2.1 U-Net Model

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

2.2 Experimental Setting

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

We train all our models on a single NVIDIA GeForce 1080 Ti GPU, using a batch size of 8, and a learning rate of $1\mathrm{e}{-4}$ with ADAM optimizer. To improve the separation quality of real piano concerto recordings, we finetune the model with TTA, which we describe in the next section.

3. TEST-TIME ADAPTATION

Supervised deep learning models addressing the MSS task typically require a large dataset that consists of isolated recordings. As a data augmentation method, one can use random mixes to provide training material for an MSS model in the case isolated stems are missing [19, 20]. While this approach cannot simulate the harmonic and rhythmic relationships between various instruments in a real recording, it helps the model to distinguish timbral characteristics of the concurrent musical sources. However, the acoustic properties of recordings (including reverberation, and background noise) play an essential role when upmixing and separating different musical tracks. For instance, in the case of poor recording conditions, (e.g., historical recordings) the properties of the test data may not be reflected well in the training set, thus resulting in a poor separation quality. Finetuning a pre-trained MSS model in the testing phase using a few samples drawn from the test data (also called *test-time adaptation (TTA)* [26]) can improve the separation quality by capturing the specific acoustic features found in a music recording.

From this perspective, separation of piano concertos is a particularly suitable scenario for applying TTA thanks to their compositional form. Depending on the period in which the work was composed, these compositions often comprise long piano-only (e.g., in the cadenza) and orchestra-only parts (e.g., in the exposition, also called *opening riternello*). Using these sections, one can create artificial mixes which come from the audio material of the given test item. As a result, the mixes share the same recording conditions as the test data.

To investigate this approach, we consider seven piano concerto recordings (see Figure 2a). The selected movements of these piano concertos have a long cadenza, which contains only the piano (see Figure 2b). Note that, with the exception of `BrahOp015`, these musical pieces also comprise a long exposition in which only the orchestra plays. For our experiments, we annotate the piano-only and orchestra-only sections, which are publicly available. [1] Exploiting the structural characteristics of piano concertos, we create random mixes of piano-only and orchestra-only sections, which serve as further training data for model adaptation for each piano concerto individually. In the next section, we investigate the improvement of qualitative and subjective separation quality via TTA.

4. EVALUATION

In this section, we report on the separation results acquired by our pre-trained model `PT` and the finetuned models `TTA`. In Section 4.1 we describe our test dataset. We discuss the quantitative empirical results in Section 4.2 and present the subjective evaluation in Section 4.3.

4.1 Test Dataset

For the evaluation of our models, we generate 30-second random mixes of piano-only and orchestra-only parts sampled from the annotated piano concertos (see Figure 2b). These are different from the artificial training set, which we use for the pre-trained model, as they share harmonic and acoustic properties originating from the same recording. Note that we ensure that the samples used for training do not overlap with the test mixtures which we use for the evaluation purposes.

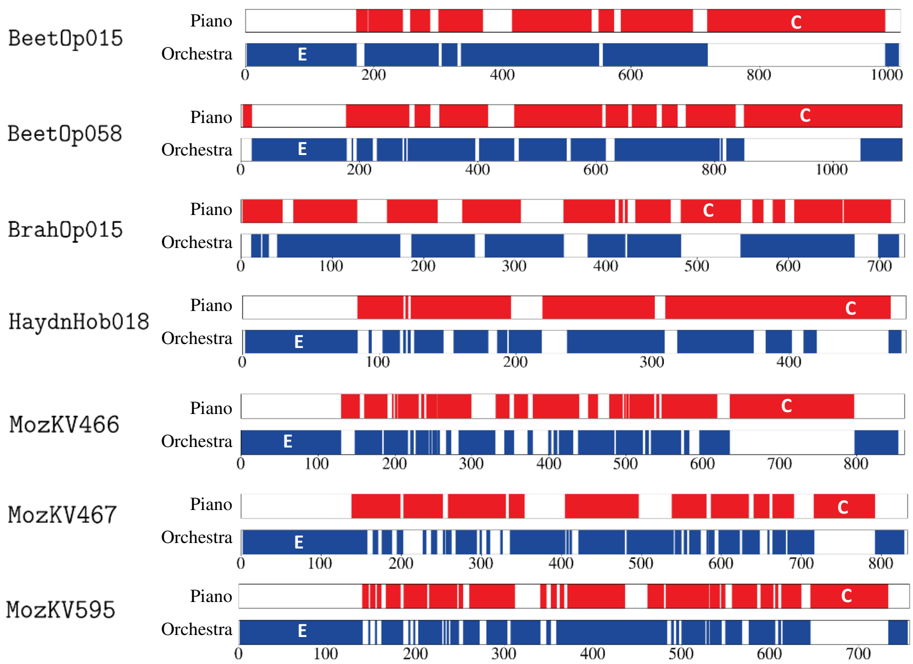
4.2 Quantitative Evaluation

To get a first impression of the performance of the models `PT` and `TTA`, we use the SDR [27] as our quantitative evaluation metric for the separation. Table 1 provides a comparison of the resulting SDR values between a baseline for the SDR values (denoted as `BL`), which we compute using the test mixture as the target signal and ground-truth sources as the reference, pre-trained `PT`, and finetuned `TTA`

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

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

(a) The table shows the composer, full name of the work, performer, identifier, recording year, movement (M.), duration (Dur.) in seconds of total recording (T), exposition (E), and cadenza (C).



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

Figure 2: Overview of the piano concertos in our test dataset.

after 100 iterations. One can observe that `PT` leads to a substantial improvement in SDR values compared to `BL` over the whole dataset, both for the piano and orchestra. It is interesting to observe that `PT` improves the SDR value of `BeetOp015` for the separation of piano from 4.48 to 4.60, which is a relatively low improvement compared to other piano concertos in the dataset. Note that `BeetOp015` is a historical recording (see Figure 2a for the recording year of the piano concertos), whose inadequate recording conditions may not be well represented in the random mixes used for the training of the `PT`, thus leading to a relatively poor separation performance.

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

In our next experiment, we investigate the performance of the finetuned models `TTA` per iteration. Figure 3 illustrates the evolution of the SDR values for each piano concerto in our test dataset. The overall convergence behavior exhibits a general trend of improvement of SDR values through TTA over `PT` for the separation of the piano and the orchestra. In particular, the SDR values for the separation of `BeetOp015` depict a rapid improvement within

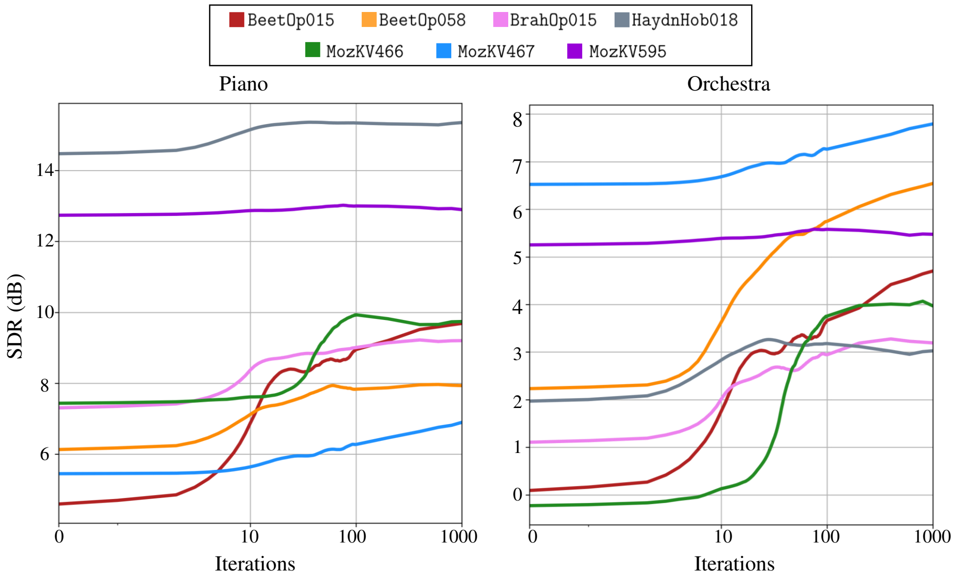


Figure 3: Evolution of SDR values based on our test dataset, applying TTA on the pre-trained model PT.

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

Table 1: Comparison of the SDR (dB) values between the baseline BL, and the separated sources by the pre-trained model PT and the finetuned model TTA after 100 iterations. The average SDR values are denoted with ϕ.

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

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

the first 10 iterations. For the other piano concertos, the improvement of SDR values generally accelerates after the 10th iteration. After the 100th iteration, the separation performance remains steady for most of the piano concertos. Furthermore, after the 100th iteration, the SDR values constantly increase in the case of BeetOp015 and MozKV467 for both piano and orchestra.

Although SDR is widely used as a quantitative evaluation metric for MSS, it is well known that it may not be suitable for determining the perceptual sound quality of separated musical sources [34]. The work by Torcoli et al. [35] provides a comparison of objective quality measures in the source separation domain. Their analysis indicates that a quantitative evaluation using the metric called *2f-model* exhibits the best correlation with ground-truth data based on the subjective ratings from MUSHRA listening tests. For a detailed account of the 2f-model, we refer to [28]. Note that the 2f-model values range from 0 to 100 following MUSHRA rating scores (see Section 4.3). Table 2 provides the resulting 2f-model values for the separated sources by PT and TTA using 100 iterations. In general, one can observe here a similar trend as for the SDR. PT mostly reveals better 2f-model scores than the baseline BL, except for the piano separation of BeetOp015, which presumably suffers from its poor recording conditions that are not well-represented in the artificial training set.

As for the SDR-based results, 2f-model values increase via TTA after 100 iterations for both the piano and the orchestra. For example, in the case of BeetOp015, PT yields a 2f-model value of 21.50 for the separated piano, improving to 28.39 after applying TTA. Interestingly, the separation of the orchestral part yields better results than the piano according to 2f-model values. This is opposed to the evaluation based on the SDR, where the separation results are significantly better in the case of piano separation (see Table 1).

4.3 Subjective Evaluation

In this section, we describe the experimental setting for our subjective evaluation to assess the perceived separation quality. We carried out two listening tests using the MUSHRA methodology following the ITU-R BS.1534-3 recommendation [36]. It is a double-blind multi-stimulus test method with a hidden reference and an additional lower anchor signal. The rating scores range from 0 to

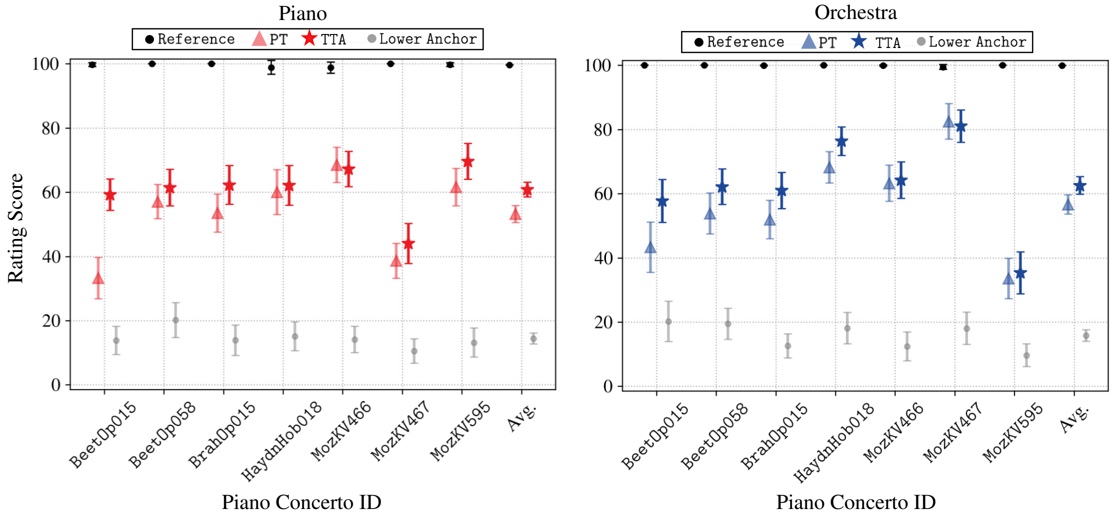


Figure 4: Results of our subjective evaluation based on MUSHRA listening tests for the piano (left) and orchestra (right). The colored markers indicate the average rating scores enclosed by 95% confidence intervals (indicated by the vertical lines).

100, involving five categories: Bad (0-20), Poor (20-40), Fair (40-60), Good (60-80), and Excellent (80-100).

In total, 34 participants were involved in our listening tests (31 experienced listeners and 3 inexperienced listeners). The MUSHRA methodology suggests a post-screening of the participants stating that participants should be excluded from the listening test if they give the hidden reference a score lower than 90 for more than 15% of the test items. Concerning our tests, one of the participants was excluded after post-screening.

Each of the two listening tests contains seven test items. For each test item, we generated four different signals with a duration of 12 seconds (maximum allowed signal duration for MUSHRA listening tests), which are excerpts from the test mixtures that we use for our quantitative evaluation. In our first listening test, we asked the participants to rate the overall audio quality for each of the four signals (also called *conditions*) with respect to a reference signal, which is a clean piano-only section. Similar to [37], we created the lower anchor signals by low-pass filtering the test mixtures to a 3.5 kHz cut-off frequency and by adding musical noise, i.e., randomly setting 20% of the remaining time/frequency coefficients to zero. The other two signals involve separated piano parts by `PT` and `TTA`. Similarly, our second listening test evaluates the overall quality of the separated orchestral parts following the same procedure as the first listening test.

Figure 4 summarizes the results from our listening tests. First, one can observe that the participants rated the reference signal with an average MUSHRA rating score of 99 and the lower anchor is significantly below the conditions `PT` and `TTA`. Remarkably, the general trend of the performances by `PT` and `TTA` support our quantitative analyses, inferring that the TTA procedure generally improves the separation of both the piano and orchestra. When observing the rating score of the piano concertos individually, one can observe that the rating of the historical recording `BeetOp015` is significantly lower than other items for `PT`. Intuitively, this is due to its poor recording conditions. After applying TTA, the average rating score of `BeetOp015` improves from 33 to 59 for the piano and from 43 to 58 for the orchestra. Furthermore, the orchestra separation led to a lower MUSHRA score in the case of `MozKV595`, both for `PT` and `TTA`. One reason may be the audible clipping artifacts in the reference signal and hidden separated orchestra, which a subset of the participants noted during the listening test.

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

5. CONCLUSION

In this paper, we investigated the separation of piano and orchestra in piano concertos. We trained our model using a U-Net architecture based on the Spleeter implementation with random mixes of solo piano and orchestral recordings, which we regarded as our baseline pre-trained model. As the main contribution, we proposed a TTA procedure to enhance the separation quality using the random mixes created from the samples found in the test data. We showed that TTA substantially improved the quantitative and subjective evaluation results, both for the piano and orchestra. For the future, we aim to explore musically plausible data augmentation methods that simulate more realistic mixtures. To further improve the separation quality, avenues of research may be to integrate a transcription model as proposed by [38] or to incorporate phase information into the network by using, e.g., a complex U-Net [39]. Moreover, we intend to further investigate objective evaluation measures for the source separation of piano concertos.

Acknowledgements: This work was supported by the German Research Foundation (DFG MU 2686/10-2). The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer Institut für Integrierte Schaltungen IIS.

6. REFERENCES

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

[2] A. Jansson, E. J. Humphrey, N. Montecchio, R. M. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep U-net convolutional networks," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Suzhou, China, 2017, pp. 745–751.

[3] N. Takahashi and Y. Mitsufuji, "Multi-scale multi-band densenets for audio source separation," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017, pp. 21–25.

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

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

[6] D. Stoller, S. Ewert, and S. Dixon, "Wave-U-net: A multi-scale neural network for end-to-end audio source separation," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Paris, France, 2018, pp. 334–340.

[7] A. Défossez, N. Usunier, L. Bottou, and F. R. Bach, "Music source separation in the waveform domain," 2019. [Online]. Available: http://arxiv.org/abs/1911.13254

[8] A. Défossez, "Hybrid spectrogram and waveform source separation," in *Proceedings of the ISMIR 2021 Workshop on Music Source Separation*, Online, 2021.

[9] H. Liu, Q. Kong, and J. Liu, "CWS-PResUNet: Music source separation with channel-wise subband phase-aware ResUNet," in *Proceedings of the ISMIR 2021 Workshop on Music Source Separation*, Online, 2021.

[10] M. Kim, W. Choi, J. Chung, D. Lee, and S. Jung, "KUIELab-MDX-Net: A two-stream neural network for music demixing," in *Proceedings of the ISMIR 2021 Workshop on Music Source Separation*, Online, 2021.

[11] R. Sawata, S. Uhlich, S. Takahashi, and Y. Mitsufuji, "All for one and one for all: Improving music separation by bridging networks," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Toronto, ON, Canada, 2021, pp. 51–55.

[12] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "The MUSDB18 corpus for music separation," 2017. [Online]. Available: https://doi.org/10.5281/zenodo.1117372

[13] R. M. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. P. Bello, "MedleyDB: A multitrack dataset for annotation-intensive MIR research," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Taipei, Taiwan, 2014, pp. 155–160.

[14] M. Schedl, D. Hauger, M. Tkalčič, M. Melenhorst, and C. C. S. Liem, "A dataset of multimedia material about classical music: PHENICX-SMM," in *Proceedings of International Workshop on Content-Based Multimedia Indexing (CBMI)*, 2016, pp. 1–4.

[15] C. Joder and B. W. Schuller, "Score-informed leading voice separation from monaural audio," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Porto, Portugal, 2012, pp. 277–282.

[16] E. Cano, G. Schuller, and C. Dittmar, "Pitch-informed solo and accompaniment separation towards its use in music education applications," *EURASIP Journal on Advances in Signal Processing*, vol. 2014, no. 23, 2014. [Online]. Available: https://doi.org/10.1186/1687-6180-2014-23

[17] Z. Rafii, A. Liutkus, F. Stöter, S. I. Mimilakis, D. FitzGerald, and B. Pardo, "An overview of lead and accompaniment separation in music," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 8, pp. 1307–1335, 2018.

[18] F. Stöter, A. Liutkus, and N. Ito, "The 2018 Signal Separation Evaluation Campaign," in *Proceedings of the International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, ser. Lecture Notes in Computer Science, vol. 10891. Springer, 2018, pp. 293–305.

[19] 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 *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, New Orleans, Louisiana, USA, March 2017, pp. 261–265.

[20] C.-Y. Chiu, W.-Y. Hsiao, Y.-C. Yeh, Y.-H. Yang, and A. W.-Y. Su, "Mixing-specific data augmentation techniques for improved blind violin/piano source separation," in *Proceedings of the Workshop on Multimedia Signal Processing (MMSP)*, 2020, pp. 1–6.

[21] Z. Rafii and B. Pardo, "Repeating pattern extraction technique (REPET): A simple method for music/voice separation." *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 1, pp. 71–82, 2013.

[22] B. Lehner, R. Sonnleitner, and G. Widmer, "Towards lightweight, real-time-capable singing voice detection," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Curitiba, Brazil, 2013, pp. 53–58.

[23] C. Dittmar, B. Lehner, T. Prätzlich, M. Müller, and G. Widmer, "Cross-version singing voice detection in classical opera recordings," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Málaga, Spain, October 2015, pp. 618–624.

[24] G. Cantisani, A. Ozerov, S. Essid, and G. Richard, "User-guided one-shot deep model adaptation for music source separation," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, New York, USA, 2021, pp. 111–115.

[25] Y. Wang, J. P. Bello, D. Stoller, and R. Bittner, "Few-shot musical source separation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Singapore, Singapore, 2022, pp. 121–125.

[26] Y. Sun, X. Wang, L. Zhang, J. Miller, M. Hardt, and A. A. Efros, "Test-time training with self-supervision for generalization under distribution shifts," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2020.

[27] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[28] T. Kastner and J. Herre, "An efficient model for estimating subjective quality of separated audio source signals," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, New York, USA, 2019, pp. 95–99.

[29] B. Series, "Method for the subjective assessment of intermediate quality level of audio systems," *International Telecommunication Union Radiocommunication Assembly*, 2014.

[30] A. Liutkus and R. Badeau, "Generalized Wiener filtering with fractional power spectrograms," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, April 2015, pp. 266–270.

[31] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention MICCAI*, Munich, Germany, 2015, pp. 234–241.

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

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

[34] E. Cano, D. FitzGerald, and K. Brandenburg, "Evaluation of quality of sound source separation algorithms: Human perception vs quantitative metrics," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, 2016, pp. 1758–1762.

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

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

[37] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2046–2057, 2011.

[38] E. Manilow, P. Seetharaman, and B. Pardo, "Simultaneous separation and transcription of mixtures with multiple polyphonic and percussive instruments," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020, pp. 771–775.

[39] 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 *Proceedings of the International Conference on Learning Representations (ICLR)*, Vancouver, Canada, 2018.

COUNTERPOINT ERROR-DETECTION TOOLS FOR OPTICAL MUSIC RECOGNITION OF RENAISSANCE POLYPHONIC MUSIC

Martha E. Thomae
McGill University
martha.thomaeelias@mail.mcgill.ca

Julie E. Cumming
McGill University
julie.cumming@mcgill.ca

Ichiro Fujinaga
McGill University
ichiro.fujinaga@mcgill.ca

ABSTRACT

In this paper, we present a music information retrieval (MIR) pipeline to aid musicologists in making editions of mensural music sources. We designed this pipeline by improving existing MIR tools and allowing for their interoperability rather than implementing a new monolithic tool. These MIR tools include technologies such as optical music recognition (OMR), automatic voice alignment for mensural notation, editorial correction software, and computational counterpoint error detection. To ease the editorial correction process necessary to obtain correctly lined-up scores, we evaluate whether the use of counterpoint error-detection tools makes the correction process more efficient. While this idea has been discussed before, this paper presents the first attempt at implementing it. The results confirm that marking illegal dissonances in the score following the rules of Renaissance counterpoint makes the process of editorial correction of scribal errors in Renaissance music more efficient by reducing the time taken and improving the accuracy of such corrections. Moreover, it also allowed us to catch OMR errors that had passed through undetected at a previous step of the pipeline. This paper is part of a larger project to preserve and increase access to a set of Guatemalan polyphonic choirbooks through digital images and symbolic scores.

1. INTRODUCTION

This paper discusses part of a larger project to preserve and increase access to a set of six Guatemalan Cathedral choirbooks (the *GuatC* collection). [1] These choirbooks, written in mensural notation, contain mostly sixteenth-century polyphonic music that was copied in the seventeenth and eighteenth centuries [1]. They document a continuous performance tradition of sacred choral music from the Renaissance until the beginning of the nineteenth century and are valuable sources for studying the transmission of music from Europe to Latin America.

[1] *GuatC*: Guatemala. Guatemala City. Catedral, Archivo Capitular. Other sigla include GCA-Gc.

Given the limited access to these manuscript sources, it is of utmost importance to digitize and encode this corpus which otherwise might be lost, damaged, or forgotten. We used a set of digitization and music information retrieval (MIR) technologies to obtain digital images and symbolic files encoding scores with editorial corrections for each of the pieces of the first choirbook of the collection, *GuatC 1* (see Figure 1). In the process, we tested the following three-step pipeline: (i) digitization, (ii) optical music recognition, and (iii) automatic voice alignment & editorial correction. The digitization step, conducted with a do-it-yourself scanner, was discussed in a previous publication [2]. In this paper, we focus on the encoding part of the pipeline, the last two steps.

1. **Optical music recognition (OMR) & correction of the results.** We perform OMR on the images to retrieve a symbolic file (a Mensural MEI file) encoding the music of the manuscript. We used the *Music Recognition Encoding and Transcription* (*MuRET*) OMR framework for this step (see Section 2.2).
2. **Automatic voice alignment & editorial corrections.** Since we are dealing with mensural notation, OMR is not enough to encode the full rhythmic information of the pieces and obtain a score with the voices properly lined up (see Section 2.1). Two additional steps are needed: (1) *automatic voice alignment*, which provides the actual duration of each note and returns a preliminary score; and (2) *editorial correction*, which allows for corrections of scribal errors, some of which can affect the alignment of the voices into a score. We used the *Measuring Polyphony* (*MP*) *Editor* for this (see Section 2.3).

This process generates symbolic scores with editorial corrections in a semi-automatic way through OMR and automatic voice alignment. The user manually corrects the output of each step, correcting the results of the OMR—the recognized symbols and their pitches—in MuRET and correcting the results of the voice alignment in the MP Editor. The latter normally implies the correction of scribal errors in the form of editorial corrections and occasional OMR errors that went undetected in the previous step of the pipeline. We evaluate whether the use of a tool that identifies illegal dissonances in Renaissance counterpoint (see Section 2.4) helps in the correction of this last step of the pipeline. The goal of this MIR pipeline is to aid musicologists in making editions of mensural music sources, with the last step aimed to reduce the challenge of aligning (possibly error-ridden) polyphonic parts by automatically scoring up the voices and by flagging areas of special attention to the human editor—given how time-consuming it can be to find scribal errors. While the idea of using counterpoint rules to detect errors in the original mensural sources has been mentioned before [3], it has not been implemented yet. This paper provides the first attempt at its implementation and evaluation by conducting a small experiment to test our hypothesis.

Figure 1: Example of a piece in GuatC 1. The voices are written in choirbook layout, where each voice is in a different area of the book opening.

2. BACKGROUND

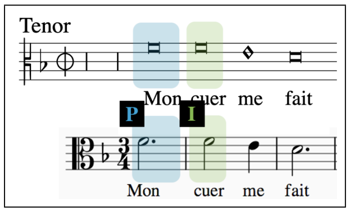
The GuatC collection consists of a set of six manuscript choirbooks written in mensural notation. The first three books have been inventoried [4–6], and the fourth one has been fully transcribed [6]. An overview of the whole collection was presented in [1] and a full inventory is expected [7]. Microfilm images for the first three books were created in the 1980s [8, pp. 3–4]; however, these images are of low quality, with cropped areas and missing folios.

We decided to test the MIR pipeline presented in Section 1 with the first choirbook (GuatC 1, one of the best-preserved manuscripts). The GuatC 1 is a book of masses. It contains twelve masses and fifteen short polyphonic pieces. Eight of the masses are from sixteenth-century composers, one mass is by a seventeenth-century composer, another by an eighteenth-century one, and two by composers whose period of activity remains unknown. On the other hand, most of the short polyphonic pieces are anonymous. Modern transcriptions of ten of the masses and four of the short pieces can be found in the *Choral Public Domain Library* (*CPDL*) wiki, although the provenance of the materials on which the transcriptions were based is shrouded in contradictory accounts. [2]

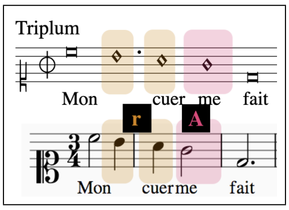
[2] The account found at the *Música Colonial Archive* page (https://www.cpdl.org/wiki/index.php/Música_Colonial_Archive) cannot be corroborated by the Centro de Investigaciones Regionales de Mesoamérica (CIRMA), the institution that holds the original microfilms, as indicated by the director of CIRMA's historical archive (Thelma Porres, personal communication, November 2018).

2.1 Mensural Notation and the Voice-Alignment Issue

Mensural notation was used for polyphonic music in Europe from the thirteenth to the seventeenth centuries. In triple meter, the duration of the notes in mensural notation depends on the context (i.e., the notes preceding or following), as shown in Figure 2. In Figure 2a, the same note shape has two different durational values, a ternary value, which is called *perfect*, and a binary value, which is called *imperfect*. In Figure 2b, another note shape represents two different durations, a regular one and an *altered* one where the note has twice the duration of its regular value. These three durational values of notes, perfect, imperfect, and altered, are common in fourteenth- to sixteenth-century mensural notation.



(a) Same note shape with a perfect (P; triple) and imperfect (I; duple) value.



(b) Same note shape with a regular (one beat) and altered (A; two beats) value.

Figure 2: Example of the different durational values of notes given the context. Both (a) and (b) show an example in mensural notation and its modern transcription below.

The context-dependent duration of mensural notes, together with the separate-parts layout of most mensural music (e.g., the choirbook layout shown in Figure 1), makes it difficult to know what notes are sounding at the same time in the different voices. We implemented an algorithm to compute the duration of notes based on the context and present the piece lined up in a score [9]. This algorithm is referred to as "automatic voice alignment" or "automatic scoring up" of mensural music. After scoring up the voices, it is still necessary to account for errors (e.g., missing or wrong values for notes and rests) to have a correct score. The scoring up of the piece helps to identify OMR and scribal errors that are obscured by the separate-parts arrangement of the music.

2.2 The Music Recognition Encoding and Transcription OMR Framework (MuRET)

There are a few OMR frameworks for early music, including the OMR workflow used by the SIMSSA project through the Rodan workflow manager with the Neon editor [10–13], the OMMR4all framework [14], Aruspix [15],