The Sound Demixing Challenge 2023 – Cinematic Demixing Track

Uhlich, Stefan; Fabbro, Giorgio; Hirano, Masato; Takahashi, Shusuke; Wichern, Gordon; Le Roux, Jonathan; Chakraborty, Dipam; Mohanty, Sharada; Li, Kai; Luo, Yi; Yu, Jianwei; Gu, Rongzhi; Solovyev, Roman; Stempkovskiy, Alexander; Habruseva, Tatiana; Sukhovei, Mikhail; Mitsufuji, Yuki

Abstract

This paper summarizes the cinematic demixing (CDX) track of the Sound Demixing Challenge 2023 (SDX'23). We provide a comprehensive summary of the challenge setup, detailing the structure of the competition and the datasets used. Especially, we detail CDXDB23, a new hid- den dataset constructed from real movies that was used to rank the submissions. The paper also offers insights into the most successful approaches employed by participants. Compared to the cocktail-fork baseline, the best-performing system trained exclusively on the simulated Divide and Remaster (DnR) dataset achieved an improvement of 1.8 dB in SDR, whereas the top-performing system on the open leaderboard, where any data could be used for training, saw a significant improvement of 5.7 dB. A significant source of this improvement was making the simulated data better match real cinematic audio, which we further investigate in detail.

OVERVIEW ARTICLE

The Sound Demixing Challenge 2023 – Cinematic Demixing Track

Stefan Uhlich,[*] Giorgio Fabbro,[*] Masato Hirano,[†] Shusuke Takahashi,[†] Gordon Wichern,[‡] Jonathan Le Roux,[‡] Dipam Chakraborty,[§] Sharada Mohanty,[¶] Kai Li,[||][**] Yi Luo,[||] Jianwei Yu,[||] Rongzhi Gu,[||] Roman Solovyev,[††] Alexander Stempkovskiy,[††] Tatiana Habruseva,[‡‡] Mikhail Sukhovei,[§§] and Yuki Mitsufuji[†]

Abstract

This paper summarizes the cinematic demixing (CDX) track of the Sound Demixing Challenge 2023 (SDX'23). We provide a comprehensive summary of the challenge setup, detailing the structure of the competition and the datasets used. Especially, we detail CDXDB23, a new hidden dataset constructed from real movies that was used to rank the submissions. The paper also offers insights into the most successful approaches employed by participants. Compared to the cocktail–fork baseline, the best–performing system trained exclusively on the simulated Divide and Remaster (DnR) dataset achieved an improvement of 1.8 dB in SDR, whereas the top–performing system on the open leaderboard, where any data could be used for training, saw a significant improvement of 5.7 dB. A significant source of this improvement was making the simulated data better match real cinematic audio, which we further investigate in detail.

Keywords: Cinematic sound separation, dialogue enhancement, sound effect extraction

1. Introduction

Cinematic source separation refers to the task of separating movie audio into *dialogue* (DX), *music* (MX), and *sound effects* (FX). While speech separation (Hershey et al., 2016; Chen et al., 2017; Yu et al., 2017) and music separation (Huang et al., 2012; Grais et al., 2014; Uhlich et al., 2015) have been studied extensively, cinematic source separation is a relatively recent field (Petermann et al., 2022) despite its numerous practical applications. These include enhancing old movies by converting them to formats like MPEG-H or Dolby Atmos, dubbing them into different languages, or generating subtitles including non-speech sounds present in

an auditory scene.

The first work in the area of cinematic separation was dialogue enhancement in (Uhle et al., 2008; Geiger et al., 2015)(Paulus et al., 2019; Torcoli et al., 2021), which employs source separation to extract and remix the dialogue signal at a desired level. The problem was further formalized in (Petermann et al., 2022), where Petermann et al. introduced cinematic separation as a three-way problem of splitting the audio into dialogue, sound effects, and music, which they referred to as the *cocktail fork problem*. They created a new dataset, *Divide and Remaster* (DnR), which was built upon LibriSpeech (Panayotov et al., 2015) for dialogue, Free Music Archive (Defferrard et al., 2016) for music, and the Freesound Dataset 50k (Fonseca et al., 2021) for sound effects. Their exploration of various separation models revealed that their proposed multi-resolution extension of X-UMX (Sawata et al., 2021, 2023), termed MRX, provided the best performance. Subsequently, the same authors extended this work in (Petermann et al., 2023) to also consider the impact of source separation on downstream tasks. They proposed a two-stage approach where an MRX separator is used to obtain preliminary separation, which is fol-

---

[*] Sony Europe B.V., Stuttgart, Germany
[†] Sony Group Corporation, Tokyo, Japan
[‡] Mitsubishi Electric Research Labs, Cambridge, USA
[§] Alcrowd, Bengaluru, India
[¶] Alcrowd, Lausanne, Switzerland
[||] Tencent AI Lab, Shenzhen, China
[**] Department of Computer Science and Technology, Tsinghua University, Beijing, China. Work done during internship at Tencent AI Lab.
[††] Institute for Design Problems in Microelectronics of Russian Academy of Sciences, Moscow, Russian Federation
[‡‡] Independent researcher, Cork, Ireland
[§§] Independent researcher, Saint-Petersburg, Russia

lowed by an activity detector to estimate the activity profile for every source. This activity information is then utilized in a second stage by a conditioned MRX, called MRX-C, to improve the separation performance. Recently, DnR was also used in (Watcharasupat et al., 2023), which extended the band-split RNN (Luo and Yu, 2023) to cinematic separation by introducing the BandIt architecture.

Cinematic separation has several unique challenges compared to speech or music separation. Firstly, the multi-channel format of most cinematic audio (stereo or 5.1 surround) necessitates a suitable augmentation during training, as many datasets are only monaural, such as the DnR dataset. Secondly, the scarcity of full-bandwidth material with sampling rate of 48 kHz for training poses a significant hurdle, as high-quality audio data is essential for effective model training. Thirdly, the lack of emotional speech within the used speech datasets presents a challenge. Separation models trained on these datasets often struggle with emotional speech in real cinematic dialogues, because it is typically absent from the training data as was already noted in earlier work (Uhle et al., 2008). Fourthly, the sound effects class, which encompasses a wide variety of sounds, is particularly challenging to extract due to its broad and diverse nature[1]. Finally, the three classes exhibit some overlap, such as the presence of vocals in music, background chatter (chatter noise), which is a sound effect but shares similarities with dialogue, or the use of musical instruments for sound design as seen in the alien communication signal in *Close Encounters of the Third Kind*, which is a sound effect made of musical notes. These challenges highlight the complexity of cinematic separation and the need for further research and development in this field.

Hence, in addition to the *music demixing* (MDX) track (Fabbro et al., 2023), which was already present in the *Music Demixing Challenge 2021* (MDX'21) (Mitsufuji et al., 2022), we have added a new *cinematic demixing* (CDX) track to the *Sound Demixing Challenge 2023* (SDX'23) in order to foster research in this direction. The challenge was facilitated through AIcrowd[2], and participants were invited to submit their systems to one of two leaderboards, depending on whether they used only DnR or additional training data. To rank the submissions, we developed a new hidden test set, called CDXDB23, derived from real movies. Through the establishment of this challenge framework, we observed substantial performance enhancements. Specifically, the top-performing system, trained solely on DnR, demonstrated an improvement of 1.8 dB compared to the cocktail-fork baseline based on MRX (Petermann et al., 2022). Remarkably, the highest-performing system on the open leaderboard, which allowed the use of any data for training, exhibited a significant improvement of 5.7 dB. These results underscore the efficacy of our challenge in driving advancements in the field of cinematic audio separation.

This paper is organized as follows: Sec. 2 outlines the competition's design, Sec. 3 discusses the training datasets and establishes the performance baseline, Sec. 4 presents the results and summarizes the most successful strategies, and Sec. 5 analyzes the differences between the provided training dataset, DnR, and hidden test set, CDXDB23. Finally Sec. 6 concludes the paper with key findings and future research directions.

2. CDX Challenge Setup
In the following, we will summarize the structure of the competition.

2.1 Task Definition
Participants in the CDX track of SDX'23 were asked to submit systems that can extract the dialogue $\mathbf{s}_{DX}(n) \in \mathbb{R}^2$, sound effects $\mathbf{s}_{FX}(n) \in \mathbb{R}^2$, and music $\mathbf{s}_{MX}(n) \in \mathbb{R}^2$ from the stereo cinematic audio

$$\mathbf{x}(n) = \mathbf{s}_{DX}(n) + \mathbf{s}_{FX}(n) + \mathbf{s}_{MX}(n), \tag{1}$$

where n denotes the time index and all stereo signals are sampled at 44.1 kHz. We used the following definition for each class[3]:

- *Dialogue* refers to all spoken content in a movie including conversations between characters, monologues, and any other spoken elements.
- *Sound effects* are sounds that are used to support or complement the action on screen. They can be split into object sounds (e.g., footsteps) and ambient sounds (e.g., wind or rain).
- *Music* refers to the soundtrack that accompanies the visuals and which is often used to provide an emotional context. It might be a single instrument (e.g., a violin in a dramatic moment) or a full orchestra/music band.

We verified these definitions with mixing engineers from Sony Pictures.

A unique aspect of this challenge was the requirement for participants to submit their pre-trained models along with the corresponding inference code, as the test dataset was kept hidden. This stands in contrast to many other challenges where participants have access to unlabeled test data and are required to submit processed files or labels.

2.2 Leaderboards
Submissions were categorized under two leaderboards:

- *Leaderboard A* was designated for models exclusively trained on the train and validation splits 'tr' and 'cv' of the Divide and Remaster (DnR) dataset (Petermann et al., 2022), while
- *Leaderboard B* was for models trained on any data.

The rationale behind this dual-leaderboard approach is threefold. Firstly, it allows individuals who may not have access to extensive datasets to participate in

the competition. Secondly, it provides a platform to explore data augmentation strategies, such as mono-to-stereo conversion, which is particularly relevant as the DnR dataset is monaural, while the hidden test set used for evaluation is in stereo format. Thus, the two leaderboards not only foster inclusivity but also encourage innovative approaches to data augmentation. Thirdly, the two leaderboards allow disentangling data improvements from algorithm improvements, as Leaderboard B performance could come from extra data or better augmentation strategies relying on additional data (e.g., room impulse responses), while Leaderboard A improvements must come from augmentations without additional data and from algorithms only. However, Leaderboard B is required to determine the true state of the art.

2.3 Ranking Metric

For the evaluation of the systems, we used the global *signal-to-distortion ratio* (SDR) which is defined for one movie clip as

$$\text{SDR} = \frac{1}{3}\Big(\text{SDR}_{\text{DX}} + \text{SDR}_{\text{FX}} + \text{SDR}_{\text{MX}}\Big), \qquad (2)$$

with $\text{SDR}_j = 10\log_{10}\frac{\sum_n \|\mathbf{s}_j(n)\|^2}{\sum_n \|\mathbf{s}_j(n) - \hat{\mathbf{s}}_j(n)\|^2}$ where $\mathbf{s}_j(n) \in \mathbb{R}^2$ and $\hat{\mathbf{s}}_j(n) \in \mathbb{R}^2$ denote the stereo target and estimate for source $j \in \{\text{DX}, \text{FX}, \text{MX}\}$. The definition in (2) is also called utterance-level SDR (cf., for example, (Luo and Yu, 2023)) and equivalent to the SDR of multi-channel BSS Eval v3 (Vincent et al., 2007). Finally, the global SDR of (2) is averaged over all clips in the hidden test dataset and the three sources DX, FX, and MX to obtain the final score. We chose this metric to rank submissions over scale-invariant metrics like SI-SDR (Le Roux et al., 2019), because systems with good SDR performance have the advantage that they can easily be blended with other models (Uhlich et al., 2017) and also allow one to compute the residual $\hat{\mathbf{s}}_{\neg j}(n) = \mathbf{x}(n) - \hat{\mathbf{s}}_j(n)$ without having to recover the correct scale.

Besides the chosen global SDR (2), there are also other metrics that were proposed in the literature for the comparison of source separation models. As part of MDX'21, a thorough comparison of different metrics was done in (Mitsufuji et al., 2022) to show that (2) highly correlates with many others, in particular those that were used in previous iterations of the SiSEC competition in 2015, 2016, and 2018. We refer the interested reader to (Mitsufuji et al., 2022) for more details.

2.4 Timeline, Challenge Phases and Prizes

The challenge took place in two phases. Phase I started on January 23rd, 2023. Phase II commenced on March 6th, 2023, as planned. However, due to the submission system experiencing difficulties in handling the surge in the number of submissions towards the end of the challenge, the end date of Phase II was ex-

tended by one week. Originally scheduled to conclude on May 1st, 2023, the challenge was extended to May 8, 2023, to ensure a fair competition for all teams.

CDXDB23 was partitioned into three sets of approximately equal size, each containing three, three, and four movies respectively. During Phase 1 of the competition, participants were able to assess the performance of their submissions using one-third of the movies from the hidden test set. In Phase 2, this was expanded to include two-thirds of the movies from the hidden test set. Upon the conclusion of Phase 2, participants were required to select three submissions for evaluation on the full hidden test set, the results of which were then displayed on the final leaderboards. This selection process was implemented to mitigate the potential impact of overfitting. In cases where participants did not explicitly select three submissions, the top three submissions from the Phase 2 leaderboard were automatically chosen for final evaluation.

For Leaderboard A, which was for models trained exclusively on the Divide and Remaster (DnR) dataset, a total of 5,000 USD was distributed among the top three submissions. The first-place winner received 2,500 USD, the second-place winner was awarded 1,500 USD, and the third-place winner received 1,000 USD. To be eligible for these prizes, participants were required to open-source both their training and inference code as well as the pretrained model. Similarly, for Leaderboard B, which was for models trained on any data, the same prize distribution was applied: 2,500 USD for the first place, 1,500 USD for the second place, and 1,000 USD for the third place, totaling 5,000 USD. For this leaderboard, participants were required to open-source their inference code as well as the pretrained model. Compliance with these open-source requirements was ensured by the organizers through a due diligence check. In the course of this evaluation, a thorough review of the source code was conducted to verify that participants in Leaderboard A exclusively trained their models using only DnR.

3. Datasets and Baseline

The following subsections offer detailed descriptions of the datasets employed throughout the challenge, as well as an overview of the baseline included in the starter kit.

3.1 Divide and Remaster (DnR) – Training Dataset

Introduced in (Petermann et al., 2022), the Divide and Remaster (DnR) dataset serves as a tool for developing and evaluating mono audio signal separation algorithms applied to podcasts, television, and movies. It includes artificial mixtures sourced from LibriSpeech (Panayotov et al., 2015), Free Music Archive (FMA) (Defferrard et al., 2016), and Freesound Dataset 50k (FSD50k) (Fonseca et al., 2021). This dataset, available in both 16 kHz and 44.1 kHz sampling rates, comes with time-stamped annotations for each class:

genre for music, audio-tags for sound effects, and transcription for speech.

The creation process of DnR was centered on addressing class overlap and relative source levels in the mix within a single-channel[4] context. It includes four categories: speech, music, foreground effects, and background effects – the latter two being merged into a single submix. All mixtures have a duration of 60 seconds, encompassing multiple full speech utterances and sufficient onsets and offsets between classes. File count for each class was set via a zero-truncated Poisson distribution, and relative amplitude levels across the classes were determined per industry standards and prior studies as discussed in (Petermann et al., 2022). Each sound file's gain was individually adjusted to add variability while preserving realistic consistency across the mix. The final dataset, divided into training, validation, and testing subsets in line with base dataset proportions, comprises 3,406 training mixtures, 487 validation mixtures, and 973 test mixtures.

While the DnR dataset took care to simulate realistic cinematic mixtures, there are some notable differences between the source material used to create DnR and actual cinematic audio:

- **Read speech vs. emotional speech** – First, LibriSpeech contains read speech from audio books, which may have significant timbral differences compared to the emotional speech typically used by film actors.
- **Vocals in music stems** – Second, many of the musical genres from the FMA dataset contain vocals. While music with vocals is used in cinema, the majority of cinematic music does not contain singing. Thus, the prevalence of music with vocals may be overrepresented in FMA compared to the hidden test data.
- **Production quality** – Finally, Librispeech, FMA, and FSD50K are all crowd-sourced datasets, and there may be significant differences in terms of recording hardware and post-production effects compared to actual movies. We will investigate this in more detail in Sec. 5.

In summary, it is expected that mismatches such as this may limit performance of separation models trained only on DnR.

For Leaderboard A, participants were required to only utilize the training and validation split of the DnR dataset in training their systems.

3.2　CDXDB23 – Hidden Test Dataset

To rank the submissions, we generated a novel dataset derived from authentic Sony Pictures movies and we will refer in the following to this dataset as *cinematic demixing database* (CDXDB23). It comprises 11 movies with a total of 156 clips each with an average length of 11 seconds, amounting overall to approximately 19.4 minutes of content. The audio was originally

in a higher sample rate, but we downsampled it to 44.1 kHz stereo to match the sample rate of the DnR dataset. This was done to avoid requiring participants in Leaderboard A to design systems that can upscale to a higher sampling rate. Fig. 1 shows the distribution of the genres as well as the release years of the eleven movies inside CDXDB23. Please note that a single movie can fall under multiple genres, such as *Animation* and *Family*. This characteristic is reflected in the bar plot, where the representation of movies in various genres contributes to the observed distribution. From Fig. 1 we can observe that they are recent movies covering a wide variety of genres.

The original data supplied by Sony Pictures was formatted as 5.1ch, 48 kHz, and 24-bit with several *stem* tracks for each movie, encompassing either dialogue, effects, music, or their combination. We manually annotated the sound events with one class label (dialogue, sound effects, or music) within each stem and carefully selected segments to ensure a balanced representation of each class in the resulting mixture. Specifically, to exclude extremely low-amplitude sound sources, we computed for each class the *root mean square* $\mathrm{RMS}_j = \left(\frac{1}{N}\sum_{n=1}^{N}\|\mathbf{s}_j(n)\|^2\right)^{\frac{1}{2}}$ and excluded segments where this value was below a threshold τ_j for any $j \in \{\mathrm{DX}, \mathrm{FX}, \mathrm{MX}\}$. Empirically, we found the thresholds

$$\tau_{\mathrm{DX}} = 0.022, \qquad \tau_{\mathrm{FX}} = 0.005, \qquad \tau_{\mathrm{MX}} = 0.003,$$

to give good test samples. On occasion, environmental noise was unintentionally recorded, or dialogue/vocal components appeared in the effects or music stems. We made diligent efforts to minimize the inclusion of such samples by manually inspecting all data.

Please note that we are unable to provide further details about the movies (e.g., title or actors) to prevent participants in a future challenge from fine-tuning their models based on this specific information. However, we made available demo samples from "Kilian's Game", a short film produced by Sony Pictures to demonstrate the latest filmmaking technologies. These samples could be used by participants to test their submissions and to see the performance on real movie audio.[5] The samples from "Kilian's Game" were not used to rank the submissions.

Ideally, we should also use authentic movie data for training models. During the preparation of CDXDB23, we noticed that this is actually not straightforward. One problem is the preparation of the three-way stems from movie audio, which is a time-intensive process. The material is not readily accessible, requiring reloading all raw tracks into the Digital Audio Workstation (DAW), deciding for each of them the class it belongs to, and finally bouncing the stems for each class. Additionally, we noticed the challenge of other sound classes infiltrating a single stem. For example, dialogue stems can contain sound effects recorded on-
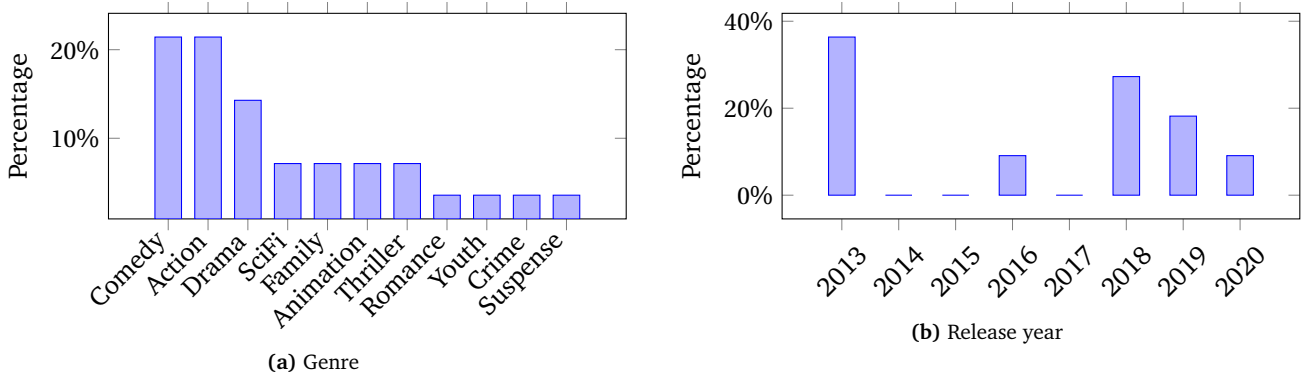
(a) Genre

(b) Release year

Figure 1: Statistics about movies in CDXDB23.

stage. Consequently, after bouncing the stems, one has to manually annotate all audio material to find suitable time regions for the three-way separation, leading to a smaller dataset suitable only for testing, as exemplified by CDXDB23.

3.3 Cocktail–Fork Baseline

As part of the challenge, MERL open-sourced their *multi-resolution CrossNet* (MRX) (Petermann et al., 2022), an improved version of *CrossNet-UMX* (X-UMX) (Sawata et al., 2021, 2023), which itself is an improved version of Open-Unmix (UMX) (Stöter et al., 2019). MRX leverages multiple *short-time Fourier transform* (STFT) resolutions of the mixture, enhancing the estimation process as it allows to better address the variety of acoustic characteristics of the three source types. The entire system is available on GitHub[6].

Using the available pre-trained model on DnR, a baseline submission was created and made available for all participants as part of the starter-kit[7]. We noticed already during the preparation of the baseline that scaling the input mixture is beneficial and, hence, apply the scaling

$$\mathbf{x}(n) \leftarrow \frac{\mathbf{x}(n)}{\max_n |\mathbf{x}(n)|}, \qquad (3)$$

i.e., the cocktail-fork model is run on the peak normalized mixture. Training of MRX utilized scale-invariant signal-to-distortion ratio (SI-SDR) loss, necessitating subsequent scale estimation using least-squares according to the formula

$$\hat{\mathbf{s}}_j(n) \leftarrow \frac{\sum_n \mathbf{x}(n)^T \hat{\mathbf{s}}_j(n)}{10^{-7} + \sum_n \|\hat{\mathbf{s}}_j(n)\|^2} \hat{\mathbf{s}}_j(n) \qquad (4)$$

for any $j \in \{DX, FX, MX\}$. Furthermore, a post-processing step was implemented to ensure mixture consistency (Wisdom et al., 2019), where we first compute the residual $\mathbf{r}(n) = \mathbf{x}(n) - \hat{\mathbf{s}}_{DX}(n) - \hat{\mathbf{s}}_{FX}(n) - \hat{\mathbf{s}}_{MX}(n)$

which is then distributed to the estimates

$$\hat{\mathbf{s}}_{DX}(n) \leftarrow \hat{\mathbf{s}}_{DX}(n),$$

$$\hat{\mathbf{s}}_{FX}(n) \leftarrow \hat{\mathbf{s}}_{FX}(n) + \frac{1}{2}\mathbf{r}(n),$$

$$\hat{\mathbf{s}}_{MX}(n) \leftarrow \hat{\mathbf{s}}_{MX}(n) + \frac{1}{2}\mathbf{r}(n).$$

This post-processing was beneficial as the residual contains mostly sound effects and background music. SDR_{FX} improved by $+1.1$ dB and SDR_{MX} by $+0.7$ dB, resulting in an overall improvement of $+0.6$ dB. The performance of the cocktail-fork baseline on CDXDB23 can be found in Table 1.

After the challenge, we revisited this baseline as many participants recognized a distribution mismatch between DnR and CDXDB23, which can also be seen in Table 1 in the lower scores of this model. In Sec. 5.2, we will present two new versions of the cocktail-fork baseline with improved performance due to adjusting the loudness or equalization of DnR during training.

4. Challenge Outcome

The CDX track saw a dynamic evolution in terms of both the number of submissions and the SDR performance. The competition attracted a total of 19 teams for Leaderboard A and 10 teams for Leaderboard B, with 369 and 179 submissions respectively. Tables 1 and 2 present the final rankings for both leaderboards. The team *aim-less* emerged as the winner of Leaderboard A, achieving an average SDR of 4.345 dB. On the other hand, Leaderboard B was topped by *Jusper-Lee*, with an impressive SDR of 8.181 dB. It is noteworthy that while all top five teams in Leaderboard A were from academic institutions, the highest scores in Leaderboard B were obtained by two commercial entities. This diversity of participants underscores the broad interest and applicability of the challenge across both academic and industry sectors. Fig. 2 shows the progress that the teams could achieve during the course of the competition. We can observe that there was a continuous improvement of the SDR for each source and, especially at the end of the competition,

Table 1: Final Leaderboard A (models trained only on DnR; top-5).

Rank	Participant	Global SDR (dB)				Submissions to Ldb A		
		Mean	Dialogue	Effect	Music	1st phase	2nd phase	
Submissions								
1.	aim-less	4.345	7.981	1.217	3.837	36	32	Code[8]
2.	mp3D	4.237	8.484	1.622	2.607	–	42	Code[9]
3.	subatomicseer	4.144	7.178	2.820	2.433	65	22	Code[10]
4.	thanatoz	3.871	8.948	1.224	1.442	21	22	
5.	kuielab	3.537	7.687	0.449	2.474	36	15	
Baseline								
Scaled Identity $\hat{\mathbf{s}}_j(n) = \frac{1}{3}\mathbf{x}(n)$		-0.019	1.562	-1.236	-0.383			
Cocktail-Fork (Petermann et al., 2022)		2.491	7.321	-1.049	1.200			

Table 2: Final Leaderboard B (models trained on any data; top-5).

Rank	Participant	Global SDR (dB)				Submissions to Ldb A + B		
		Mean	Dialogue	Effect	Music	1st phase	2nd phase	
Submissions								
1.	JusperLee	8.181	14.619	3.958	5.966	42	102	
2.	Audioshake	8.077	14.963	4.034	5.234	–	97	
3.	ZFTurbo	7.630	14.734	3.323	4.834	25	31	Code[11]
4.	aim-less	4.345	7.981	1.217	3.837	36	53	Code[8]
5.	mp3D	4.237	8.484	1.622	2.607	–	48	Code[9]

there is a steady improvement visible as participants tuned their submissions.

To investigate whether this improvement resulted from participants overfitting to the visible portion of the test set, Fig. 3 presents the difference between the hidden SDR (the SDR for all clips of CDXDB23 hidden from the participants) and the visible SDR (the SDR for all clips of CDXDB23 shown to the participants). Comparing two subsequent submissions where the newer one is worse in this difference than the previous one indicates that a participant is obtaining less improvement/more degradation on the hidden SDR than for the visible SDR hinting at a possible overfitting to the displayed global SDR. Hence, seeing "trajectories" of consecutive submissions in Fig. 3 with negative slopes can be used to detect overfitting. Intriguingly, some degree of overfitting is apparent for the submissions to Leaderboard B towards the end of the challenge but less overfitting is observed for submissions to Leaderboard A. For example, looking at the results for the teams *JusperLee* and *Audioshake*, we can see that there is a negative trend in their submissions towards the end of the challenge. Especially for team *Audioshake*, this is visible as the models extracting sound effects and music seem to be tuned in the last week of the challenge period. Consequently, to reduce the potential effect of overfitting, participants needed to select three submissions at the end of the challenge which were then eval-

uated on the full CDXDB23 as discussed in Sec. 2.4.

The substantial improvement upon the provided cocktail-fork baseline by the participants is noteworthy. This was achieved not only through the implementation of enhanced architectures, such as MRX-C (Petermann et al., 2023) used by team *mp3d*, but also through the identification and rectification of two issues inherent in the DnR dataset. Firstly, the presence of vocals in the *music* category necessitated dataset cleaning. Secondly, the difference in loudness exhibited by DnR resulted in suboptimal performance of systems trained on this dataset, necessitating the consideration of this factor as discussed in 4.3.2 by using a suitable input normalization. Interestingly, none of the top teams explored mono-to-stereo augmentations, which presents an intriguing avenue for future research.

Comparing the results for Leaderboard A and B, we can observe that especially *dialogue* gains from having access to additional training data. This is in our opinion due to the access to much more speech and vocals material, which can be used as training material for *dialogue*. Particularly, the inclusion of vocal material proves advantageous due to its similarity to emotional speech. Additionally, the processing pipelines employed in cinematic production may align closely with those utilized in music production, further enhancing the benefit of vocal material.
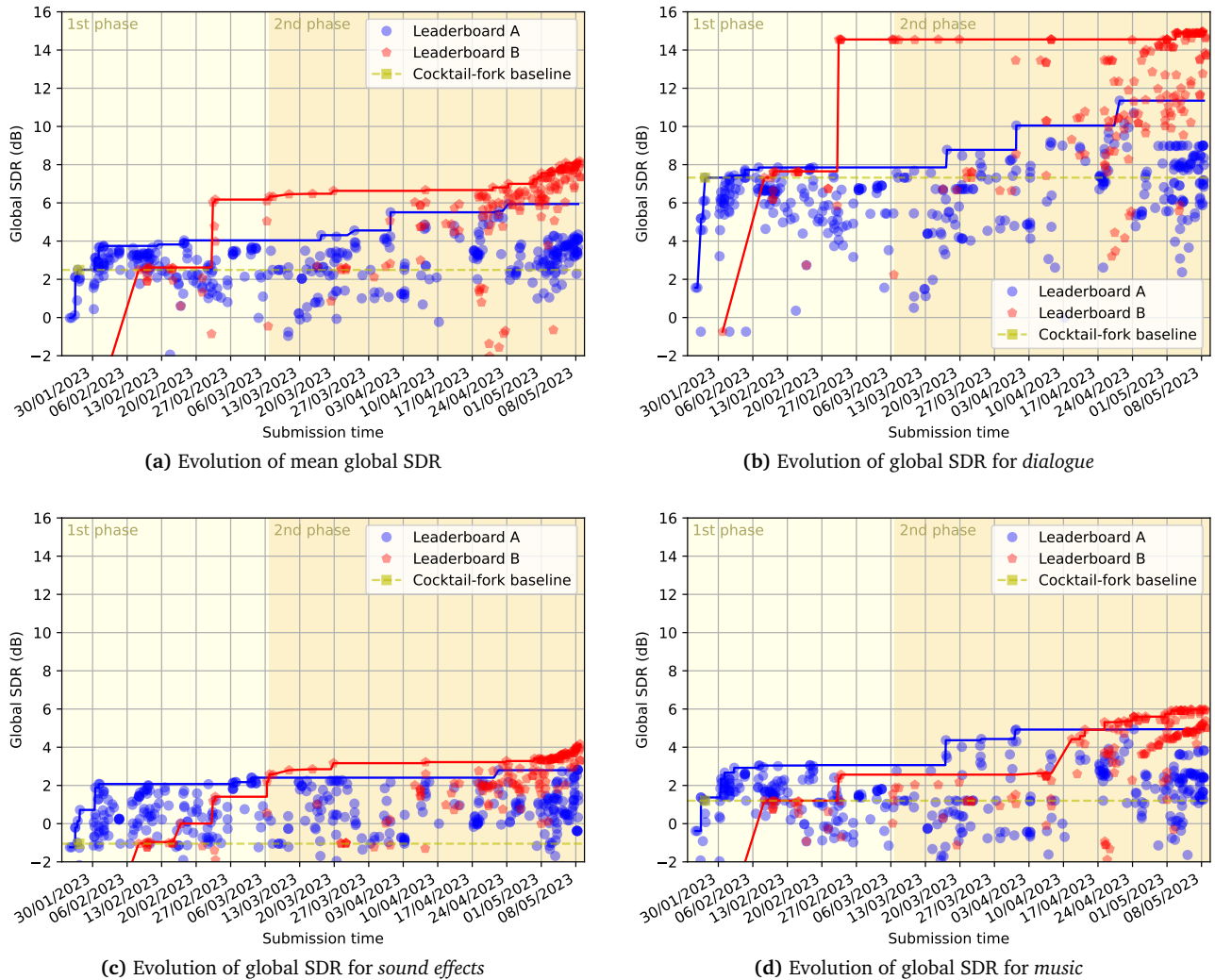
(a) Evolution of mean global SDR

(b) Evolution of global SDR for *dialogue*

(c) Evolution of global SDR for *sound effects*

(d) Evolution of global SDR for *music*

Figure 2: Performance of submissions on full CDXDB23 over time.

In order to gain more insight into the benefit of additional data, we show in Fig. 4 the performance of the winning submissions on both leaderboards in comparison to the cocktail-fork baseline. Please note that there is only a single clip for movie "000" and, hence, the box plot collapses to a horizontal line. Notably, the most significant disparities between the models trained on DnR and the winning entry in Leaderboard B are observed in animation movies ("002", "006") and action movies ("003", "008").

After the conclusion of the challenge, we contacted the top three teams in each leaderboard and invited them to contribute to this manuscript with a description of their approaches. In the following, the teams accepting our invitation present their submissions and discuss them. For the team *subatomicseer*, which ranked 3rd in Leaderboard A, we refer the interested reader to (Fabbro et al., 2023) where the team explains in detail their approach.

4.1 Team *JusperLee* (Kai Li, Yi Luo, Jianwei Yu, Rongzhi Gu)

Final ranking: Leaderboard A: —, Leaderboard B: 1st

4.1.1 Dataset

We used the public Divide and Remaster (DnR) (Petermann et al., 2022) dataset, the public deep noise suppression (DNS) dataset (Dubey et al., 2022), the public MUSDB18-HQ dataset (Rafii et al., 2019), and some extra internal data for system training. The extra internal speech data include 150 hours of data used for a text-to-speech task, the extra internal sound effect data include 10 hours of cinematic sound effects data, and the extra internal music data include 100 hours of cinematic background music data.

One important step in our data preprocessing pipeline was that we found that the effects and music signals in both the DnR dataset and our internal dataset may contain human voice. We thus used a music source separation (MSS) model to preprocess all the effect and music signals to remove the "speech" or "vo-
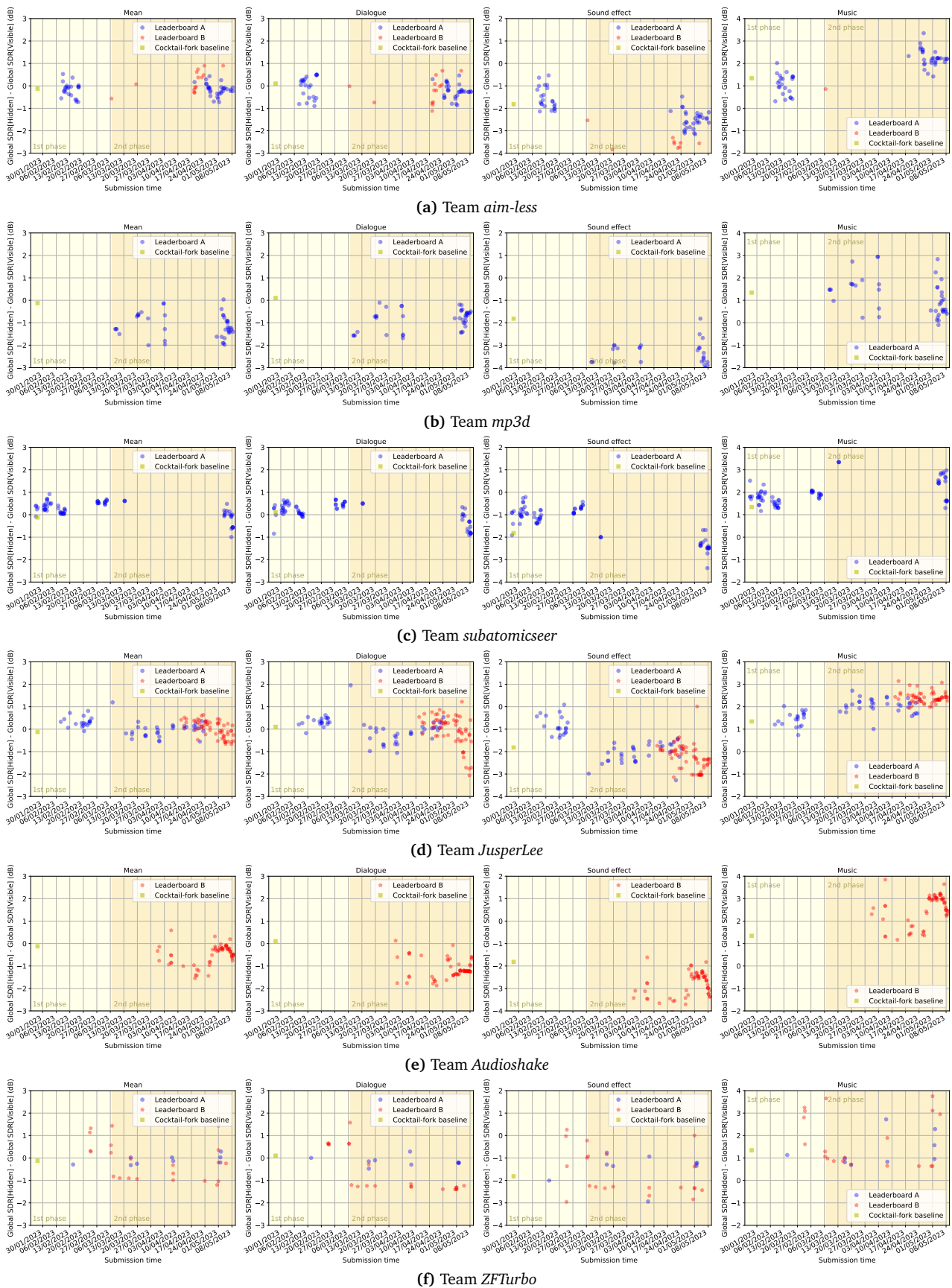
(a) Team *aim-less*

(b) Team *mp3d*

(c) Team *subatomicseer*

(d) Team *JusperLee*

(e) Team *Audioshake*

(f) Team *ZFTurbo*

Figure 3: Analysis of overfitting of global SDR. *y*-axis shows difference between global SDR on hidden test set and global SDR displayed to the participants (trajectories with negative slope indicate overfitting).
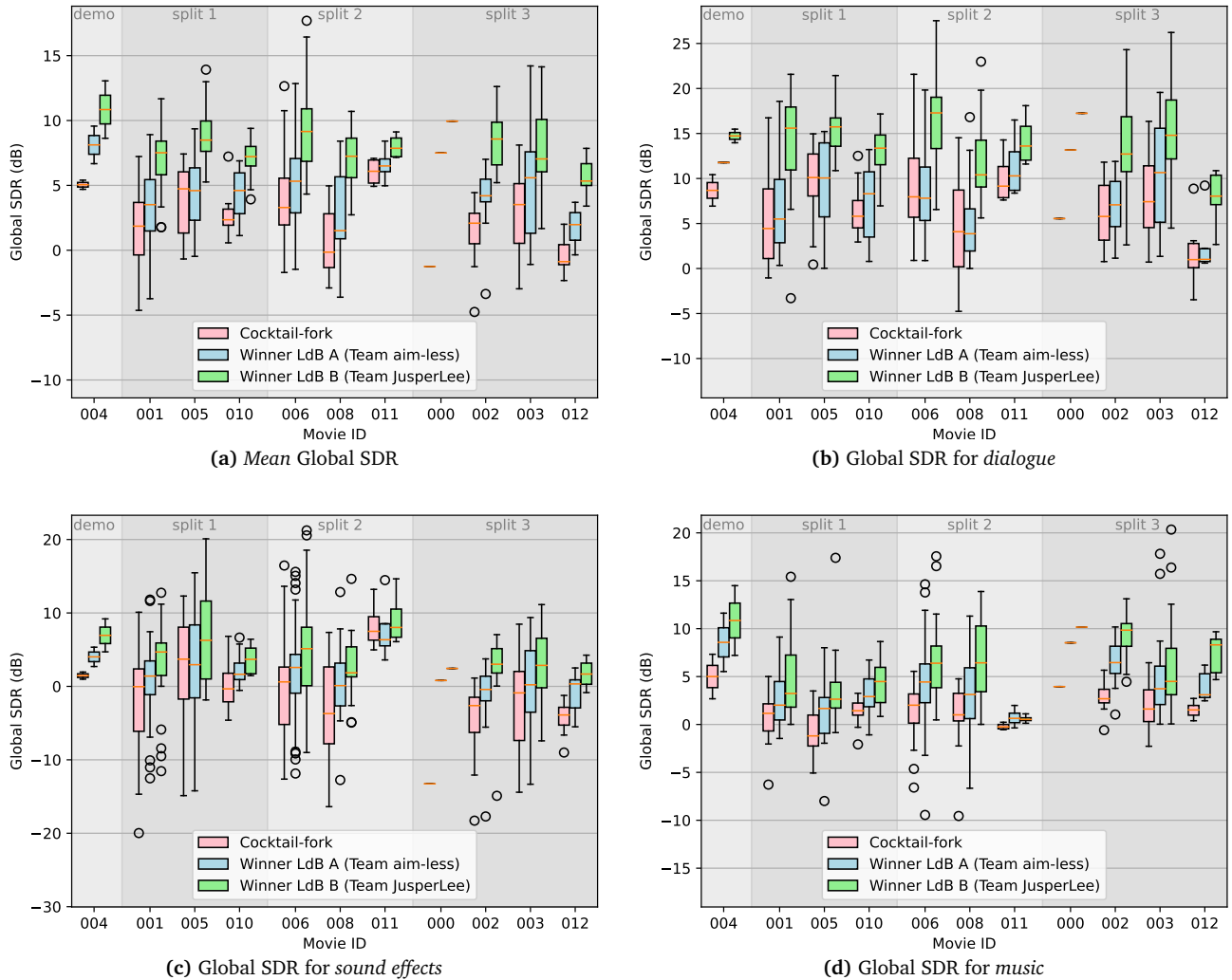
Figure 4: Comparison of the cocktail-fork baseline with winning submissions on both leaderboards for individual movies. For movie "000", we only have one clip and, hence, the box plot collapses to a horizontal line. Circles represent outliers that are outside the whiskers of the boxplot.

cal" signals from them. We found that doing this significantly improved the systems' performance compared to directly using the original signals for training.

4.1.2 Methods

a) On-the-fly Data Mixing – We performed on-the-fly data mixing during training to increase the variety of the training data mixtures. For each mixture utterance, we randomly sampled $0 - 1$ speech or vocal signals (we also treated s vocal signal as a form of dialogue signal in our setting), $0 - 2$ music signals, and $0 - 3$ effect signals, and rescaled each of them by a random energy gain of $[-10, 10]$ dB. We truncated the signals to be 3-seconds long and then added them up to form the mixture. The sum of individual music and effects signals were set as the training targets for the two tracks, respectively.

b) Model Design – Our system consists of three independent models, one for each of the dialog, effects, and music sources. All models share the same architecture, which is the band-split RNN (BSRNN) architecture we proposed for the MSS task (Luo and Yu, 2023). For the dialogue track, we directly use a BSRNN model trained for the music source separation task instead of the CDX task, as we eventually found that using an MSS model trained on music-only data that extracts the vocal track from the accompaniment can lead to better SDR score on the hidden test set than a speech extraction model trained on speech data (please see the discussion section for more on this observation). For the effects source and the music source, we used two separate BSRNN models trained on the aforementioned dataset, while we used the MSS model to first subtract the separated dialogue signal from the mixture to create a pseudo music-effects-only mixture, and then trained the two models on this mixture to perform a slightly simpler separation task. We found that this could lead to better performance than training the two

models on mixtures containing dialogue data, and also better than training on mixtures without speech or vocal signals.

We used the standard BSRNN architecture, for which we do not include a detailed description here for the sake of brevity. The band-split scheme we used for all models was identical to the one we proposed in the original paper (Luo and Yu, 2023). The number of sequence and band modeling modules in the effects and music models were 8 and 12, respectively, and the feature dimension N was set to 64 and 128, respectively.

c) Training Configurations – All models were trained with the Adam optimizer (Kingma and Ba, 2014) with an initial learning rate of 0.001. We used 8 GPUs for each model with a per-GPU batch size of 2. Each training epoch contained 10k iterations, and the learning rate was decayed by 0.98 every two epochs. We did not apply early stopping as the evaluation was done on the hidden test set, and we submitted the latest model to the grading system every day to find the best model.

4.1.3 Results and Discussions

Our system achieved #1 on the Final Leaderboard B in the CDX challenge. Comparing with other top-ranking systems, our system performed significantly better on the music source and on par or slightly worse on the two other sources, and the overall improvement mainly came from the gain from the music source.

To better understand the effect of our vocal-removal preprocessing on DnR, we did an ablation study where we trained two BSRNN models: one using the original DnR dataset, and the other using DnR after applying vocal-removal preprocessing to the music and sound effect sources. Both models were configured identically and their performance was evaluated on CDXDB23 using the AIcrowd evaluation system. Compared to the model trained on the original DnR dataset, the one trained on the vocal-removed DnR dataset achieved 1.32 dB overall SDR improvement on the challenge's test set. This confirmed our hypothesis and proved that vocal-removal for the music and sound effect class during training with DnR is an important step in our pipeline.

Another interesting observation we had was about the dialogue source – we initially tried to treat the "dialogue separation" task as a "speech enhancement" task which aims at removing any non-speech components out of the mixture, and we trained systems based on both our speech enhancement system which ranked 3rd in the 5th DNS challenge (Yu et al., 2022, 2023) and our MSS system (Luo and Yu, 2023) with the extra cinematic data. We perceptually evaluated the systems' performance on internal movie data and found the quality of their outputs satisfying. However, all model weights trained with this fashion could not achieve 13 dB SDR on the hidden test set, no matter

Table 3: Comparison of 2-stem and 3-stem HT demucs models trained on DnR and evaluated on CDXDB23 (Team *ZFTurbo*)

Model	Global SDR (dB)			
	Mean	Dialogue	Effects	Music
HT demucs trained on 2-stems mix	7.560	14.532	3.355	4.794
HT demucs trained on 3-stems mix	6.692	14.530	3.277	2.269
Ensemble of 2- and 3-stems HT demucs	7.630	14.734	3.323	4.834

how we adjusted the training pipeline or the model design. Later we tried to directly submit the original MSS system trained on music-only data (MUSDB18-HQ and another internal music dataset), and the performance of the dialogue source on Leaderboard B suddenly reached 15 dB SDR. One possible explanation is that there might still exist non-speech human sounds that are categorized as noise by the speech enhancement system but identified as vocals by the music separation system, possibly due to the differences in the training data as well as the data mixing strategies used during training.

4.2 Team *ZFTurbo* (Roman Solovyev, Alexander Stempkovskiy, Tatiana Habruseva)

Final ranking: Leaderboard A: —, Leaderboard B: 2nd

4.2.1 Approach

Our approach is based on an ensemble of models suited best for a particular stem. As we noticed that dialogue can be extracted with high quality by a vocals model that was trained originally for music separation, we separate dialogue by this model first, and then apply a model trained on the DnR dataset to the remaining part (music and sound effects). The source code is publicly available on GitHub[12].

To compare different models, we developed new benchmarks and leaderboards for sound demixing (Solovyev et al., 2023)[13]. It can be seen from this leaderboard that various *hybrid transformer demucs* (HT demucs) (Rouard et al., 2023) models dominate all stems except for vocals separation. Models based on the MDX algorithm (Kim et al., 2021) are the best for separating the vocals. Therefore, ensembles of different models used for vocal and non-vocal stems are expected to provide the top overall performance.

To separate the dialogue, we used a combination of three pre-trained vocals models: UVR-MDX1[14] and UVR-MDX2[15] from the Ultimate Vocal Remover project[16], and HT demucs (finetuned)[17]. The vocals were separated independently by all of these models and the results were combined with weights:

$$\hat{s}_{DX,1} = \text{UVR-MDX1}(\mathbf{x}, \text{overlap}=0.6),$$
$$\hat{s}_{DX,2} = \text{UVR-MDX2}(\mathbf{x}, \text{overlap}=0.6),$$
$$\hat{s}_{DX,3} = \text{HT-demucs}(\mathbf{x}, \text{'demucs\_ft'}, \text{shifts}=1, \text{overlap}=0.6).$$

We tried different weights for the DX ensemble and

Table 4: Comparison of single model HT demucs with final ensemble model (Team *ZFTurbo*).

Model	Global SDR on val1 (dB)				Global SDR on val2 (dB)				Global SDR on CDXDB23 (dB)			
	Mean	Dialogue	Effects	Music	Mean	Dialogue	Effects	Music	Mean	Dialogue	Effects	Music
HT demucs (single)	6.387	13.887	2.781	2.494	9.634	14.151	7.740	7.012	2.602	6.650	0.648	0.507
CDX23 best ensemble model	8.922	14.927	3.780	8.060	7.585	9.949	6.377	6.429	7.630	14.734	3.323	4.834

optimized them by considering two datasets that we created: "Multisong MVSep" and "Synth MVSep" as detailed in (Solovyev et al., 2023). The weights 10, 4, and 2 for UVR-MDX1, UVR-MDX2, and HT demucs, respectively, produced the best results on these two datasets. Interestingly, we observed that the models with the best SDR for vocals extraction were also the best for dialogue and, hence, we can rely on the results from (Solovyev et al., 2023).

After obtaining the high-quality dialogue part, we can subtract it from the original track to obtain the non-dialogue part. To separate it into two remaining stems, we trained two versions of the HT demucs model (Rouard et al., 2023) on the DnR dataset. The first HT demucs model was trained using the standard protocol for all three stems, while the second HT demucs model was trained only on two stems: sound effects and music, excluding dialogue. Table 3 shows the global SDR on CDXDB23 and we can observe that the 2-stem model yields better scores. Especially music benefits from the simplified training mixtures as it improves by 2.5 dB. Interestingly, blending both models is still beneficial as can also be seen from Table 3, where we blended four checkpoints from the 2-stem HT demucs training with seven checkpoints of the 3-stem HT demucs training, giving each the same weight. Please note that we also updated the vocals model in this submission and, hence, there is also a slight improvement for vocals if compared to the individual models. Consequently, for the final submission, we used several checkpoints of each of the 2-stem and 3-stem models to average predictions and obtain better generalization.

4.2.2 Results

Table 4 shows the results of the ablation study with and without a separate dialogue removal. We used two validation sets: *val*1 – validation on two tracks provided by the organizers, and *val*2 – a subset of 20 random tracks from the DnR test set. The single HT demucs model showed promising results on both validation sets (see Table 4). However, the model performance was poor on CDXDB23 and the best results can only be obtained with our ensemble model which first extracts dialogue with a vocals model from music separation and then employs an HT demucs for sound effects and music separation. *val*1 correlated better with CDXDB23 dataset than *val*2, which is based on the DnR dataset. Still, metrics on *val*1 did not strongly correlate with the final results, presumably due to the tiny size of the demo set.

4.2.3 Discussion

During the competition, we noticed that the DnR dataset contains vocals in the music part sometimes. Our SDR for vocals on the leaderboard is very high, but our vocals model extracts all vocals from audio. Based on this, we made a conclusion that *music* in the competition dataset most likely never contains vocals.

4.3 Team *mp3d* (Mikhail Sukhovei)

Final ranking: Leaderboard A: 2ⁿᵈ, Leaderboard B: 5ᵗʰ

4.3.1 Approach

Since the training data for Leaderboard A was restricted to only the DnR dataset, we focused on identifying the shortcomings of the baseline multi-resolution crossnet (MRX) (Petermann et al., 2022) model and improving it. We implemented a modification of this model called conditional multi-resolution crossnet (MRX-C) (Petermann et al., 2023). The essence of this modification is to train an additional CRNN model that predicts source activity labels. The output of the MRX model (which estimates music, dialogue, and sound effects) is converted into a mel-spectrogram and concatenated with the mel-spectrogram of the original audio to form a $(4, n_{mels}, n_{freq})$ tensor, which is then fed into the CRNN.

To further improve our solution, we analyzed the effect of the Wiener filter on the final score as well as the influence of post-processing source scaling on the final SDR score.

4.3.2 Results

During the competition, we observed that validation on DnR is significantly different from metrics obtained on CDXDB23. This difference arises due to the dependence of the MRX and MRX-C model performance on the volume of the input signal. For testing on DnR, the optimal value was found to be −27 LUFS, which yielded the maximum SDR value. To obtain the optimal input volume for real world data, we propose a *Realistic Evaluation Dataset* (RED) consisting of 26 stereo audio tracks of 20 seconds each. These audio samples were manually compiled, with an average sample volume of approximately −15 LUFS and a sampling rate of 44.1 kHz. Additionally, each audio file includes separate tracks for dialogue, music, and effects, all with the same duration and with average sample volumes of −24.4 ± 4.5 LUFS for dialogues, −18.8 ± 0.5 LUFS for effects and −18.4 ± 5.0 LUFS for music. All the original audio files are sourced from the open archive.org platform. The RED dataset was used only to select the
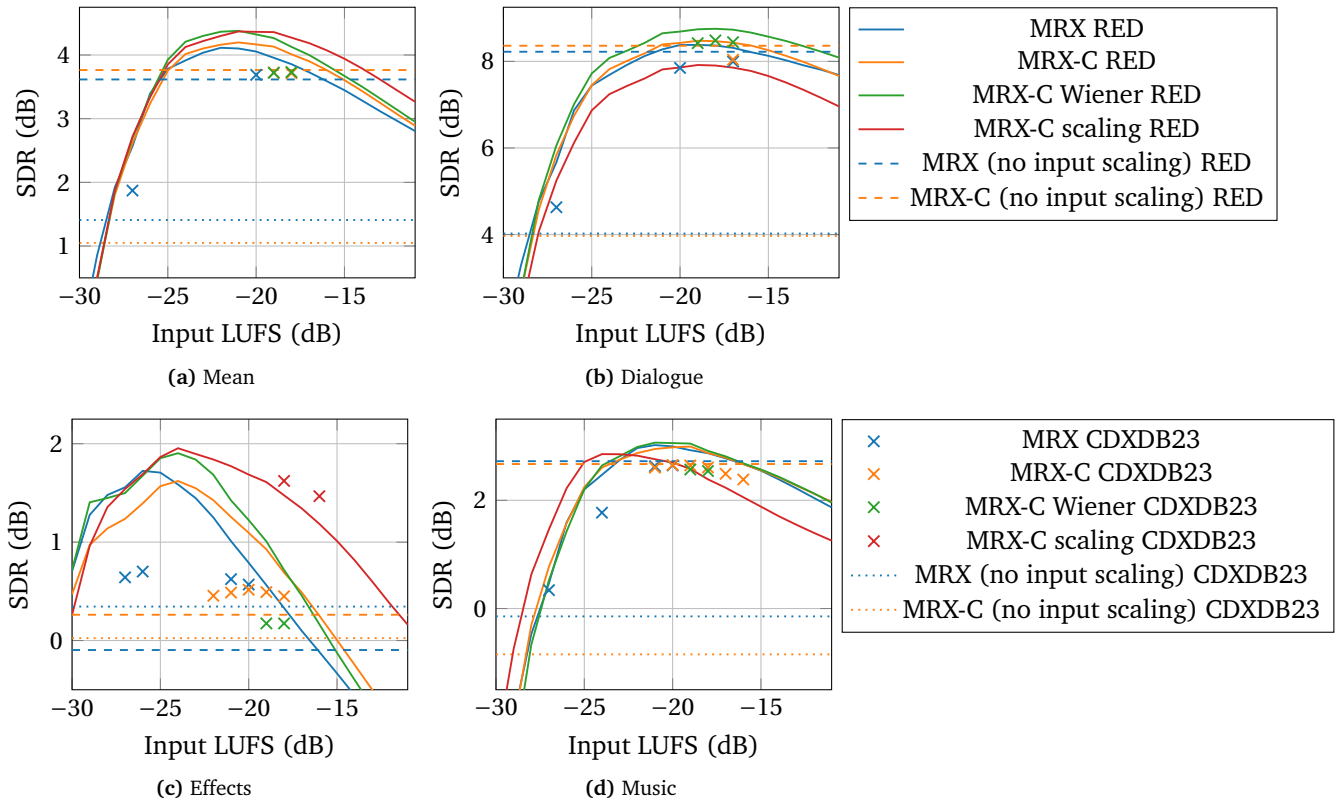
Figure 5: SDR dependencies on the input volume in LUFS for music, dialog, and effects. A solid line shows SDR values on RED, crosses mark on CDXDB23. Horizontal dashed and dotted lines show SDR for models without converting the volume of the input signal. The MRX model is blue, MRX-C is orange, MRX-C with a Wiener filter is green, and MRX-C with post-processing scaling is red. In the case of testing MRX-C scaling on the CDXDB23, the SDR values are only available for effects (Team *mp3d*).

optimal value of the input signal volume of the model.

As shown in Fig. 5, the optimal volume value differed significantly from that of DnR, both in the case of RED and CDXDB23. Furthermore, the SDR metric on RED was more consistent with CDXDB23. After separating the sources, they were brought back to the original volume and a Wiener filter and post-processing scaling were applied. Post-processing scaling involved multiplying the estimated sources by a factor of $\frac{1}{\alpha}$, where $\alpha = \frac{\sum_n \mathbf{x}(n)^T \hat{\mathbf{s}}(n)}{10^{-7} + \sum_n \|\hat{\mathbf{s}}(n)\|^2}$, $\mathbf{x}(n)$ is the mixture and $\hat{\mathbf{s}}(n)$ is an estimated source.

Table 5 summarizes the SDR metrics on RED for the baseline MRX solution, MRX-C, MRX-C with Wiener filter, and MRX-C with source scaling. As shown in the table, the MRX-C model yields a 0.1 dB improvement in dialogue and a 0.1 dB decrease in effects. The Wiener filter yields a 0.1 dB improvement in music, a 0.3 dB improvement in dialogue, and a 0.3 dB improvement in effects. Post-processing scaling results in a 0.2 dB decrease in music, a 0.6 dB decrease in dialogue, and a 0.3 dB increase in effects. Our final solution involved applying the Wiener filter only to the dialogue and post-processing scaling only to the effects.

4.3.3 Discussion
In the future, we plan to study the effect of adding additional activity labels contained in the DnR data on the MRX-C model accuracy. Additionally, to use the model on real data, we need to make the result of source separation independent of the input volume. Unlike the approaches of other teams in this competition, we focused on training a single model rather than an ensemble of models. A key aspect of our solution is to apply the normalization of the mixture at the input of the model.

5. Distribution Mismatch between DnR and CDXDB23

In the following, we will analyze the distribution mismatch between DnR and CDXDB23 that was noticed by the participants in Sec. 4. This will give us insight into the recording and production differences as well as allow us to train two improved cocktail-fork models with an adjusted version of DnR.

5.1 Difference in Signal Statistics
First, we will compare the signal characteristics between DnR and CDXDB23. Our focus will be on loudness, equalization, stereo panning, and dynamic range

Table 5: SDR values obtained during testing on RED for MRX, MRX-C, MRX-C with Wiener filter, and MRX-C with scaling. SDR values from the table are maximum possible values from all input volumes (Team *mp3d*).

Model	Global SDR (dB)			
	Mean	Dialogue	Effect	Music
MRX	4.38	8.38	1.72	3.02
MRX-C	4.36	8.48	1.62	2.99
MRX-C Wiener	4.57	8.75	1.90	3.07
MRX-C scaling	4.24	7.92	1.95	2.85

compression as they are key elements in audio mixing (Martínez-Ramírez et al., 2022).[18]

Loudness—We measured the loudness for each audio clip in both datasets using ITU-R BS. 1770-4 (1770-4, 2015) with the help of `pyloudnorm` (Steinmetz and Reiss, 2021). The average loudness values are shown in Table 6 and the histograms can be found in Fig. 6. We can observe that CDXDB23 utilizes the full range of loudness for all three classes whereas DnR has a more limited range. On average, DnR is 4 LUFS louder than CDXDB23. Notably, CDXDB23 uses the same loudness level for effects and music which is 5 LUFS lower than dialogue. This balance is likely due to a post-production step which was not considered in DnR.

Equalization—To assess equalization differences, we normalized each waveform to −24 LUFS and calculated the magnitude STFT spectrogram using a Hann window of 4096 samples with 75% overlap. The average equalization curves are displayed in Fig. 7. It shows that CDXDB23 generally has a faster decay at low and high frequencies but more energy in the mid-frequency range compared to DnR. We attribute this to the use of parametric EQs containing low and high shelf filters in the post production process for CDXDB23, while DnR consisted mostly of web content, which likely lacked professional post production.

Amplitude panning—We calculated the Stereo Panning Spectrum as outlined in (Tzanetakis et al., 2007; Avendano, 2003) and as used in (Martínez-Ramírez et al., 2022). Using the magnitude spectrogram from the earlier equalization analysis, we computed

$$\Psi(f) = 2 \frac{X_L(f) X_R(f)}{X_L(f)^2 + X_R(f)^2}, \tag{5a}$$

$$\begin{aligned} \Delta(f) &= \text{sign}\left(\Psi_L(f) - \Psi_R(f)\right) \\ &= \text{sign}\left(\frac{X_L(f) X_R(f)}{X_L(f)^2} - \frac{X_L(f) X_R(f)}{X_R(f)^2}\right), \end{aligned} \tag{5b}$$

where $X_{L/R}(f)$ denotes the left/right channel magnitude spectrogram. $\Psi(f)$ measures whether frequencies are panned, regardless of direction, while $\Delta(f)$ measures which direction frequencies are panned to. Figures 8 and 9 show the average values for both datasets.

It is important to note that DnR is monaural, which results in a horizontal line for $\Psi(f)$ and $\Delta(f)$. From these figures, we see that in the audio mixing process, dialogue is typically centered, while effects are more often panned to one side. Music shows the most varied panning. Figure 9 indicates that there is no specific preferred direction for panning in the datasets.

Dynamic range compression (DRC)—Lastly, we analyzed the used DRC by calculating the average peak value as DRC usually alters the transients. We started by normalizing the loudness of the audio waveform to −24 LUFS. Then, we used the high frequency content (HFC) method for onset detection, as described by (Masri, 1996; Brossier et al., 2019) and as implemented in (Martínez-Ramírez et al., 2022). For each audio clip, we calculated the average peak level $\mathscr{P}_\mu$. This measure helps us understand the extent of DRC applied; larger $\mathscr{P}_\mu$ values indicate less compression since the peaks are more pronounced at the same loudness level. From the data in Table 6, we see that CDXDB23 exhibits more uniform compression compared to DnR. Notably, in DnR, effects are less compressed compared to dialogue and music. This contrast in compression is not observed in CDXDB23, where compression is more consistently applied across all three classes due to being professionally produced.

5.2 Improving the Cocktail–Fork Baseline

Using our understanding of the distribution differences from the previous section, we aim to enhance the cocktail-fork model from Sec. 3.3. We created two updated versions of DnR, modifying either the average loudness or the average equalization for each audio source. For the loudness adjustment, we changed the loudness of each source stem by a specific amount. For example, we altered the loudness of each dialogue stem by 4 LUFS (from −28.4 to −24.4 LUFS). This adjustment was made so that the average loudness of the modified DnR matches that of CDXDB23. Regarding equalization, we designed a 101-tap FIR filter for each audio source. The magnitude response of this filter is the square root of the difference between the average equalization of CDXDB23 and DnR. We then applied this filter to each stem using forward-backward filtering (`filtfilt`), as mentioned in (Martínez-Ramírez et al., 2022). This process modifies the amplitude without altering the phase of the audio.

Table 7 shows the improvements to the cocktail-fork model. Please note that an additional feature for mixture loudness normalization was introduced with version 1.1, setting the mixture to −27 LUFS. We will discuss the results without considering this normalization, although similar trends are observed with it. The results in Table 7 indicate that adjusting the loudness is particularly effective. The mean SDR improved from −0.1 dB to 1.3 dB, primarily due to enhanced performance in dialogue. Adjusting the equalization also

Table 6: Loudness and Dynamic Range Compression (DRC) statistics for DnR and CDXDB23.

	Divide and Remaster (DnR)			CDXDB23		
	Dialogue	Effect	Music	Dialogue	Effect	Music
Loudness (LUFS)	-24.4 ± 1.3	-29.7 ± 1.9	-31.4 ± 1.8	-28.4 ± 3.1	-33.9 ± 8.0	-33.6 ± 7.1
DRC (dB)	-10.7 ± 0.9	-5.1 ± 2.4	-12.6 ± 1.4	-11.4 ± 1.3	-10.6 ± 3.7	-11.2 ± 2.3

(a) Dialogue   **(b)** Effect   **(c)** Music
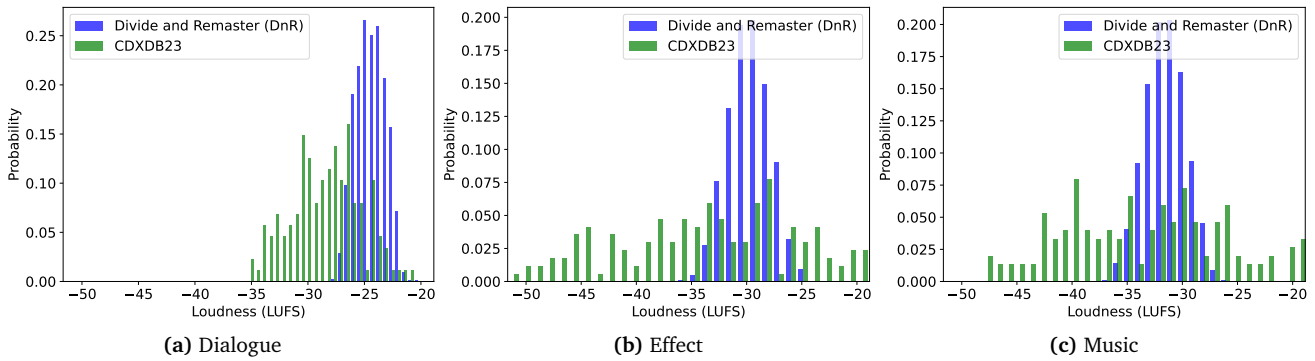
Figure 6: Comparison of loudness between DnR and CDXDB23.

(a) Dialogue   **(b)** Effect   **(c)** Music
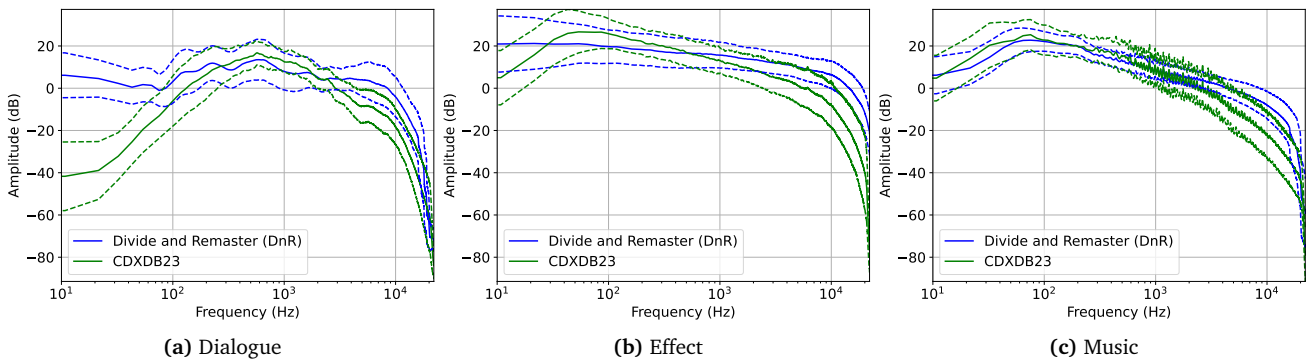
Figure 7: Comparison of average equalization between DnR and CDXDB23. Dashed curves give one standard deviation above/below average.

(a) Dialogue   **(b)** Effect   **(c)** Music
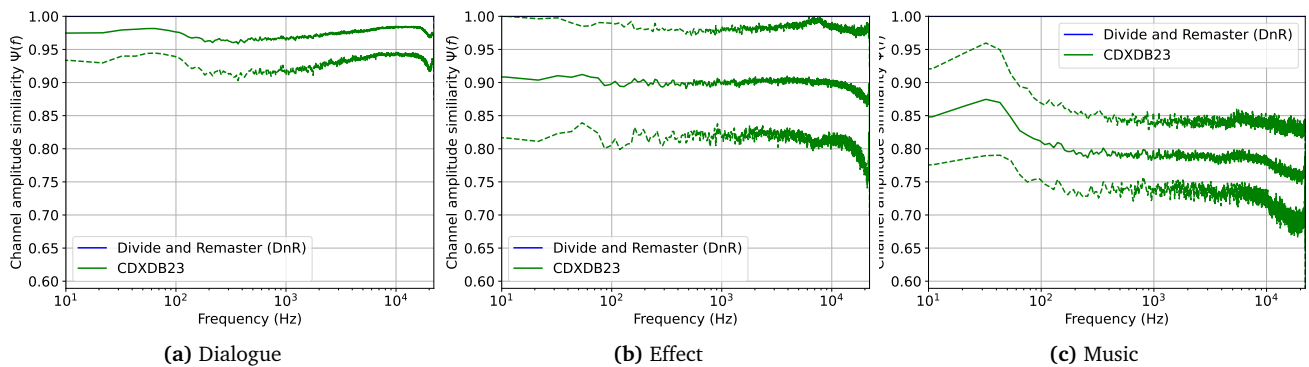
Figure 8: Comparison of average amplitude panning between DnR and CDXDB23. Channel amplitude similarity $\Psi(f)$ can take values $0 \leq \Psi(f) \leq 1$ where $\Psi(f) = 1$ refers to panning frequency f to the center whereas $\Psi(f) < 1$ denotes a panning to either side. Dashed curves give one standard deviation above/below average. Please note that DnR is monaural and, hence, $\Psi(f)$ collapses to a horizontal line at $\Psi(f) = 1$.
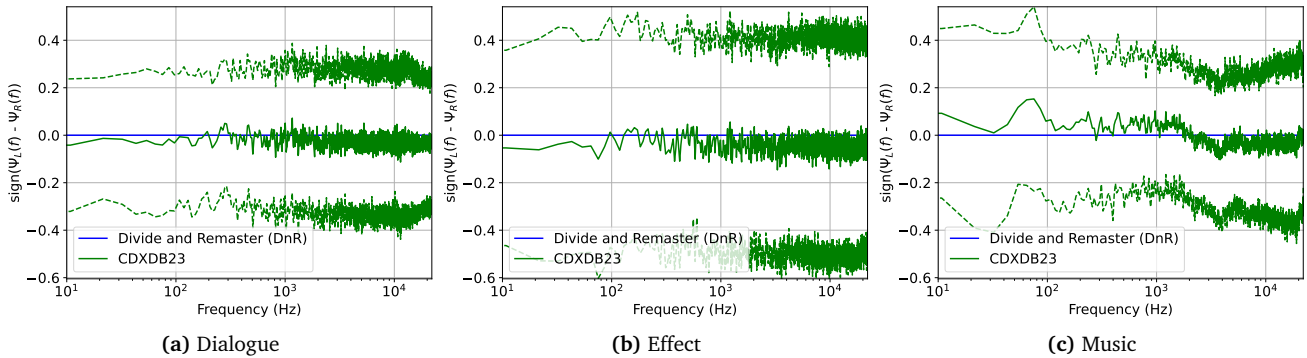
(a) Dialogue   **(b)** Effect   **(c)** Music

Figure 9: Comparison of average amplitude panning between DnR and CDXDB23. $\Delta(f) = \text{sign}\left(\Psi_L(f) - \Psi_R(f)\right)$ denotes the panning direction where $\Delta(f) < 0$ refers to panning to the left and $\Delta(f) > 0$ to a panning to the right. Dashed curves give one standard deviation above/below average. Please note that DnR is monaural and, hence, $\Delta(f)$ collapses to a horizontal line at $\Delta(f) = 0$.

showed benefits, with an overall improvement of 0.3 dB. Here, both dialogue and effects showed improvement, but there was a slight decrease of 0.2 dB in music. We believe this decrease in music is linked to the fluctuation in the equalization curve shown in Fig. 7, which correlates with the frequencies of musical notes. This fluctuation could inadvertently act as a marker on the music stems, making them easier for the model to separate, leading to a slight decline in music performance. Both trained models are available on the cocktail-fork GitHub[6].

In summary, by aligning the DnR dataset more closely with a more realistic dataset like CDXDB23, we significantly enhanced the performance of the model. This approach presents a promising avenue for future research in this field. Besides using mono-to-stereo augmentation for stereo panning and compressors to adjust the DRC, also combining all of them should be considered to close the distribution mismatch as much as possible. Using the reported statistics in Sec. 5.1 will help to choose realistic parameters for the data augmentation.

6. Summary and Outlook

The CDX track of SDX'23 has provided valuable insights into the current state-of-the-art in cinematic audio separation and has highlighted areas for future research and development.

Looking at the results for Leaderboard A, we can observe that models suffered from the constraint of only being allowed to utilize the DnR dataset. This caused a "simulation-to-reality" gap where models were trained on simulated data, but evaluated on real-world data (CDXDB23). In particular, the following challenges were identified by the participants:

- The DnR dataset contains sometimes vocals within the music, leading to confusion during model training. Preemptively removing these vocals prior to training was found to enhance model performance.
- The dialogue data in the DnR dataset, being read

speech, lacks emotional speech elements such as shouting, as well as other human sounds like breathing or humming. This absence posed a challenge for the models.
- Lastly, a mismatch in loudness was observed between the training and evaluation data. If not accounted for, this mismatch could lead to suboptimal model performance as we also saw in Sec. 5.

Hosting a competition like SDX'23 allows to identify and address these issues, thereby contributing significantly to the field. Moreover, the allowance for participants to utilize additional data, as seen in Leaderboard B, proved beneficial. Particularly, an improvement of approximately 6 dB was observed for dialogue when comparing the results of Leaderboard B to those of Leaderboard A.

Another interesting observation was the successful application of cascaded approaches, which initially filtered out dialogue. The effectiveness of this strategy can likely be attributed to two factors. First, the substantial amount of available data for vocals and the existence of highly efficient models, honed through research in the field of music separation (Stöter et al., 2018; Mitsufuji et al., 2022; Fabbro et al., 2023), provide a strong foundation for dialogue extraction. Second, vocals, which include sounds like breathing, bear a close resemblance to dialogue, thereby facilitating the extraction of dialogue from movies. As cinematic separation is a relatively young field, further advances are required, particularly in the extraction of sound effects and music.

Looking forward to the next challenge, we anticipate further advances in the field. Exploring uncharted areas such as mono-to-stereo augmentations is interesting and will have a positive impact on performance. We also aim to encourage participants to develop models that are robust to variations in the input data, such as loudness invariance. The goal remains to push the boundaries of what is possible in cinematic audio sep-

Table 7: Results on CDXDB23 for training the cocktail-fork model with adjusted DnR versions where we matched either the average loudness or the average equalization from CDXDB23. "input norm" refers to the loudness normalization to -27 LUFS introduced with version 1.1 of the cocktail-fork model.

Training Dataset	Global SDR w/o input norm (dB)				Global SDR w/ input norm (dB)			
	Mean	Dialogue	Effect	Music	Mean	Dialogue	Effect	Music
DnR	−0.104	4.108	−2.018	−2.401	0.325	4.662	−1.979	−1.707
DnR w/ adapted loudness	1.287	6.535	−1.506	−1.168	1.539	6.727	−1.278	−0.832
DnR w/ adapted equalization	0.176	4.621	−1.470	−2.623	0.544	4.922	−1.212	−2.078

aration and to continue fostering innovation in this exciting field. This first edition of the challenge has demonstrated the utilization of models, data, and concepts from music separation to enhance cinematic separation. We believe that this represents an initial stage, and look forward to future development of more specialized ideas and approaches exploiting also the signal statistics presented in Sec. 5.1.

Notes

[1] Frequently, sound effects for a movie must be created when no suitable option is at hand as they need to provide "sound" for new objects. A notable illustration is the crafting of the Godzilla Roar (Sound Effects Wiki, 2024), where a contrabass string was rubbed with gloves soaked in pine tar to produce the distinctive sound used as Godzilla's roar. Numerous instances exist where sound effects had to be invented for movies, contributing significantly to the overall diversity within the sound effects category.

[2] `https://www.aicrowd.com/challenges/sound-demixing-challenge-2023`

[3] There was a related discussion about "music" vs. "sound effects" and "dialogue" during the competition which can be found here: https://discourse.aicrowd.com/t/class-label-definition-of-sound-effects-and-music/8490.

[4] Given the absence of universally accepted rules for multi-channel spatialization, the DnR dataset does not incorporate it, leading to a disparity with the hidden test set (CDXDB23). As a result, participants are compelled either to develop suitable data augmentation during training or to employ channel-wise processing which can be combined with a multi-channel Wiener filter during inference as already proposed in (Petermann et al., 2022).

[5] The short movie "Kilian's Game" and related content can be accessed via the following links. The full movie is available at `https://www.youtube.com/watch?v=PxKB8NKQj3U`, and a behind-the-scenes look can be found at `https://www.youtube.com/watch?v=NdUuiwmHsKU`. Participants could also view the output of their system on demo clips, available at `https://www.youtube.com/embed/PxKB8NKQj3U?start=39&end=62` and `https://www.youtube.com/embed/PxKB8NKQj3U?start=32&end=39`.

[6] `https://github.com/merlresearch/cocktail-fork-separation`

[7] `https://gitlab.aicrowd.com/aicrowd/challenges/sound-demixing-challenge-2023/sdx-2023-cinematic-sound-demixing-starter-kit`

[8] `https://gitlab.aicrowd.com/yoyololicon/cdx-submissions`

[9] `https://gitlab.aicrowd.com/mikhail_sukhovey/mrxc`, `https://gitlab.aicrowd.com/mikhail_sukhovey/sdx-2023-cinematic-sound-demixing-starter-kit/tree/3695fd3e2cf85cddad6446decf276fc8dc46d27d`

[10] `https://github.com/naba89/iSeparate-SDX`

[11] `https://github.com/ZFTurbo/MVSEP-CDX23-Cinematic-Sound-Demixing`, `https://drive.google.com/file/d/1AQt2uNMdTI_aGcyFEjKGe-GCetKK3xo0/view?usp=sharing`

[12] `https://github.com/ZFTurbo/MVSEP-CDX23-Cinematic-Sound-Demixing`

[13] `https://mvsep.com/quality_checker/`.

[14] Checkpoint "Kim_Vocal_1.onnx" available at `https://github.com/TRvlvr/model_repo/releases/download/all_public_uvr_models/Kim_Vocal_1.onnx`

[15] Checkpoint "UVR–MDX–NET–Inst_HQ_2.onnx" available at `https://github.com/TRvlvr/model_repo/releases/download/all_public_uvr_models/UVR-MDX-NET-Inst_HQ_2.onnx`

[16] `https://github.com/Anjok07/ultimatevocalremovergui`

[17] Checkpoint "htdemucs_ft" available at `https://github.com/facebookresearch/demucs`

[18] As in (Martínez-Ramírez et al., 2022), we do not consider the reverberation of the signals as blind estimation of reverberation features such as reverberation time is an open research area which has mostly been worked on only for speech signals.

Acknowledgements

References

1770-4, I.-R. B. (2015). Algorithms to measure audio programme loudness and true-peak audio level. *International Telecommunications Union*.

Avendano, C. (2003). Frequency-domain source identification and manipulation in stereo mixes for enhancement, suppression and re-panning applications. In *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (IEEE Cat. No. 03TH8684)*, pages 55–58. IEEE.

Brossier, P., Tintamar, Müller, E., Philippsen, N., Seaver, T., Fritz, H., cyclopsian, Alexander, S., Williams, J., Cowgill, J., and Cruz, A. (2019). aubio/aubio: 0.4.9.

Chen, Z., Luo, Y., and Mesgarani, N. (2017). Deep attractor network for single-microphone speaker separation. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 246–250.

Defferrard, M., Benzi, K., Vandergheynst, P., and Bresson, X. (2016). FMA: A dataset for music analysis. *arXiv preprint arXiv:1612.01840*.

Dubey, H., Gopal, V., Cutler, R., Aazami, A., Matusevych, S., Braun, S., Eskimez, S. E., Thakker, M., Yoshioka, T., Gamper, H., et al. (2022). ICASSP 2022 Deep Noise Suppression Challenge. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 9271–9275.

Fabbro, G., Uhlich, S., Lai, C.-H., Choi, W., Martinez-Ramirez, M., Liao, W., Gadelha, I., Ramos, G., Hsu, E., Rodrigues, H., Stoeter, F.-R., Defossez, A., Luo, Y., Yu, J., Chakraborty, D., Mohanty, S., Solovyev, R., Stempkovskiy, A., Habruseva, T., Goswami, N., Harada, T., Kim, M., Lee, J. H., Dong, Y., Zhang, X., Liu, J., and Mitsufuji, Y. (2023). The Sound Demixing Challenge 2023 – Music demixing track. Under review.

Fonseca, E., Favory, X., Pons, J., Font, F., and Serra, X. (2021). FSD50K: An open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:829–852.

Geiger, J. T., Grosche, P., and Parodi, Y. L. (2015). Dialogue enhancement of stereo sound. In *2015 23rd European Signal Processing Conference (EUSIPCO)*, pages 869–873.

Grais, E. M., Sen, M. U., and Erdogan, H. (2014). Deep neural networks for single channel source separation. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3734–3738.

Hershey, J. R., Chen, Z., Le Roux, J., and Watanabe, S. (2016). Deep clustering: Discriminative embeddings for segmentation and separation. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 31–35.

Huang, P.-S., Chen, S. D., Smaragdis, P., and Hasegawa-Johnson, M. (2012). Singing-voice separation from monaural recordings using robust principal component analysis. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 57–60.

Kim, M., Choi, W., Chung, J., Lee, D., and Jung, S. (2021). Kuielab-mdx-net: A two-stream neural network for music demixing. *arXiv preprint arXiv:2111.12203*.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Le Roux, J., Wisdom, S., Erdogan, H., and Hershey, J. R. (2019). SDR–half-baked or well done? In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 626–630.

Luo, Y. and Yu, J. (2023). Music source separation with band-split RNN. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1893–1901.

Martínez-Ramírez, M. A., Liao, W.-H., Fabbro, G., Uhlich, S., Nagashima, C., and Mitsufuji, Y. (2022). Automatic music mixing with deep learning and out-of-domain data. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)*.

Masri, P. (1996). *Computer modelling of sound for transformation and synthesis of musical signals*. PhD thesis, University of Bristol.

Mitsufuji, Y., Fabbro, G., Uhlich, S., Stöter, F.-R., Défossez, A., Kim, M., Choi, W., Yu, C.-Y., and Cheuk, K.-W. (2022). Music Demixing Challenge 2021. *Frontiers in Signal Processing*, 1:18.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Paulus, J., Torcoli, M., Uhle, C., Herre, J., Disch, S., and Fuchs, H. (2019). Source separation for enabling dialogue enhancement in object-based broadcast with MPEG-H. *Journal of the Audio Engineering Society*, 67(7/8):510–521.

Petermann, D., Wichern, G., Subramanian, A. S., Wang, Z.-Q., and Le Roux, J. (2023). Tackling the cocktail fork problem for separation and transcription of real-world soundtracks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31.

Petermann, D., Wichern, G., Wang, Z.-Q., and Le Roux, J. (2022). The cocktail fork problem: Three-stem audio separation for real-world soundtracks. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 526–530.

Rafii, Z., Liutkus, A., Stöter, F.-R., Mimilakis, S. I., and Bittner, R. (2019). MUSDB18-HQ - an uncompressed version of musdb18.

Rouard, S., Massa, F., and Défossez, A. (2023). Hybrid transformers for music source separation. In *Proc.*

IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pages 1–5.

Sawata, R., Takahashi, N., Uhlich, S., Takahashi, S., and Mitsufuji, Y. (2023). The whole is greater than the sum of its parts: Improving DNN-based music source separation. *arXiv preprint arXiv:2305.07855*.

Sawata, R., Uhlich, S., Takahashi, S., and Mitsufuji, Y. (2021). All for one and one for all: Improving music separation by bridging networks. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 51–55.

Solovyev, R., Stempkovskiy, A., and Habruseva, T. (2023). Benchmarks and leaderboards for sound demixing tasks. *arXiv preprint arXiv:2305.07489*.

Sound Effects Wiki (2024). Godzilla roar. `https://soundeffects.fandom.com/wiki/Godzilla_Roar` [Accessed: 2024-01-15].

Steinmetz, C. J. and Reiss, J. (2021). pyloudnorm: A simple yet flexible loudness meter in python. In *Audio Engineering Society Convention 150*. Audio Engineering Society.

Stöter, F.-R., Liutkus, A., and Ito, N. (2018). The 2018 signal separation evaluation campaign. In *Proc. Intl. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)*, pages 293–305. Springer.

Stöter, F.-R., Uhlich, S., Liutkus, A., and Mitsufuji, Y. (2019). Open-unmix - a reference implementation for music source separation. *Journal of Open Source Software*, 4(41):1667.

Torcoli, M., Simon, C., Paulus, J., Straninger, D., Riedel, A., Koch, V., Wits, S., Rieger, D., Fuchs, H., Uhle, C., et al. (2021). Dialog+ in broadcasting: First field tests using deep-learning-based dialogue enhancement. *arXiv preprint arXiv:2112.09494*.

Tzanetakis, G., Jones, R., and McNally, K. (2007). Stereo panning features for classifying recording production style. In *ISMIR*, pages 441–444.

Uhle, C., Hellmuth, O., and Weigel, J. (2008). Speech enhancement of movie sound. In *Audio Engineering Society Convention 125*. Audio Engineering Society.

Uhlich, S., Giron, F., and Mitsufuji, Y. (2015). Deep neural network based instrument extraction from music. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2135–2139.

Uhlich, S., Porcu, M., Giron, F., Enenkl, M., Kemp, T., Takahashi, N., and Mitsufuji, Y. (2017). Improving music source separation based on deep neural networks through data augmentation and network blending. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 261–265.

Vincent, E., Sawada, H., Bofill, P., Makino, S., and Rosca, J. P. (2007). First stereo audio source separation evaluation campaign: Data, algorithms and results. In *Proc. Intl. Conf. on Independent Component Analysis and Signal Separation*, pages 552–559. Springer.

Watcharasupat, K. N., Wu, C.-W., Ding, Y., Orife, I., Hipple, A. J., Williams, P. A., Kramer, S., Lerch, A., and Wolcott, W. (2023). A generalized bandsplit neural network for cinematic audio source separation. *IEEE Open Journal of Signal Processing*.

Wisdom, S., Hershey, J. R., Wilson, K., Thorpe, J., Chinen, M., Patton, B., and Saurous, R. A. (2019). Differentiable consistency constraints for improved deep speech enhancement. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 900–904.

Yu, D., Kolbæk, M., Tan, Z.-H., and Jensen, J. (2017). Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 241–245.

Yu, J., Chen, H., Luo, Y., Gu, R., Li, W., and Weng, C. (2023). TSpeech-AI system description to the 5th Deep Noise Suppression (DNS) challenge. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–2.

Yu, J., Luo, Y., Chen, H., Gu, R., and Weng, C. (2022). High fidelity speech enhancement with band-split RNN. *arXiv preprint arXiv:2212.00406*.