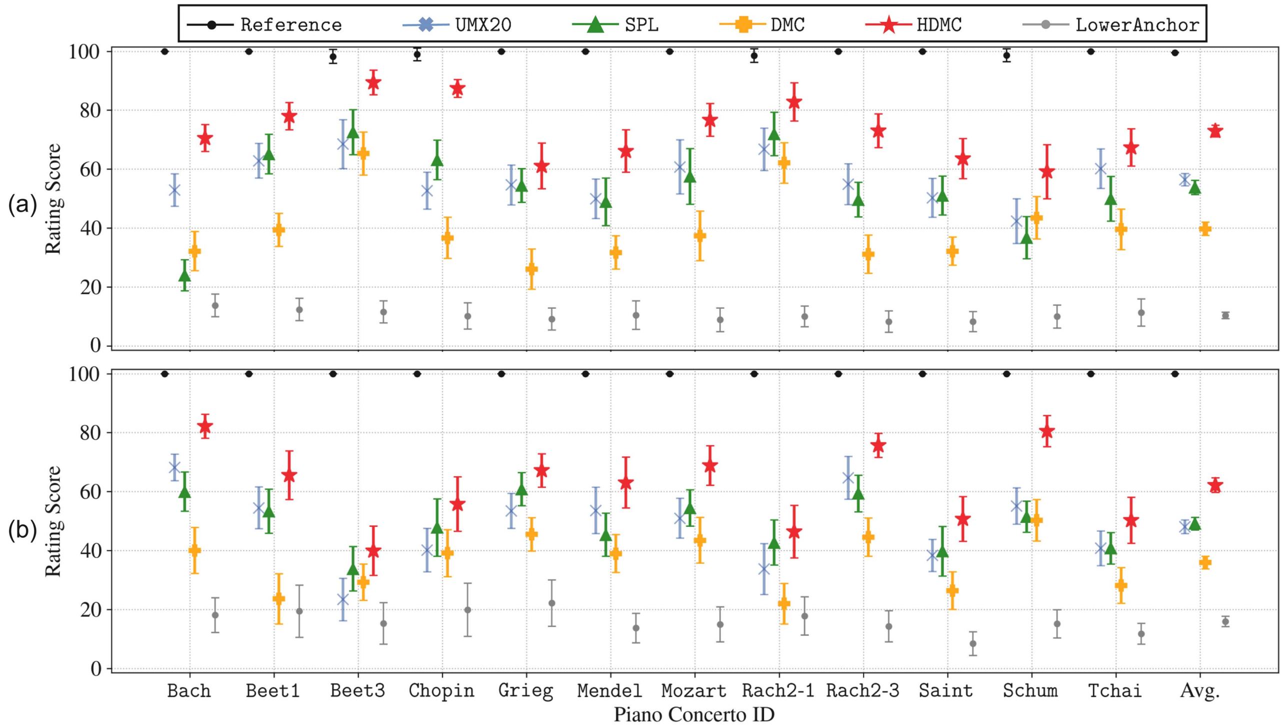

Fig. 4. Results of our listening tests based on the MUSHRA framework for the **(a)** piano and **(b)** orchestral tracks. The listening test employs models that all incorporate the complete data augmentation approach (R_H_HU_HUS). The colored markers indicate the average rating scores enclosed by 95% confidence intervals (shown as the vertical lines).

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).

E. 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. 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 IV 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 V-C 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

TABLE IV
Mean SDR Values and Variances of Different Models Trained With Various Data Augmentation Methods

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	$\mathbf{9.30 \pm 4.00}$	3.86 ± 2.34	4.17 ± 2.13	$\mathbf{4.53 \pm 2.46}$

The mean values and variances are computed over all test items.

TABLE V
Mean SDR Values and Variances of Different Models Trained With Various Data Augmentation Methods

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	$\mathbf{9.41 \pm 4.18}$	4.16 ± 2.46	4.33 ± 2.22	$\mathbf{4.61 \pm 2.39}$

The mean values and variances are computed over all test items.

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 V-A.

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 V. 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 II, 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.

VI. Conclusion

In this work, 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 multi-track 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.

Acknowledgment

Special thanks to Jakob Abeßer, Vlora Arifi-Müller, Christian Dittmar, Michael Krause, Gabriel Meseguer-Brocal, Geoffroy Peeters, Jordi Pons, Simon Schwär, and Nazif Can Tamer for fruitful discussions. The authors are with the International Audio Laboratories Erlangen, a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer Institute for Integrated Circuits IIS.

References

[1] W. Cole, *The Form of Music*. London, U.K.: The Associated Board of the Royal Schools of Music (ABRSM), 1997.

[2] Y. Özer and M. Müller, "Source separation of piano concertos with test-time adaptation," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2022, pp. 493–500.

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

[4] F. Stöter, S. Uhlich, A. Liutkus, and Y. Mitsufuji, "Open-Unmix–A reference implementation for music source separation," *J. Open Source Softw.*, vol. 4, no. 41, 2019, Art. no. 1667. [Online]. Available: https://doi.org/10.21105/joss.01667

[5] R. Hennequin, A. Khlif, F. Voituret, and M. Moussallam, "Spleeter: A fast and efficient music source separation tool with pre-trained models," *J. Open Source Softw.*, vol. 5, no. 50, Art. no. 2154, 2020. [Online]. Available: https://doi.org/10.21105/joss.02154

[6] A. Jansson, E. J. Humphrey, N. Montecchio, R. M. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep U-net convolutional networks," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2017, pp. 745–751.

[7] D. Stoller, S. Ewert, and S. Dixon, "Wave-U-Net: A multi-scale neural network for end-to-end audio source separation," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2018, pp. 334–340.

[8] A. Défossez, N. Usunier, L. Bottou, and F. R. Bach, "Music source separation in the waveform domain," 2021. [Online]. Available: http://arxiv.org/abs/1911.13254

[9] A. Défossez, "Hybrid spectrogram and waveform source separation," in *Proc. ISMIR Workshop Music Source Separation*, 2021, pp. 1–13.

[10] X. Song, Q. Kong, X. Du, and Y. Wang, "CatNet: Music source separation system with mix-audio augmentation," 2021. [Online]. Available: https://arxiv.org/abs/2102.09966

[11] S. Rouard, F. Massa, and A. Défossez, "Hybrid transformers for music source separation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2023, pp. 1–5.

[12] Y. Luo and J. Yu, "Music source separation with Band-Split RNN," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 31, pp. 1893–1901, 2023.

[13] F. Stöter, A. Liutkus, and N. Ito, "The 2018 signal separation evaluation campaign," in *Proc Int. Conf. Latent Variable Anal. Signal Separation*, 2018, pp. 293–305.

[14] Z. Rafii, A. Liutkus, F. Stöter, S. I. Mimilakis, and R. Bittner, "The MUSDB18 corpus for music separation," 2019. [Online]. Available: https://doi.org/10.5281/zenodo.3338373

[15] Y. Mitsufuji et al., "Music demixing challenge 2021," *Front. Signal Process.*, vol. 1, pp. 1–14, 2022.

[16] M. Miron, J. Janer, and E. Gómez, "Monaural score-informed source separation for classical music using convolutional neural networks," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2017, pp. 55–62.

[17] M. Gover and P. Depalle, "Score-informed source separation of choral music," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2020, pp. 231–239.

[18] D. Petermann, P. Chandna, H. Cuesta, J. Bonada, and E. Gómez, "Deep learning based source separation applied to choir ensembles," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2020, pp. 733–739.

[19] K. Chen et al., "Improving choral music separation through expressive synthesized data from sampled instruments," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, Bengaluru, India, 2022, pp. 726–732.

[20] R. M. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. P. Bello, "MedleyDB: A multitrack dataset for annotation-intensive MIR research," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, Taipei, Taiwan, 2014, pp. 155–160.

[21] M. Müller, T. Prätzlich, and C. Dittmar, "Freischütz digital–when computer science meets musicology," in *Proc. Festschrift fürJoachimVeitzum Geburtstag*, 2016, pp. 551–573.

[22] M. Schedl, D. Hauger, M. Tkalčič, M. Melenhorst, and C. C. S. Liem, "A dataset of multimedia material about classical music: PHENICX-SMM," in *Proc. Int. Workshop Content-Based Multimedia Indexing*, 2016, pp. 1–4.

[23] B. Li, X. Liu, K. Dinesh, Z. Duan, and G. Sharma, "Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications," *IEEE Trans. Multimedia*, vol. 21, no. 2, pp. 522–535, Feb. 2019.

[24] S. Rosenzweig, H. Cuesta, C. Weiß, F. Scherbaum, E. Gómez, and M. Müller, "Dagstuhl ChoirSet: A multitrack dataset for MIR research on choral singing," *Trans. Int. Soc. Music Inf. Retrieval*, vol. 3, no. 1, pp. 98–110, 2020.

[25] C. Böhm, D. Ackermann, and S. Weinzierl, "A multi-channel anechoic orchestra recording of Beethoven's symphony no. 8 op. 93," *J. Audio Eng. Soc.*, vol. 68, no. 12, pp. 977–984, 2021.

[26] S. Sarkar, E. Benetos, and M. Sandler, "EnsembleSet: A new high quality dataset for chamber ensemble separation," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, Bengaluru, India, 2022, pp. 625–632.

[27] C.-Y. Chiu, W.-Y. Hsiao, Y.-C. Yeh, Y.-H. Yang, and A. W.-Y. Su, "Mixing-specific data augmentation techniques for improved blind violin/piano source separation," in *Proc. Workshop Multimedia Signal Process.*, 2020, pp. 1–6.

[28] Y. Özer, S. Schwär, V. Arifi-Müller, J. Lawrence, E. Sen, and M. Müller, "Piano concerto dataset (PCD): A multitrack dataset of piano concertos," *Trans. Int. Soc. Music Inf. Retrieval*, vol. 6, no. 1, pp. 75–88, 2023.

[29] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jun. 2006.

[30] T. Kastner and J. Herre, "An efficient model for estimating subjective quality of separated audio source signals," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, New Paltz, New York, USA, 2019, pp. 95–99.

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

[32] B. Series, "Method for the subjective assessment of intermediate quality level of audio systems," in *Proc. Int. Telecommun. Union Radiocommunication Assem.*, 2015.

[33] A. Liutkus and R. Badeau, "Generalized Wiener filtering with fractional power spectrograms," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Brisbane, Australia, 2015, pp. 266–270.

[34] H.-S. Choi, J.-H. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee, "Phase-aware speech enhancement with deep complex U-Net," in *Proc. Int. Conf. Learn. Representations*, 2018, pp. 1–20.

[35] C. Trabelsi et al., "Deep complex networks," in *Proc. Int. Conf. Learn. Representations*, 2018, pp. 1–19.

[36] H. Liu, Q. Kong, and J. Liu, "CWS-PResUNet: Music source separation with channel-wise subband phase-aware ResUNet," in *Proc. ISMIR Workshop Music Source Separation*, 2021, pp. 1–5.

[37] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, "TF-GRIDNET: Making time-frequency domain models great again for monaural speaker separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Rhodes Island, Greece, 2023, pp. 1–5.

[38] W. Choi, M. Kim, J. Chung, D. Lee, and S. Jung, "Investigating U-Nets with various intermediate blocks for spectrogram-based singing voice separation," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2020, pp. 192–198.

[39] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 2, pp. 236–243, Apr. 1984.

[40] Z. Pruša and P. L. Søndergaard, "Real-time spectrogram inversion using phase gradient heap integration," in *Proc. Int. Conf. Digit. Audio Effects*, Brno, Czech Republic, 2016, pp. 17–21.

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

[42] M. Kim, W. Choi, J. Chung, D. Lee, and S. Jung, "KUIELab-MDX-Net: A two-stream neural network for music demixing," in *Proc. ISMIR Workshop Music Source Separation*, 2021, pp. 1–7.

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

[44] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[45] S. Uhlich et al., "Improving music source separation based on deep neural networks through data augmentation and network blending," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, New Orleans, Louisiana, USA, 2017, pp. 261–265.

[46] E. Gusó, J. Pons, S. Pascual, and J. Serrá, "On loss functions and evaluation metrics for music source separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 306–310.

[47] R. Sawata, S. Uhlich, S. Takahashi, and Y. Mitsufuji, "All for one and one for all: Improving music separation by bridging networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Toronto, ON, Canada, 2021, pp. 51–55.

[48] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, Munich, Germany, 2015, pp. 234–241.

[49] G. Meseguer-Brocal and G. Peeters, "Conditioned-U-Net: Introducing a control mechanism in the U-Net for multiple source separations," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, Delft, Netherlands, 2019, pp. 159–165.

[50] A. Cohen-Hadria, A. Roebel, and G. Peeters, "Improving singing voice separation using deep U-Net and wave-U-Net with data augmentation," in *Proc. Eur. Signal Process. Conf.*, 2019, pp. 1–5.

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

[52] C. Hawthorne et al., "Enabling factorized piano music modeling and generation with the MAESTRO dataset," in *Proc. Int. Conf. Learn. Representations*, New Orleans, Louisiana, USA, 2019, pp. 1–12. [Online]. Available: https://openreview.net/forum?id=r1lYRjC9F7

[53] H. Zhang, J. Tang, S. R. M. Rafee, S. Dixon, and G. Fazekas, "ATEPP: A dataset of automatically transcribed expressive piano performance,"

in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, Bengaluru, India, 2022, pp. 446–453.
[54] Z. Duan and B. Pardo, "Soundprism: An online system for score-informed source separation of music audio," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 6, pp. 1205–1215, Oct. 2011.
[55] J. J. Bosch, R. Marxer, and E. Gómez, "Evaluation and combination of pitch estimation methods for melody extraction in symphonic classical music," *J. New Music Res.*, vol. 45, no. 2, pp. 101–117, 2016.
[56] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Music genre database and musical instrument sound database," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, Baltimore, Maryland, USA, 2003, pp. 229–230.
[57] S. Yuan et al., "Improved singing voice separation with chromagram-based pitch-aware remixing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 111–115.
[58] H. Kim, J. Park, T. Kwon, D. Jeong, and J. Nam, "A study of audio mixing methods for piano transcription in violin-piano ensembles," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Rhodes Island, Greece, 2023, pp. 1–5.
[59] F.-R. Stöter, S. Bayer, and B. Edler, "Unison source separation," in *Proc. Int. Conf. Digit. Audio Effects*, Erlangen, Germany, 2014, pp. 235–241.
[60] F. Stöter, A. Liutkus, R. Badeau, B. Edler, and P. Magron, "Common fate model for unison source separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Shanghai, China, 2016, pp. 126–130.
[61] S. Salvador and P. Chan, "FastDTW: Toward accurate dynamic time warping in linear time and space," in *Proc. KDD Workshop Mining Temporal Sequential Data*, 2004.
[62] S. Dixon and G. Widmer, "MATCH: A music alignment tool chest," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, London, U.K., 2005, pp. 492–497.
[63] R. B. Dannenberg and N. Hu, "Polyphonic audio matching for score following and intelligent audio editors," in *Proc. Int. Comput. Music Conf.*, San Francisco, USA, 2003, pp. 27–34.
[64] M. Müller, *Fundamentals of Music Processing–Using Python and Jupyter Notebooks*, 2nd ed. Berlin, Germany: Springer, 2021.
[65] S. Ewert, M. Müller, and P. Grosche, "High resolution audio synchronization using chroma onset features," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Taipei, Taiwan, 2009, pp. 1869–1872.
[66] Y. Özer, M. Ištvánek, V. Arifi-Müller, and M. Müller, "Using activation functions for improving measure-level audio synchronization," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, Bengaluru, India, 2022, pp. 749–756.
[67] S. Böck, F. Korzeniowski, J. Schlüter, F. Krebs, and G. Widmer, "madmom: A new python audio and music signal processing library," in *Proc. ACM Int. Conf. Multimedia*, Amsterdam, The Netherlands, 2016, pp. 1174–1178.
[68] M. Müller, Y. Özer, M. Krause, T. Prätzlich, and J. Driedger, "Sync toolbox: A python package for efficient, robust, and accurate music synchronization," *J. Open Source Softw.*, vol. 6, no. 64, Art. no. 3434, 2021.
[69] T. Prätzlich, J. Driedger, and M. Müller, "Memory-restricted multiscale dynamic time warping," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Shanghai, China, 2016, pp. 569–573.
[70] J. Driedger, M. Müller, and S. Ewert, "Improving time-scale modification of music signals using harmonic–percussive separation," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 105–109, Jan. 2014.
[71] J. L. Flanagan and R. M. Golden, "Phase vocoder," *Bell Syst. Tech. J.*, vol. 45, pp. 1493–1509, 1966.
[72] W. Verhelst and M. Roelands, "An overlap–add technique based on waveform similarity (WSOLA) for high quality time–scale modification of speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Minneapolis, USA, 1993, pp. 554–557.
[73] W. E. Caplin., *Classical Form: A. Theory of Formal Functions for the Instrumental Music of Haydn, Mozart, and Beethoven*. London, U.K.: Oxford Univ. Press, 1998.
[74] Y. Sun, X. Wang, L. Zhang, J. Miller, M. Hardt, and A. A. Efros, "Test-time training with self-supervision for generalization under distribution shifts," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 9229–9248.
[75] K. Schulze-Forster, C. S. J. Doire, G. Richard, and R. Badeau, "Weakly informed audio source separation," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, New Paltz, USA, 2019, pp. 273–277.
[76] M. Torcoli and E. A. P. Habets, "Better together: Dialogue separation and voice activity detection for audio personalization in TV," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Rhodes Island, Greece, 2023, pp. 1–5.
[77] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. Int. Conf. Mach. Learn.*, New York, NY, USA, 2009, pp. 41–48. [Online]. Available: https://doi.org/10.1145/1553374.1553380
[78] E. Cano, D. FitzGerald, and K. Brandenburg, "Evaluation of quality of sound source separation algorithms: Human perception vs quantitative metrics," in *Proc. Eur. Signal Process. Conf.*, 2016, pp. 1758–1762.
[79] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2046–2057, Sep. 2011.

Yigitcan Özer received the B.Sc. degree in electrical and electronics engineering from Bilkent University, Ankara, Turkey, in 2013, and the M.Sc. degree in communications engineering from the Technical University of Munich, Munich, Germany, in 2015. He is currently working toward the Ph.D. degree with International Audio Laboratories Erlangen, Erlangen, Germany, under the supervision of Prof. Meinard Müller. In January 2021, he joined the International Audio Laboratories Erlangen. Before his Ph.D. studies, he was a Research Associate with the Spoken Language Processing Group, Fraunhofer Institute for Integrated Circuits IIS, Erlangen, Germany. His research interests include audio source separation, music synchronization, and text-to-speech synthesis.

Meinard Müller (Fellow, IEEE) received the Diploma in mathematics and the Ph.D. degree in computer science from the University of Bonn, Bonn, Germany, in 1997 and 2001, respectively. After his postdoctoral studies during 2001–2003, in Japan, and his habilitation during 2003–2007, in multimedia retrieval in Bonn, he was a senior Researcher with Saarland University, Saarbrücken, Germany, and the Max-Planck Institut für Informatik during 2007–2012. Since 2012, he has been holding a professorship of semantic audio signal processing with the International Audio Laboratories Erlangen, a joint institute of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and the Fraunhofer Institute for Integrated Circuits IIS. His research interests include music processing, music information retrieval, audio signal processing, and motion processing. He was a Member of the IEEE Audio and Acoustic Signal Processing Technical Committee during 2010–2015, a Member of the Senior Editorial Board of the IEEE Signal Processing Magazine during 2018–2022, and a Member of the Board of Directors, International Society for Music Information Retrieval during 2009–2021, being its president in 2020/2021. In 2020, he was elevated to IEEE Fellow for contributions to music signal processing.