of audio effect modeling and three architectures originally proposed for music source separation.[1] For the latter models, instead of multiple output channels for different sources, we use only one output channel and consider them as general audio-to-audio transformation architectures.

CRAFx was proposed as a system for modeling time-varying audio effects with a neural network [9, 18]. The end-to-end model operates on the signal in the time domain and is divided into an adaptive front-end (encoder), a bi-directional long-short-term-memory (BLSTM)-based structure that applies the modeled effect in the latent space, and a synthesis back-end (decoder). In contrast to the other DNN-based models presented in this section, this architecture employs architectural priors (e.g., learnable nonlinear activations) in the context of audio effects.

Open-Unmix (UMX) was introduced as a reference implementation for music source separation [19]. The architecture is based on the BLSTM model from [20] and uses magnitude spectrograms as input features. The essential element of Open-Unmix is its three-layer BLSTM network that enables to learn both long- and short-time dependencies [21].

An element-wise multiplication of the input spectrograms with the estimated masks yields the final output. Commonly, spectrogram-based source separation models are compared with the oracle performance of an ideal ratio mask (IRM) that is defined as the ratio between the reference and the test spectrogram [22] in decibels. For reconstruction, the phase of the input signal is used. The model was adapted for a sampling rate of $f_s = 16\,\mathrm{kHz}$. We include this model in our evaluation as a standard frequency domain MSS model that relates to the BLSTM-based declipping model from [16] (cf. Sec. 3.2).

Wave-U-Net was proposed as one of the first end-to-end approaches for music source separation based on time domain signals [23]. Hence, it incorporates not only the magnitude but also the phase of music signals. It adapts the U-Net architecture [15] to one-dimensional audio signals. We decreased the models' number of learnable parameters from 17M to approximately 1M by reducing the number of layers from 12 to 8 resulting in a reduced receptive field, as we experienced overfitting for our datasets. Since the model was successfully employed not only for source separation but automatic mixing as well [3], we assume a general suitability for audio-to-audio transformation tasks. Moreover, the model is highly related to the U-Net-based declipping model from [14] (cf. Sec. 3.2).

Demucs was initially designed to be an end-to-end model for music source separation in the time domain [24]. While it builds on the Wave-U-Net model, it introduces several improvements to the architecture. The model comprises a convolutional encoder, a BLSTM structure, and a convolutional decoder. As for Wave-U-Net, we decreased the number of trainable parameters from 66M to 1M by reducing the number of blocks from 12 to 6 resulting in a reduced receptive field.

A-SPADE was introduced as a sparsity-based declipping algorithm that outperforms previous similar approaches [25]. For each time frame of the clipped signal $\mathbf{y}$, it approximates a solution of the following problem:

$$\min_{\mathbf{x},\mathbf{z}} ||\mathbf{z}||_0 \;\; \text{s.t.} \;\; \mathbf{x} \in \Gamma(\mathbf{y}) \;\; \text{and} \;\; ||\mathcal{F}(\mathbf{x}) - \mathbf{z}||_2 \leq \epsilon, \qquad (4)$$

where $\mathbf{z}$ denotes the unknown discrete Fourier coefficients of each time frame and $\mathcal{F}$ the Fourier transform operator. Γ is defined as the feasible space of solutions (i.e., clipping consistency constraint). We included the algorithm in the evaluation of the declipping task as a baseline that delivers state-of-the-art performance [11, 13].

5. EXPERIMENTS

Our experiments focus on the following three scenarios: **a)** We conducted experiments on guitar recordings that were processed using the overdrive algorithm of the audio editing software SoX [10] (CEG-OD). **b)** We performed the same experiments as in the previous scenario while processing the same clean guitar recordings with hard-clipping (CEG-HC). **c)** We tested the systems on an extensive dataset comprising various hard-clipped sounds (SignalTrain-HC) to evaluate their performance against a popular declipping algorithm when the models are trained given an increase in the amount and variety of data.

5.1 Data

The models were trained on two different datasets to assess the distortion audio effect removal capabilities. The audio signals from a dataset that contains a single instrument class (e.g., electric guitar) exhibit consistent signal statistics. In order to restrict the statistics that need to be modeled in a first step, we chose to concentrate on dry guitar samples as target data.

Since a large-scale, polyphonic dataset from clean electric guitar sounds is, to the best of our knowledge, not available[2], we used an internal dataset, which we refer to as CEG (Clean Electric Guitar) dataset. The monaural data were gathered from various sources, mainly commercial audio loop packages and recordings of solo guitar, and has a duration of $1.68\,\mathrm{h}$. All signals were re-sampled to a common sampling rate of $16\,\mathrm{kHz}$ in order to speed up convergence during training. To create the input dataset CEG-OD, the overdrive algorithm of SoX was applied to the data using five uniformly sampled gain levels in the range of $\gamma \in [20, 50]\mathrm{dB}$. Likewise, the input dataset for the hard-clipping task, CEG-HC, was created using hard-clipping (see (3)) with gain levels from the same distribution. Both datasets have a total length of $8.4\,\mathrm{h}$.

Although CEG represents a good source of data for our experiments due to its specificity of timbre, it remains a

[1] Two of the methods under evaluation, Demucs and Wave-U-Net, do not explicitly filter the signals in the audio domain, rather they perform a nonlinear mapping. Nevertheless, they were both successfully employed for MSS, which is a filtering problem.

[2] A publicly available guitar dataset for the recognition of audio effects exists (IDMT-SMT-Audio-Effects [26]). However, the dataset contains primarily homogeneous monophonic sounds and therefore, we choose to use the CEG dataset instead.

limited resource in terms of size and variety. Before attempting to train a system to handle recordings in real environments (e.g., a commercial song), we need to investigate how the current models at our disposal handle the availability of more and diverse training data. For this purpose, we also performed experiments on the SignalTrain dataset, which consists of more than 24 h of music and randomly-generated test signals [27]. By applying hard-clipping to these clean data using a uniformly-sampled input SDR value in the range $\mathrm{SDR}_{\mathrm{inp}} \in [1, 20]\mathrm{dB}$, we created SignalTrain-HC. During evaluation, we applied each input SDR in the set $\mathrm{SDR}_{\mathrm{inp}} \in \{1, 3, 5, 7, 10, 15, 20\}\mathrm{dB}$ to each sample in the test set.

Each dataset was split into non-overlapping subsets for training (80%), validation (10%) and testing (10%). We evaluated the models on the test split.

5.2 Experimental Setup

During the supervised training procedure, Adam [28] was used as optimizer with initial learning rates according to the model's original implementations. The learning rate was reduced by a factor of 10 after 150 epochs of no decrease in the validation loss. All models were optimized using the source-to-distortion ratio (SDR) between their full output sequences $\hat{\mathbf{x}}$ and the respective target sequences $\mathbf{x}$ in each batch $\mathcal{B}$ of N elements (not to be confused with the definition in the `BSS_eval` toolkit [29]):

$$\mathcal{L}_{\mathcal{B}}(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{N} \sum_{i \in \mathcal{B}} 10 \log_{10} \left(||\mathbf{x}_i||^2 / ||\mathbf{x}_i - \hat{\mathbf{x}}_i||^2 \right). \quad (5)$$

We stopped all trainings after 1000 epochs. All models processed audio sequences that are randomly extracted from each clip in the dataset; the length of the extracted sequences is equal to 2 s (2.3 s for CRAFx due to its architecture). We used a batch size of 16 for all experiments.

5.3 Objective Metrics

In speech enhancement and source separation, the ubiquitous measure to estimate the quality of a system is the SDR. However, applying (2) to a signal does not retain its energy. Therefore, we follow the approach of [30]: the **scale-invariant SDR (SI-SDR)** is obtained by rescaling the target signal s such that the residual $s - \hat{s}$ is orthogonal to s by using the optimal scaling factor $\hat{s}^T s / ||s||^2$.

An evaluation exclusively based on the objective similarity of the signals does not necessarily imply a correlation with human perception [31, 32]. Accordingly, we observed that the SI-SDR scores occasionally disagreed with our qualitative evaluation. Therefore, we also considered three metrics based on human perception.

The **perceptual evaluation of audio quality (PEAQ)** [33] is a widely used perceptual metric [11, 13, 34, 35] that measures the amount of degradation between two audio signals. The output of PEAQ is an Overall Difference Grade (ODG), which can reach values between 0 (*imperceptible impairment*) and -4 (*very annoying impairment*). Even though PEAQ is used in declipping and audio restoration studies [11, 13, 34], it was developed for audio codecs.

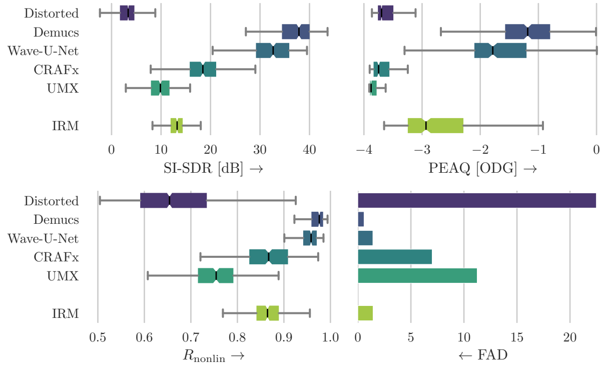

Figure 2. Box plot of scores for the CEG-OD dataset. The boxes show the first and third quartile of the data while the median is indicated with a line in the box. Higher scores indicate superior performance except for FAD. The results suggest that U-Net-based models performing convolutions in the time domain are the most promising approach to solving the task of overdrive removal.

The **R-nonlin metric** [36], in contrast, was developed specifically for detecting nonlinear distortions and, like PEAQ, considers the human auditory system. R_{nonlin} is defined between 0 (*high distortion*) and 1 (*no distortion*).

The **Fréchet audio distance (FAD)** was recently proposed as a reference-free evaluation metric for music enhancement algorithms. It has shown to correlate more with human perception than the SDR [37]. In order to obtain the FAD, the embedding statistics of both the whole clean and distorted test set are generated using a VGGish model [38]. The FAD is calculated based on the Fréchet distance between two multivariate Gaussians computed from both the test and the reference embeddings. [39].

6. RESULTS

In this section, we provide the results of the experiments introduced in the previous section.

6.1 De-Overdrive (CEG-OD)

Fig. 2 shows the results of the models that remove overdrive from guitar tracks.

Firstly, in SI-SDR, both Demucs and Wave-U-Net perform exceptionally well and even outperform the ideal-ratio-mask by more than 24 dB. While CRAFx yields considerably worse performance, it also surpasses the IRM oracle. UMX is the least performing of all the models in our comparison. It should be noted that for UMX, a model that operates on magnitude spectrograms, the IRM represents its upper limit in performance.

Similar results are obtained with PEAQ, R_{nonlin} and FAD: while Demucs and Wave-U-Net yield the best scores and surpass the IRM, the ones for CRAFx and UMX are considerably worse. It seems that directly processing the signals in the time domain using a U-Net-based architecture represents the most promising approach for the removal of the overdrive effect.

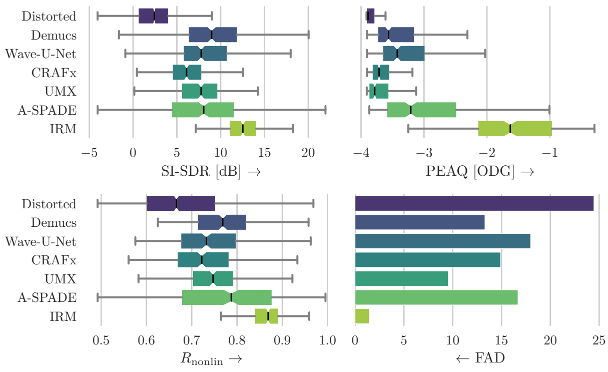

Figure 3. Box plot of scores for the CEG-HC dataset. It is difficult to clearly determine the best model, although Demucs represents a good compromise for all the metrics.

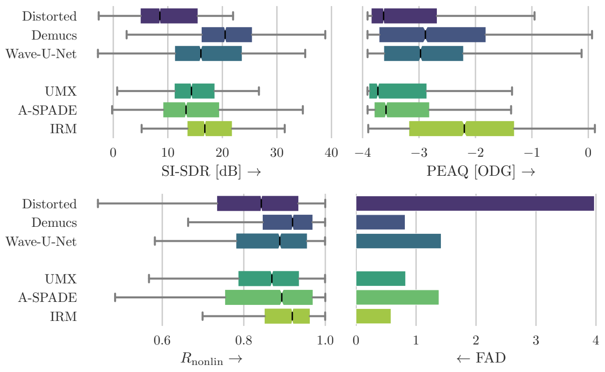

Figure 4. Box plot of scores for the SignalTrain-HC dataset. Demucs can be regarded as the best model regarding the median score across all metrics.

6.2 Declipping (CEG-HC)

Fig. 3 shows the results on the task of declipping on guitar recordings. Generally, we experienced a drop in the scores: now the models have been trained on declipping, which is an ill-posed problem, as missing parts of the signal need to be reconstructed. Additionally, we report scores for our declipping baseline, A-SPADE.

Despite the general performance drop, Demucs surpasses the results of A-SPADE in terms of SI-SDR by almost $1\,\mathrm{dB}$. While UMX and Wave-U-Net yield similar performances, CRAFx is the method with lowest scores.

Regarding PEAQ, no method surpasses the IRM and the A-SPADE algorithm. As before, Demucs and Wave-U-Net have similar performance: while PEAQ slightly favors the latter, SI-SDR favors the former. In contrast to the previous results, CRAFx has no significant advantage over UMX.

While Demucs achieves the best score for R_{nonlin} among the neural models, it does not surpass A-SPADE since the difference in their score is marginal. Interestingly, UMX achieves better results than CRAFx and Wave-U-Net. As the R_{nonlin} metric was explicitly designed to detect nonlinear distortions, we conclude that UMX' outputs contain fewer nonlinear distortions than the outputs of CRAFx and Wave-U-Net.

Surprisingly, when computing the FAD, UMX outperforms all other methods, including A-SPADE. This might be accounted to the fact that the FAD is based on the mel spectrum, whereas UMX optimizes the magnitude spectrogram. While Demucs and CRAFx outperform A-SPADE in FAD as well, the score for Wave-U-Net is slightly worse. The FAD for the IRM is relatively small because the difference between the reference and test embeddings from the VGGish model can be traced back to the masking operation and quantization. Demucs seems to constitute a good compromise regarding performance since it yields first- or second-best results for all metrics.

6.3 Declipping (SignalTrain-HC)

Fig. 4 shows the results for declipping SignalTrain-HC. Due to the lower gains that are used to prepare the data (see Sec. 5.1), the overall performance seems to be superior but cannot be directly compared to the previous results. Because of its considerably worse performance in the previous task, CRAFx was left out of the evaluation.

Demucs surpasses all other models in SI-SDR, including IRM and A-SPADE. Moreover, Wave-U-Net and UMX both surpass A-SPADE but not the IRM. PEAQ gives a similar ranking of the methods: only UMX cannot reach the A-SPADE baseline. None of the neural methods surpasses the IRM. In terms of R_{nonlin}, the same ranking is obtained, with Demucs surpassing, Wave-U-Net reaching, and UMX just missing A-SPADE. The FAD highlights that UMX and Demucs deliver comparable performance, outperforming the baseline, but not surpassing the IRM.

When only looking at SI-SDR or PEAQ, we notice the superiority of the time domain models. Future research should investigate whether the waveform in the time domain is the best input representation for the task, compared to, e.g., the real and imaginary part of a spectrogram [40] or both the waveform and the spectrogram [41]. Ultimately, Demucs can be considered the best model in our experiments for the task of declipping on SignalTrain.

6.4 Discussion

Qualitative Evaluation Although the abundance of evaluation metrics in the literature has the potential to analyze the results in very detailed ways, it does not always aid the judgement of which method is best given the individual context. Often, different applications have different requirements concerning the sound quality and can afford some types of distortions or artifacts to be left in the signal. Therefore, we report some qualitative considerations that need to be taken into account concerning our results, with the aim of finding some descriptive patterns among them. Fig. 5 shows spectrograms of a hard-clipped guitar signal and the outputs of all models under evaluation.

We found that the characteristics of the artifacts that each model produces are consistent, independent from the task it is applied to. Nevertheless, for the time domain-based models, the artifacts are less prominent in the deoverdrive task. Here, Demucs often produces outputs that

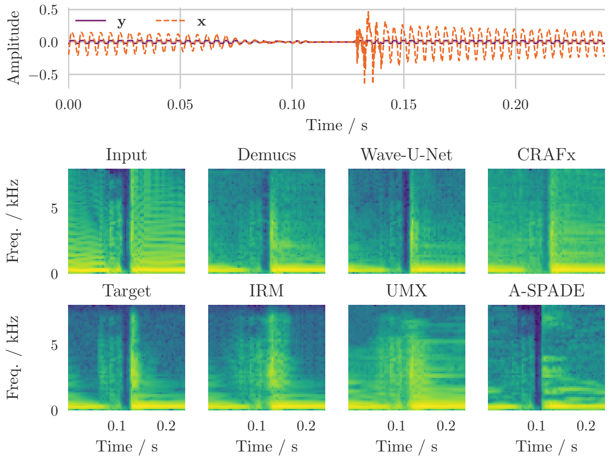

Figure 5. Spectrograms of a clipped guitar signal ($\gamma = 45\,\text{dB}$), its target signal, and the output of each method. The respective time domain input and target signal are shown at the top. While all methods significantly reduce the harmonics, it can be observed that IRM and UMX smear the transients. A-SPADE relies on the strongest bins in the clipped spectrum, which leads to tonal artifacts.

are virtually indistinguishable from the original signal.

Generally, Demucs is the model that most often produces high quality results. Especially for inputs with a low amount of distortion, it can reconstruct the original sound without any perceptual artifact. Wave-U-Net behaves similarly, although it often cannot reach the same quality.

UMX generally removes the distortion characteristics very well at the cost of strong phasing artifacts: Despite the absence of input distortions, the spectral features of the output are not necessarily consistent with the ones of the target. This is most likely due to UMX re-using the phase of the distorted input to reconstruct the signal in the time domain and the frame-wise processing. Moreover, Fig. 5 highlights that the transients are smeared by re-using the phase of the degraded signal.

CRAFx does not suffer from phasing artifacts, but occasionally leaves part of the distortion features in the output. In some cases, the model fails to reconstruct the onset of some notes, penalizing the listening experience.

Finally, A-SPADE is the model that exhibits the strongest and most frequent artifacts, especially for strongly clipped signals. Although it considers phase information by working in the complex frequency domain, it leads to non-optimal solutions. Nevertheless, distortion features (like those left by Demucs) or transient smearing (left by UMX) do not occur.

Influence of the Dry Signal We have observed a substantial difference in performance between models that need to remove overdrive (CEG-OD dataset) and models that need to remove hard-clipping (CEG-HC dataset). The superior performance of the models trained and tested on CEG-OD can be mainly traced back to the presence of the non-distorted signal in the overdrive output and not to the soft-clipping character of the specific overdrive implementation. We verified this hypothesis by training Wave-U-Net on hard-clipped data superimposed with the clean signal (even when the amplitude of the clean signal is considerably low). We obtained results similar to those in the de-overdrive task (median results without superposition: SI-SDR $= 7.2\,\text{dB}$, with superposition: SI-SDR $= 34.4\,\text{dB}$). While the performance drop is present for all metrics, it is less pronounced for UMX which seems not to utilize the additional information in the signal.

Inference Speed The benefits of Demucs in the context of declipping go beyond the quality of its outputs: using a neural approach also has advantages regarding inference speed. Inference with A-SPADE is comparably slow (real-time factor on CPU $\times\text{RT} \in [4.2, 27.3]$ depending on SDR_{inp} [13]), being an iterative approach that requires a computation of the Fourier transform and its inverse at each iteration. Demucs ($\times\text{RT} = 0.072$), Wave-U-Net ($\times\text{RT} = 0.113$) and UMX ($\times\text{RT} = 0.026$) instead, allow for fast inference on the CPU and even surpass real-time constraints independently on SDR_{inp} without sacrificing the quality of the results.

Evaluation Metrics The results highlight how our evaluation metrics focus on different aspects of the reconstructions: While SI-SDR measures differences between two audio signals in the time domain, PEAQ focuses on the perceptual quality without differentiating between degradation related to nonlinear distortions and artifacts/quality. In contrast, R_{nonlin} specifically highlights nonlinear distortions that have not been removed. Finally, FAD focuses primarily on the degradation that is observable in the mel spectrum. Hence, each metric proves beneficial in measuring specific aspects in the analysis of audio effect removal systems and can be advised for further investigations.

7. CONCLUSION

We showed that recovering the clean signal from clipped or overdriven audio signals can efficiently solved with neural networks designed for source separation. We found that Demucs achieves high quality according to the chosen evaluation metrics, especially when the distortion algorithm to be removed blends the distorted sound with the original one. This outcome highlights the potential of the proposed approach for other audio effects that mix dry and wet signals (e.g., parallel compression, reverberation, delay, modulation effects).

Moreover, we showed that Demucs, Wave-U-Net, and UMX outperform one state-of-the-art declipping method on our test data. This outcome is promising, considering that the dataset to train such a system is potentially much larger than the size of the dataset we used. By discussing the results, we stressed the usefulness of multiple evaluation metrics suitable to assess distortion removal systems.

Future work should include gathering more clean electric guitar data and generating a dataset using high-quality distortion emulations, which is required to improve generalization on real-world data. Furthermore, the knowledge from sparsity-based declipping algorithms could yield a valuable prior for declipping through DNNs.

8. REFERENCES

[1] U. Zölzer, *DAFX: digital audio effects*, 2nd ed. Chichester, West Sussex, England: Wiley, 2011.

[2] R. Stables, B. De Man, and J. D. Reiss, "Ten years of automatic mixing," in *Proceedings of the 3rd Workshop on Intelligent Music Production, Salford, UK, 15 September 2017*, 2017.

[3] M. A. Martínez-Ramírez, D. Stoller, and D. Moffat, "A Deep Learning Approach to Intelligent Drum Mixing With the Wave-U-Net," *Journal of the Audio Engineering Society*, vol. 69, no. 3, pp. 142–151, Mar. 2021.

[4] C. J. Steinmetz, J. Pons, S. Pascual, and J. Serra, "Automatic Multitrack Mixing With A Differentiable Mixing Console Of Neural Audio Effects," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* Toronto, ON, Canada: IEEE, Jun. 2021, pp. 71–75.

[5] Y. Mitsufuji, G. Fabbro, S. Uhlich, F.-R. Stöter, A. Défossez, M. Kim, W. Choi, C.-Y. Yu, and K.-W. Cheuk, "Music Demixing Challenge 2021," *Frontiers in Signal Processing*, vol. 1, Jan. 2022.

[6] W. Choi, M. Kim, M. A. Martínez-Ramírez, J. Chung, and S. Jung, "AMSS-Net: Audio Manipulation on User-Specified Sources with Textual Queries," in *Proceedings of the 29th ACM International Conference on Multimedia*, Apr. 2021, pp. 1775–1783.

[7] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "The MUSDB18 corpus for music separation," Dec. 2017.

[8] T. Wilmering, D. Moffat, A. Milo, and M. B. Sandler, "A History of Audio Effects," *Applied Sciences*, vol. 10, no. 3, p. 791, Jan. 2020.

[9] M. A. Martínez-Ramírez, E. Benetos, and J. Reiss, "Deep Learning for Black-Box Modeling of Audio Effects," *Applied Sciences*, vol. 10, p. 638, Jan. 2020.

[10] C. Bagwell, et al., "Sound eXchange (SoX)," Feb. 2015.

[11] P. Záviška, P. Rajmic, A. Ozerov, and L. Rencker, "A Survey and an Extensive Evaluation of Popular Audio Declipping Methods," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 1, pp. 5–24, Jan. 2021.

[12] A. Janssen, R. Veldhuis, and L. Vries, "Adaptive interpolation of discrete-time signals that can be modeled as autoregressive processes," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 2, pp. 317–330, Apr. 1986.

[13] C. Gaultier, S. Kitić, R. Gribonval, and N. Bertin, "Sparsity-Based Audio Declipping Methods: Selected Overview, New Algorithms, and Large-Scale Evaluation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1174–1187, 2021.

[14] H. B. Kashani, M. M. Goodarzi, A. Jodeiri, and S. G. Firooz, "Image to Image Translation based on Convolutional Neural Network Approach for Speech Declipping," *4th Conference on Technology In Electrical and Computer Engineering (ETECH 2019)*, 2019.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham: Springer International Publishing, 2015, pp. 234–241.

[16] W. Mack and E. A. P. Habets, "Declipping Speech Using Deep Filtering," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2019, pp. 200–204.

[17] T. Tanaka, K. Yatabe, M. Yasuda, and Y. Oikawa, "APPLADE: Adjustable Plug-and-Play Audio Declipper Combining DNN with Sparse Optimization," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* Singapore, Singapore: IEEE, May 2022, pp. 1011–1015.

[18] M. A. Martínez-Ramírez, E. Benetos, and J. D. Reiss, "A general-purpose deep learning approach to model time-varying audio effects," in *22nd International Conference on Digital Audio Effects (DAFx-19)*, Jun. 2019.

[19] F.-R. 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, p. 1667, Sep. 2019.

[20] 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 *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* New Orleans, LA: IEEE, Mar. 2017, pp. 261–265.

[21] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[22] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7092–7096.

[23] 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 19th ISMIR Conference, Paris, France, September 23-27, 2018*, Paris, Jun. 2018.

[24] A. Défossez, N. Usunier, L. Bottou, and F. Bach, "Music Source Separation in the Waveform Domain," *arXiv:1911.13254 [cs, eess, stat]*, Apr. 2021.

[25] S. Kitić, N. Bertin, and R. Gribonval, "Sparsity and Cosparsity for Audio Declipping: A Flexible Non-convex Approach," in *Latent Variable Analysis and Signal Separation*, E. Vincent, A. Yeredor, Z. Koldovský, and P. Tichavský, Eds. Cham: Springer International Publishing, 2015, pp. 243–250.

[26] M. Stein, J. Abeßer, C. Dittmar, and G. Schuller, "Automatic Detection of Audio Effects in Guitar and Bass Recordings," in *AES 128th Convention*, London, UK, May 2010.

[27] S. H. Hawley, B. Colburn, and S. I. Mimilakis, "SignalTrain: Profiling Audio Compressors with Deep Neural Networks," in *AES 147th Convention*, May 2019.

[28] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *3rd International Conference on Learning Representations, {ICLR} 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

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

[30] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – Half-baked or Well Done?" in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Brighton, United Kingdom: IEEE, May 2019, pp. 626–630.

[31] E. Cano, D. FitzGerald, and K. Brandenburg, "Evaluation of quality of sound source separation algorithms: Human perception vs quantitative metrics," in *2016 24th European Signal Processing Conference (EUSIPCO)*. Budapest, Hungary: IEEE, Aug. 2016, pp. 1758–1762.

[32] 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, Sep. 2011.

[33] T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. G. Beerends, and C. Colomes, "PEAQ - The ITU Standard for Objective Measurement of Perceived Audio Quality," *Journal of the Audio Engineering Society*, vol. 48, no. 1/2, pp. 3 EP – 29, Feb. 2000.

[34] S. Lattner and J. Nistal, "Stochastic Restoration of Heavily Compressed Musical Audio Using Generative Adversarial Networks," *Electronics*, vol. 10, no. 11, p. 1349, Jun. 2021.

[35] J. You, U. Reiter, M. M. Hannuksela, M. Gabbouj, and A. Perkis, "Perceptual-based quality assessment for audio–visual services: A survey," *Signal Processing: Image Communication*, vol. 25, no. 7, pp. 482–501, Aug. 2010.

[36] C.-T. Tan, B. C. J. Moore, N. Zacharov, and V.-V. Mattila, "Predicting the Perceived Quality of Nonlinearly Distorted Music and Speech Signals," *Journal of the Audio Engineering Society*, vol. 52, no. 7/8, pp. 699–711, Jul. 2004.

[37] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi, "Fréchet Audio Distance: A Reference-Free Metric for Evaluating Music Enhancement Algorithms," in *Interspeech 2019*. ISCA, Sep. 2019, pp. 2350–2354.

[38] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "CNN Architectures for Large-Scale Audio Classification," *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jan. 2017.

[39] D. Dowson and B. Landau, "The Fréchet distance between multivariate normal distributions," *Journal of multivariate analysis*, vol. 12, no. 3, pp. 450–455, 1982.

[40] W. Choi, M. Kim, J. Chung, D. Lee, and S. Jung, "Investigating U-Nets with various Intermediate Blocks for Spectrogram-based Singing Voice Separation," in *Proceedings of the 21th International Society for Music Information Retrieval Conference*, Oct. 2020.

[41] A. Défossez, "Hybrid Spectrogram and Waveform Source Separation," in *Proceedings of the ISMIR 2021 Workshop on Music Source Separation*, Nov. 2021.

END-TO-END FULL-PAGE OPTICAL MUSIC RECOGNITION FOR MENSURAL NOTATION

Antonio Ríos-Vila
U.I for Computer Research
University of Alicante, Spain
arios@dlsi.ua.es

José M. Iñesta
U.I for Computer Research
University of Alicante, Spain
iñesta@dlsi.ua.es

Jorge Calvo-Zaragoza
U.I for Computer Research
University of Alicante, Spain
jcalvo@dlsi.ua.es

ABSTRACT

Optical Music Recognition (OMR) systems typically consider workflows that include several steps, such as staff detection, symbol recognition, and semantic reconstruction. However, fine-tuning these systems is costly due to the specific data labeling process that has to be performed to train models for each of these steps. In this paper, we present the first segmentation-free full-page OMR system that receives a page image and directly outputs the transcription in a single step. This model requires only the annotations of full score pages, which greatly alleviates the task of manual labeling. The model has been tested with early music written in mensural notation, for which the presented approach is especially beneficial. Results show that this methodology provides a solution with promising results and establishes a new line of research for holistic transcription of music score pages.

1. INTRODUCTION

The majority of music creations, both historical and modern, are only available through music scores. Unfortunately, most of them have never been stored in a structured digital format, such as MEI [1] or Humdrum **kern [2], that allows their indexing, retrieval, and digital processing. The high cost of performing a manual transcription of these documents demands processes that solve this task automatically.

Optical Music Recognition (OMR) is the research field that studies how to computationally read music scores [3]. Most OMR literature is framed within a multi-stage workflow, where several steps, such as image binarization [4], staff-line removal [5], symbol classification [6, 7], and notation assembly [8], are considered. These pipelines, although effective to some extent, are hard to manage in a production environment, as they imply the development of several models to obtain a complete OMR system.

With the advent of machine learning technologies, namely those related to deep neural networks, OMR systems have evolved towards alternatives that simplify these complex workflows. Specifically, two trends emerged from this advance. On the one hand, complex multi-stage pipelines for symbol isolation and retrieval have been replaced by object detection algorithms [9], where symbols are directly located in the image. On the other hand, holistic strategies—commonly referred to as end-to-end approaches [10, 11]—are also used to transcribe music scores. In this case, the system directly outputs the transcription of a single-staff image.

Although the advances brought about by these new technologies have simplified these pipelines, these systems still rely on other models to complete the transcription of scores. In the case of the object detection approach, models for notation reconstruction and sequence encoding need to be implemented to obtain a readable representation of the music score. End-to-end approaches, in turn, are only able to transcribe single music staves, so staff detection algorithms have to be implemented to retrieve all the information from the page and be recognized individually [12, 13]. The latter pipeline is represented in the upper workflow of Fig. 1.

Following the path of end-to-end approaches, in this paper we present the first segmentation-free strategy for full-page music transcription that unifies the two required steps—see the lower workflow in Fig. 1—based on evolving the current state of the art. For this model to be trained, only page-level annotations—which could be available in any music encoding—are required. This greatly alleviates the labeling of music scores, as additional OMR-specific data preprocessing steps are avoided.

The motivation for implementing this idea is to solve some issues of pipeline-based end-to-end approaches. The main inconvenience is that they are composed of two models that are trained specifically on independent tasks—that is, each document has to be manually labeled twice. First, bounding boxes and region classes have to be inserted to train the staff detection model. Once this information has been retrieved, the corresponding symbolic music representation must be written for each staff present in the dataset to train the staff-level recognition system. Therefore, labeling a corpus to train a production-ready OMR system for full-page transcription is costly—due to the need for manual segmentation and transcription of individual music staves—and time-consuming [14]. Another issue is that there has to be assumed an accumulated er-

© F. Author, S. Author, and T. Author. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0).
Attribution: F. Author, S. Author, and T. Author, “End-to-End Full-Page Optical Music Recognition for Mensural Notation”, in *Proc. of the 23rd Int. Society for Music Information Retrieval Conf.,* Bengaluru, India, 2022.

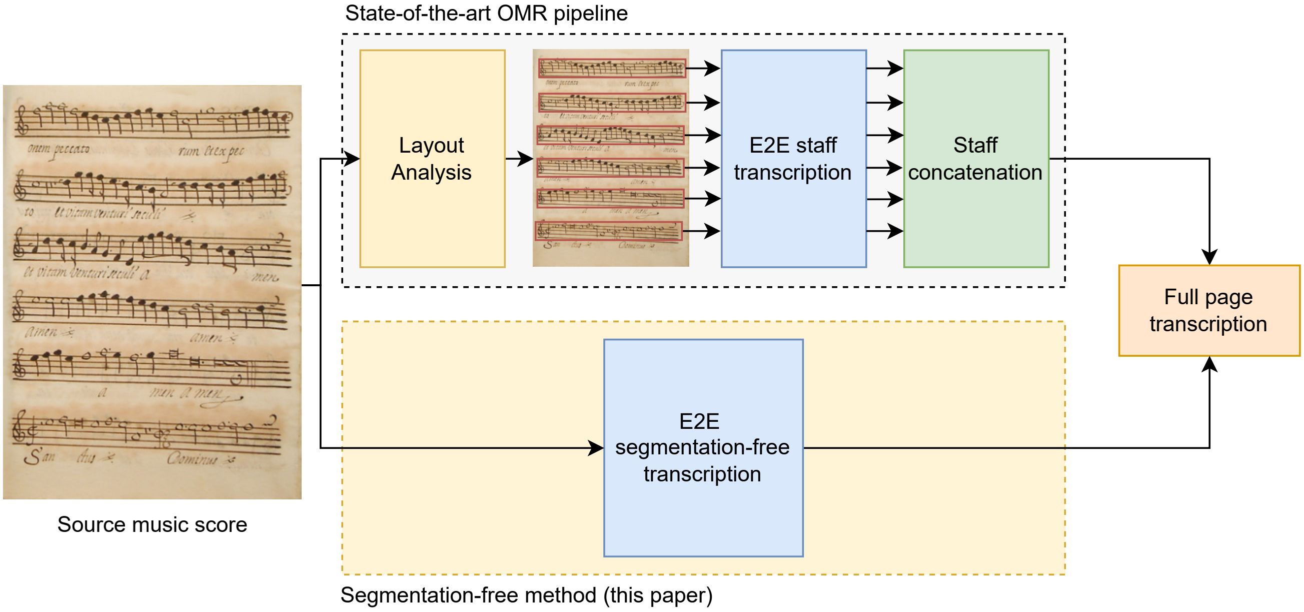
State-of-the-art OMR pipeline
Layout Analysis
E2E staff transcription
Staff concatenation
Full page transcription
Source music score
E2E segmentation-free transcription
Segmentation-free method (this paper)

Figure 1: General overview of the current full page OMR pipeline (top) in contrast to this paper proposal (bottom), where a previous layout analysis is not needed to transcribe the music symbols in the score.

ror that is carried over from one step to the next, as some works have studied [12, 15]. That is, the performance of a specific step is highly related to the effectiveness of the previous one, which is also dependent on their predecessors, thus creating a snowball effect.

2. EXTENDING OMR TRANSCRIPTION SYSTEMS

In this section, we explain how state-of-the-art end-to-end OMR models work, what issues they present when transcribing a complete music document, and how they can be adapted to a full-page scenario.

2.1 End-to-end staff transcription systems

The state-of-the-art end-to-end music staff transcription systems are neural models that directly output the music notation sequence of a given single-staff image. These networks are based on two main blocks: first, there is an encoder block, which filters an input image to learn and establish its relevant features. Commonly, this part is approached with Convolutional Neural Networks (CNN). Then, the output of this module is processed by the decoder block, which models temporal dependencies in the obtained features—which are represented as a vector sequence—to enhance results. This second module is commonly implemented with Recurrent Neural Networks (RNN). The whole model, referred to in the literature as a Convolutional Recurrent Neural Network (CRNN), is trained using the Connectionist Temporal Classification (CTC) [16] loss strategy, which maximizes through expectation-maximization the probability of obtaining a music sequence in a given notation alphabet. That is, the CTC method forces the network to align the output sequence to the available information extracted from the image, retrieving the full transcription in a single step. One important step to understand the model is how the output of the encoder is adapted to be processed as a sequence in the decoder. State-of-the-art implementations [10] reshape this obtained feature map—which is a 3D structure of size (h, w, c), where h is the height of the feature map, w is its width, and c is the number of filters (*channels*) obtained from the last convolutional layer—, by concatenating all the consecutive frames on the height axis of the image. That is, we obtain a sequence-like 2D structure that now has a shape of $(w, c \times h)$. This methodology has proven to be effective to transcribe both printed music staves [17] and handwritten ones [10, 11].

2.2 The issues with full-page images

The methodology described in the previous section is ideal for single-staff images, as it solves the transcription problem by reading the obtained feature map from left to right. However, in the case of page images, this formulation cannot be applied. The main issue with this formulation is that a single frame—which is an image column—contains more than one symbol. This is troublesome in two ways. First, transcription and retrieval of the score would result in a costly task, as ground-truth should be written and read from top to bottom and then from left to right, introducing undesirable post-processing methods to obtain a readable score. The second problem is related to the training of the system, as the CTC method should maximize the probability of non-consecutive temporal symbols in a few time steps. That is, the network has to classify symbols from different staves in the same image columns. This implicit interpretation is wrong, as these symbols are located in farther positions of the sequence, as they belong to different staves.