LEVERAGING SYNTHETIC DATA FOR IMPROVING CHAMBER ENSEMBLE SEPARATION

Saurjya Sarkar, Louise Thorpe, Emmanouil Benetos, Mark Sandler

Centre for Digital Music
Queen Mary University of London
London, UK

ABSTRACT

In this work, we tackle the challenging problem of separating monophonic instrument mixtures found in chamber music from monaural recordings. This task differs from the Music Demixing Challenge where the task is to separate vocals, drums, and bass stems from mastered stereo tracks. In our task, we separate the instruments in a permutation invariant fashion such that our model is capable of separating any two monophonic instruments, including mixtures of the same instrument. This task is particularly difficult due to label ambiguity and high spectral overlap. In this paper, we present a pre-training strategy and data augmentation pipeline using the multi-mic renders from the synthetic chamber ensemble dataset EnsembleSet and evaluate its impact using real-world chamber ensemble recordings from the URMP dataset. Our data augmentation pipeline, using synthetic data, has resulted in up to a remarkable +5.14 dB cross-dataset performance improvement for time-domain separation models when tested on real data. Our fine-tuning strategy in conjunction with our data augmentation pipeline results in up to +10.62 dB performance improvement w.r.t. our baseline for chamber ensemble separation. We report a strong negative correlation between pitch overlap and separation performance with an average of 5 dB performance drop for examples with pitch overlaps. We also show that pre-training our model with string, wind, and brass ensembles helps with separation of vocal harmony mixtures from Bach Chorales and Barbershop Quartet datasets with up to +17.92 dB SI-SDR improvement for 2 source vocal harmony mixtures.

***Index Terms*—** leveraging synthetic data, domain adaptation, cross-dataset evaluation, monaural source separation, chamber ensembles

1. INTRODUCTION

Audio source separation is the task of extracting individual sound sources from a recorded mixture. Our work primarily focuses on the challenge of source separation of musical mixtures, where each source is a musical instrument. Our sources are strongly correlated with each other by virtue of musical structure and context. The current paradigm of music source separation is highly focused on stem-based decomposition [1, 2, 3] as the research field was largely limited by the available datasets. While specific sub-tasks in the speech domain like speech denoising, multi-speaker separation, and dereverberation have been thoroughly explored, music separation research has largely been focused on the demixing task aided by the popular MUSDB dataset [4]. The demixing challenge is targeted at solving the problem of separating vocals, bass, and drums from mixed and mastered pop songs. Very little research has focused on other musical decomposition tasks such as separating harmonized sources. Even though the problem of music source separation has seen great strides recently reaching up to +8.11 dB [5] output SDR, we find enough room for improvement when compared to state-of-the-art speech separation performance which is able to achieve up to +22 dB output SDR using time-domain separation [6, 7] methods. TasNet [6] based approaches have not been successful for music separation and have been observed to introduce artifacts [3]. Chamber ensemble typically refers to a sub-genre of classical music with a small number of performers, mostly playing monophonic instruments in a highly synchronized fashion. We define the task of separating such mixtures with sources that suffer from label ambiguity and high timbral similarity as *ensemble separation*.

There are some separation tasks that fit into our definition of ensemble separation that have been explored recently. One such task is vocal harmony separation which has been tackled by [8, 9, 10, 11]. While the label ambiguity problem does exist for this task, some approaches have circumvented this issue by either looking at the problem in a class-based separation fashion by categorizing the constituent sources based on their vocal registers i.e. alto, soprano, bass, and tenor. One method of tackling this problem is called permutation invariant training (PIT) [12] which has been the preferred solution to tackle the label ambiguity problem for speech separation research [6, 7]. PIT has been utilized for ensemble separation in vocal harmony separation [10, 13] and chamber ensemble separation [14].

In this work, we analyze the impact of different types of mixtures on the task of separation from a single-channel recording. The mixtures consist of two harmonized monophonic sources. We first introduce a data augmentation method enabled by the multi-microphone renders available in EnsembleSet [14] and evaluate its impact on cross-dataset generalisability for time-domain and complex-domain separation models. We then introduce a pre-training strategy using synthetic data from EnsembleSet followed by fine-tuning on real-world datasets. We evaluate the impact of pre-training using EnsembleSet for both same-domain (chamber ensemble instruments from URMP dataset) and cross-domain data (harmonized vocals from Bach Chorales and Barbershop Quartet datasets). We train the models using PIT to separate any two monophonic sources, regardless of instrument type, and evaluate their ability to separate mixtures of the same or different instruments. Finally, we test the effectiveness of our separation strategy for mixtures with pitch overlaps between sources. Our key contributions are listed below:

- We present an instrument-agnostic source separation model trained using PIT that can separate any monophonic instrument from a mixture with a fixed number of sources.
- We propose a novel data augmentation pipeline using multi-mic renders from EnsembleSet that improve cross-dataset generalisability.

979-8-3503-2372-6/23/$31.00

- Pre-training ensemble separation models with EnsembleSet, followed by fine-tuning with a limited amount of target domain data, improves performance by up to +12.92 dB compared to baseline trained only on EnsembleSet and up to +2.1 dB compared to training only on target domain data.
- Pre-training vocal harmony separation models on chamber ensemble instruments improves performance by up to +1.08 dB.
- Overlapping pitches between sources in musical mixtures significantly decrease separation performance of models trained with PIT, by up to -5.5 dB regardless of instrument type.

2. EXPERIMENTAL SETUP

2.1. Problem definition

We train our models to separate mixtures of a given number of monophonic musical sources regardless of the type of instrument/source, unlike other music source separation tasks. Using a PIT objective forces our model to learn how to separate mixtures based on onsets and pitch tracking instead of timbres. This approach also enables our model to be able to separate mixtures of identical instruments (eg: 2 violins), similar sounding instruments (eg: violin+viola, or 2 singers), and also unseen instruments/sources. Another advantage of using PIT is related to the amount of training data, where we are able to generate $\binom{N}{2}$ training examples from a piece with N concurrent sources which greatly improves the total training data size.

There are a few drawbacks of our problem formulation as well. Firstly, there are some monophonic instruments (such as violins) where there are rare instances of a performer playing multiple notes at an instance, in which case our model is confounded since it expects the sources to be monophonic. The second drawback is that due to the nature of PIT, each model is constrained to the number of instruments present in the mixture to be the number of output nodes of the model.

2.2. Datasets

EnsembleSet [14] is a multi-track chamber ensemble music dataset synthesized by passing the MIDI transcriptions from the RWC Classical Music Database [15] and lilypond scores from Mutopia [16] through a realistic sample library Spitfire BBC Symphony Orchestra (BBCSO) [17]. This dataset presents 18 unique multi-mic recordings and 2 professional mixes for each individual source which can be used as data augmentation to avoid overfitting models to the synthesised dataset. EnsembleSet contains a total of 25 hours of multi-mic instrument renders and is focused around string ensembles (24 hours) with a limited amount of wind and brass instruments (30 minutes each). The synthetic nature of the dataset, combined with its relatively large size, enables us to test if pre-training our model on a large variety of musical contexts with limited timbral diversity can benefit ensemble separation models.

Since our models are pre-trained on synthetic data, we use real-world recordings from the URMP dataset [18] which is a multi-modal, multi-track dataset comprising audio-visual recordings of 44 chamber ensemble pieces. Unlike most other multi-track datasets of chamber ensembles, this dataset takes particular care to ensure that the individual instrument recordings do not contain bleed. In order to achieve this, each instrument was recorded in a separate take, subsequently, each of these recordings were dereverberated and downmixed together with the other instruments with reverb.

To study the transferability of features learned from chamber ensemble instruments to vocals, which have significantly different dynamics and modulations compared to bowed and wind instruments, we use the Bach Chorales and Barbershop Quartets (BCBQ) datasets. These are multi-track datasets of *a capella* recordings from [19] which we use for fine-tuning our pre-trained model from ensemble separation to vocal harmony separation. They include 26 songs from Bach Chorales (BC) and 22 songs from Barbershop Quartets (BQ). This gives us a total of 104 minutes of 4 parts: Soprano, Alto, Tenor, and Bass (SATB) recordings, where BC contains 2 male (tenor and bass) and 2 female (soprano and alto) vocalists, and BQ contains all 4 male vocalists.

2.3. Models

We choose two different baselines for our experiments, one for time-domain end-to-end separation (DPTNet [7]), and one for complex domain separation (DCUNet [20]), both of which have shown comparable results for speech separation using PIT.

We use the DPTNet models (9.9M parameters) as described in [14, 10] for our experiments with URMP and Choral Music, respectively. We train our models at 44.1 kHz except for experiments related to vocal harmony mixtures, where we pre-train our model at 22.05 kHz as the test dataset is bandlimited and we have observed noisy separation when models were trained at a higher sample rate than the contents of the data [10].

DCUNet (7.7M parameters) builds upon the original U-Net [21] by introducing a phase-aware complex-valued masking framework. We use the asteroid [22] implementation of this model with PIT as a baseline to compare and contrast our observations.

2.4. Training

We train our models with synchronised pairs of musical instruments. Our data pre-processing involves activity detection on the source monophonic instrument audio files and identifying frames of 2.97 seconds (131072 samples at 44.1 kHz) where both instruments are concurrently active for at least 40% of the frame. We generate a train-validation split by randomly choosing 10% of the training frames presented to the dataloader as the validation set. We generate our input mixtures by linearly downmixing the augmented versions of our reference sources.

All of our models are trained at 44.1 kHz, except the experiments associated with vocal harmony separation which are trained and evaluated at 22.05 kHz. We use the SI-SDR [23] metric as our loss function with PIT. We train our DPTNet models for 100 epochs with early stopping patience of 10 epochs. We start with a learning rate of $5 \times e^{-3}$ with a scheduler that halves the learning rate if the validation loss does not improve for 3 epochs. We train our models on 4 x NVIDIA A100 GPUs with a batch size of 3 per GPU using a distributed data parallel backend.

2.5. Data Augmentation

The data augmentation is applied on-the-fly using torch-audiomentations [24] and is applied across all the experiments, except using multi-mic renders from EnsembleSet. We stochastically apply gain modulations to each of the sources in a mixture separately in the range of $+5$dB to -15dB, pitch shift by up to ± 2 semitones, followed by channel swaps for the reference targets.

The experiments using EnsembleSet for training have the opportunity to use the 20 unique microphone and mix configurations

that are presented in the dataset for each source. While most of these renders are stereo and utilize multiple microphones, we downmix them to mono for our experiments. We believe this exposes our models to a good variety of recording and microphone configurations which helps improve generalisability. To enable this, our dataloader selects one of the 20 available renders at random at each training iteration and the data augmentation described before is applied subsequently.

2.6. Fine-tuning/Pre-training

To enable our model to adapt to unseen acoustics and instruments, we propose using EnsembleSet with multi-mic augmentation as a pre-training step before training the model on limited test-domain data for improved performance. For the pre-training stage, we use the same training configuration as our EnsembleSet baseline experiments, where the train and validation sets are generated with a random split and we train the model for 100 epochs. We then use the learned model weights as the initial weights for re-training on target domain data with a low learning rate of $1 \times e^{-6}$. For URMP cross-dataset performance experiments, we use a single song from URMP, and 10 songs each from BC and BQ datasets as our fine-tuning data while using the remaining tracks as test data. We do not freeze any layers during this fine-tuning stage due to the nature of joint optimization of the free-filterbank and the separator stack in DPTNet.

2.7. Evaluation Framework

We use the "Close" mic-render for evaluating our models' performance on EnsembleSet. For our cross-dataset evaluation and domain adaptation experiments, we keep aside a few audio recordings (songs selected at random) from each of our datasets such that they are never presented to the model during training. For our smaller datasets, we use the same data for both training models from scratch and fine-tuning models pre-trained using EnsembleSet. We downmix the normalised individual sources to generate the test input mixtures with an input SNR of roughly 0 dB. We use the asteroid [22] implementations of SI-SDR and SDR to report our results. We then subsequently categorize our input mixtures based on instrument types and use pYIN [25] to detect pitch overlaps in test mixtures to segregate our test set and obtain deeper insights.

3. RESULTS

3.1. Impact of microphone augmentation

We trained the DPTNet and DCUNet models on EnsembleSet with and without multi-mic augmentation and test them on EnsembleSet and URMP. Table 1 shows the results of our models trained on EnsembleSet alone with and without multi-render data augmentation and tested on a separate test set from EnsembleSet and real-world data from URMP. We observe that both models suffer from overfitting and poor cross-dataset performance when tested on URMP data when trained without using multi-mic augmentation (MicAug). However, we observe a significant improvement in cross-dataset performance when using multi-mic augmentation only on DPTNet, while DCUNet results do not show any noticeable difference. We observe a drop in the same dataset performance of DPTNet models trained with MicAug, as the test data for EnsembleSet used the Close microphone, which was the same microphone configuration used for training the models trained without MicAug.

Table 1: Output SI-SDR for 2-source Chamber Ensemble Separation models trained on EnsembleSet with and without multi-mic augmentation (MicAug), tested on EnsembleSet (ES) and URMP.

Model	Sample Rate	MicAug	ES	URMP
DPTNet	22.05 kHz	✓	+13.67 dB	+9.39 dB
DPTNet	22.05 kHz	✗	+18.39 dB	+5.74 dB
DPTNet	44.1 kHz	✓	+13.24 dB	+9.23 dB
DPTNet	44.1 kHz	✗	+18.84 dB	+3.54 dB
DCUNet	44.1 kHz	✓	+14.43 dB	+4.49 dB
DCUNet	44.1 kHz	✗	+14.43 dB	+4.65 dB

3.2. Cross-dataset performance

Table 2 shows our models' evaluation results on cross-domain real-world datasets after pre-training on EnsembleSet with and without fine-tuning[1]. Our experiments demonstrate that pre-training using EnsembleSet leads to better separation results for both chamber ensemble and vocal harmony separation. Interestingly, even though choral singing is significantly different from chamber ensemble instruments, pre-training on chamber ensembles provides a +1.08 dB performance improvement over training on harmonized vocal data alone. While the SI-SDR achieved for vocal harmony separation is higher, it must be noted that the vocal harmony separation experiments were run at 22.05 kHz (instead of 44.1 kHz for other experiments) due to the band-limited nature of the BCBQ datasets as noted in [10].

Table 2: SI-SDR for 2-source Ensemble Separation models trained on EnsembleSet with fine-tuning on respective test datasets.

Model	SR	Train	Test	FT	SI-SDR
DPTNet	22.05 kHz	ES	BCBQ	✗	4.99 dB
DPTNet	22.05 kHz	ES	BCBQ	✓	17.92 dB
DPTNet	22.05 kHz	BCBQ	BCBQ	✗	16.84 dB
DPTNet	44.1 kHz	ES	URMP	✗	9.23 dB
DPTNet	44.1 kHz	ES	URMP	✓	14.87 dB
DPTNet	44.1 kHz	URMP	URMP	✗	13.25 dB
DCUNet	44.1 kHz	ES	URMP	✗	4.49 dB
DCUNet	44.1 kHz	ES	URMP	✓	12.71 dB
DCUNet	44.1 kHz	URMP	URMP	✗	10.61 dB

3.3. Musical context vs. Separation performance

Figure 1 provides deeper insights into the performance of our DPTNet-based model with fine-tuning for different mixture types. We use the URMP dataset's test data and divide it into 4 categories: same instruments vs. different instruments, mixtures with pitch overlaps and without pitch overlaps. Pitch overlaps are detected using pYIN [25] on each instrument's ground truth audio tracks. We find that examples with pitch overlap perform significantly worse ($\approx$ -5 db) than examples without pitch overlaps across all our models. We also find that mixtures of the same instruments perform slightly worse ($\approx$ -1 db) than mixtures with separate instruments.

[1] Audio examples: http://c4dm.eecs.qmul.ac.uk/EnsembleSet/

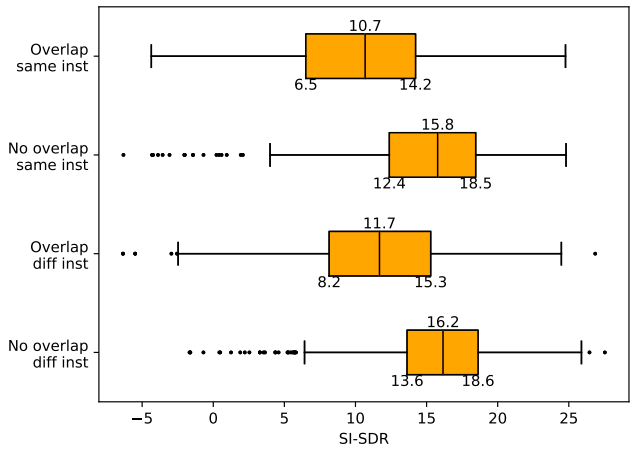

Figure 1: 2-source separation performance w.r.t. pitch overlap of DPTNet trained on EnsembleSet with FT, tested on URMP.

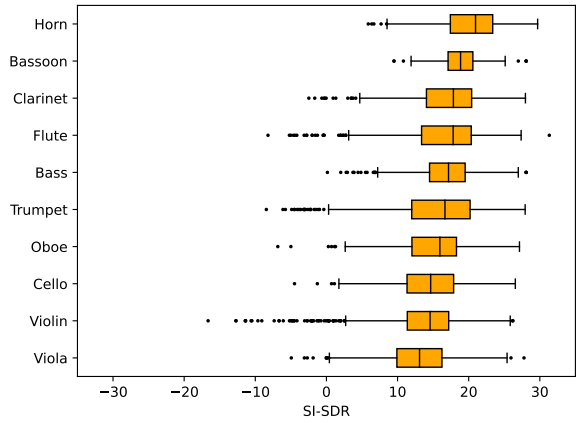

Figure 2: Instrument-wise median output SI-SDR of DPTNet trained on EnsembleSet with fine-tuning on URMP tested on 2 source mixtures from URMP dataset.

3.4. Instrument-agnostic performance

Even though EnsembleSet is 24 hours of strings and 1 hour of all other instruments, in Figure 2 we observe minimal variance across instruments. In fact, we see that most of the rarer instruments on average perform better than the most dominant instruments which are violin, viola, and cello. To test our hypothesis, we trained an EnsembleSet model with multi-render augmentation, excluding French horn training examples (see Figure 3). The average SI-SDR on URMP data was +8.8 dB, slightly lower (-0.49 dB) than the baseline. However, we found similar instrument-wise SI-SDR performance across different instruments, including the French Horn, which is an unseen instrument for the model. This suggests that models trained with PIT can separate unseen instruments.

4. DISCUSSION

We tested two models for monaural chamber ensemble separation: a TasNet-based architecture and a complex domain masking-based

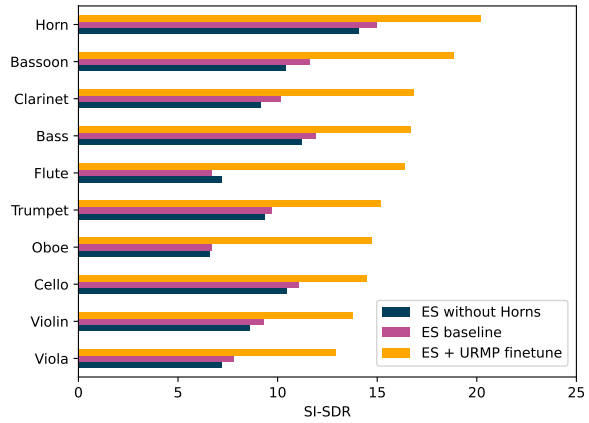

Figure 3: Average performance of DPTNet models tested on 2 source URMP mixtures.

architecture. We found that using multi-mic renders in EnsembleSet helped TasNet models generalize better. We also observe that the same augmentation has no effect on complex-domain models, this may be due to the fact that the 20 different microphone and mix renders present a larger variation in the input representation for TasNet-based models as compared to complex spectrogram representations. Our models trained with PIT were able to separate rare and unseen instruments and the same instruments mixtures. This suggests that the models trained with PIT are not learning specific timbral characteristics for each source but instead learning a generalizable separation strategy that focuses on pitch onsets and trajectories. This is further strengthened by our observation of a 5 dB average performance drop for examples with pitch overlaps regardless of the combination of instruments present in the mixture. We also show that pre-training models on large amounts of synthetic chamber ensemble data followed by fine-tuning on real data improves real-world performance for both chamber ensembles and vocal harmony mixtures by up to +2 dB. The fact that we observe performance improvement for vocal harmony separation which is significantly different from chamber ensembles from a source timbre and dynamics perspective also adds to our understanding that models trained with PIT tend to be timbre-agnostic. Hence, such models can be used effectively to separate any musical source with the only constraint that the instruments should be monophonic.

Our experiments show that separating mixtures of monophonic sources in an instrument-agnostic fashion is indeed successful using PIT. With fine-tuning, we are able to achieve very impressive results of 20+ dB separation improvement for many examples. The output quality for the better-performing examples is indeed approaching levels where it may be considered a pre-processing step for music production. However, we also find that the variance in performance is quite high and has significant dependence on the musical context of the mixture, such as when the sources play the same note. Indeed it is also a question worth considering, whether separating multiple instruments playing the same note is a source separation problem or a timbre disentanglement problem. In the future, we would explore further how to improve our models such that the performance is more consistent across all mixture scenarios, or even explore different loss functions and models to handle the case of timbre disentanglement which opens up many new avenues for source separation, and its impact on music production.

5. ACKNOWLEDGMENTS

S. Sarkar & L. Thorpe are research students at the UKRI Centre for Doctoral Training in Artificial Intelligence and Music, supported jointly by UK Research and Innovation [grant number EP/S022694/1] and Queen Mary University of London. E. Benetos & M. Sandler were supported by The Alan Turing Institute through Turing Fellowships. The computations described in this research were performed using the Baskerville Tier 2 HPC service (https://www.baskerville.ac.uk/), funded by the EPSRC and UKRI through the World Class Labs scheme (EP/T022221/1) and the Digital Research Infrastructure programme (EP/W032244/1) and is operated by Advanced Research Computing at the University of Birmingham.

6. REFERENCES

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

[2] R. Hennequin, A. Khlif, F. Voituret, and M. Moussalam, "Spleeter: A fast and state-of-the art music source separation tool with pre-trained models," in *Proc. International Society for Music Information Retrieval Conference*, 2019.

[3] A. Défossez, N. Usunier, L. Bottou, and F. Bach, "Music source separation in the waveform domain," *arXiv preprint arXiv:1911.13254*, 2019.

[4] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "Musdb18-a corpus for music separation," 2017.

[5] A. Défossez, "Hybrid spectrogram and waveform source separation," *arXiv preprint arXiv:2111.03600*, 2021.

[6] Y. Luo and N. Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 696–700.

[7] J. Chen, Q. Mao, and D. Liu, "Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation," *Proc. Interspeech 2020*, pp. 2642–2646, 2020.

[8] D. Petermann, P. Chandna, H. Cuesta, J. Bonada, and E. Gomez, "Deep learning based source separation applied to choir ensembles," *Proceedings of the 21st International Society for Music Information Retrieval Conference; 2020 Oct 11-16; Montréal, Canada. p. 733-9*, 2020.

[9] M. Gover and P. Depalle, "Score-informed source separation of choral music," Ph.D. dissertation, Ph. D. dissertation, McGill University, 2019.

[10] S. Sarkar, E. Benetos, and M. Sandler, "Vocal Harmony Separation Using Time-Domain Neural Networks," in *Proc. Interspeech 2021*, 2021, pp. 3515–3519.

[11] P. Chandna, H. Cuesta, D. Petermann, and E. Gómez, "A Deep-Learning Based Framework for Source Separation, Analysis, and Synthesis of Choral Ensembles," *Frontiers in Signal Processing*, vol. 2, 2022. [Online]. Available: https://www.frontiersin.org/article/10.3389/frsip.2022.808594

[12] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 241–245.

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

[14] S. Sarkar, E. Benetos, and M. Sandler, "Ensembleset: a new high quality synthesised dataset for chamber ensemble separation," in *Proc. of the 23rd International Society for Music Information Retrieval Conference (ISMIR)*, 2022, pp. 625–632.

[15] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "Rwc music database: Popular, classical and jazz music databases." in *Proc. of the 2nd International Society for Music Information Retrieval Conference (ISMIR)*, vol. 2, 2002, pp. 287–288.

[16] E. Praetzel, "Mutopia project: Free sheet music for everyone," 2000. [Online]. Available: https://www.mutopiaproject.org/

[17] SpitfireAudio, "User Manual BBC Symphony Orchestra Professional," 2019. [Online]. Available: https://d1t3zg51rvnesz.cloudfront.net/p/files/product-manuals/4126/1648649726/BBCSOPro_Manual_v2.0.pdf

[18] 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 Transactions on Multimedia*, vol. 21, no. 2, pp. 522–535, 2018.

[19] R. Schramm, E. Benetos, *et al.*, "Automatic transcription of a cappella recordings from multiple singers." Audio Engineering Society, 2017.

[20] 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 *International Conference on Learning Representations*, 2018.

[21] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep u-net convolutional networks," in *18th International Society for Music Information Retrieval Conference*, 2017, pp. 23–27.

[22] M. Pariente, S. Cornell, J. Cosentino, S. Sivasankaran, E. Tzinis, J. Heitkaemper, M. Olvera, F.-R. Stöter, M. Hu, J. M. Martín-Doñas, D. Ditter, A. Frank, A. Deleforge, and E. Vincent, "Asteroid: the PyTorch-based audio source separation toolkit for researchers," in *Proc. Interspeech*, 2020.

[23] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr – half-baked or well done?" in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 626–630.

[24] I. Jordal, "torch-audiomentations: Audio data augmentation in pytorch," 2021. [Online]. Available: https://github.com/asteroid-team/torch-audiomentations

[25] M. Mauch and S. Dixon, "pyin: A fundamental frequency estimator using probabilistic threshold distributions," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 659–663.