# SOURCE SEPARATION METHODS FOR COMPUTER-ASSISTED ORCHESTRATION

Luke DZWONCZYK (dz.luke@berkeley.edu) 1, Léo CHÉDIN<sup>1, 2, 3</sup>, Alejandro SALDARRIAGA-FUERTES<sup>1</sup>, Max SHERR<sup>1</sup>, Hélène-Camille CRAYENCOUR<sup>3</sup>, and Carmine-Emanuele CELLA<sup>1</sup>

### **ABSTRACT**

2

5

6

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

In this paper, we will study the possibility of adding 41 source separation as a pre-processing step to the computerassisted orchestration process. We first discuss the motivation of this addition and its potential to increase the 44 quality of orchestrations of multi-layered sounds. Second, 45 we select several state-of-the-art models for both music source separation (separation of instruments) and universal sound separation (separation of arbitrary sounds of different types), and compare their effectiveness for the task of orchestration. We assess which methods best suit the needs of orchestration by applying them on hand picked target sounds, orchestrating the separated outputs, and finally comparing them to the orchestration of the same target without separation. Our experiments show that the quality of orchestrations improves, both qualitatively and quantitatively, indicating that our approach is promising. Finally, we compare unsupervised methods to supervised methods for separation, and comment on the effect of training data selection on performance of supervised methods.

## 1. INTRODUCTION

Computer-assisted composition is a field that focuses on the creation of computational tools to aid in the musical composition process [1, 2]. Within this field, target-based computer-assisted musical orchestration is an example of how musical creativity can be supported by tools such as machine learning. Orchestration has been traditionally taught in an intuitive and non-formalized way, and computer-assisted orchestration seeks to help composers by accelerating certain creative processes. In particular, target-based computer-assisted orchestration helps composers explore new timbral possibilities by combining instrumental samples in such a way as to mimic the timbre of a given target sound. Formally, it is the process of creating an orchestral score that best matches an arbitrary target sound given a similarity metric and constraints [3].

Computer-assisted orchestrations can be divided into two categories: static and dynamic. Static orchestration con-

Copyright: © 2021 the Authors. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

siders the target as a single vector of timbral features and creates a single onset of note(s) as the solution. Dynamic orchestration considers a time series of features, which allows it to create an orchestration that evolves over time with multiple onsets. It uses time segmentation to cut the target in time into multiple segments, and orchestrates each segment.

Orchidea is a framework and set of tools that is currently considered the state-of-the-art for computer-assisted orchestration. It embeds the target in a feature space and uses a jointly-optimized heuristic and constraint solver to find a combination of samples that best match the target [4, 5]. Orchidea is able to do both static and dynamic orchestration. In order to better observe the effects of source separation, we perform only static orchestrations in this paper. The reasoning for this is that we wish to disentangle the two difficult problems of source separation and time segmentation. By only performing static orchestrations, we reduce the complexity of the problem by removing time segmentation and can better observe the effects of source separation.

Sound source separation is the process by which a single audio file is separated into multiple sound sources. A perfect separation is able to dissect a sound exactly into its constituent parts, in which each part is an independent sound event. Music source separation is a specific application of source separation in which the input audio is expected to be music that is comprised of a subset of instruments or sounds. For example, one music source separation method could attempt to split a song into three parts: voice, bass, and drums. Another aimed at orchestral music may try to separate the input into families of instruments: woodwinds, brass, strings, and percussion. In contrast, universal sound separation does not operate under the assumption that the input is of a musical nature, and can be applied to arbitrary sounds. In this case, the number of sources expected is specified, and the method will attempt to divide the input into the given number of sources.

Using source separation techniques as a pre-processing step has been shown to improve results in several music processing tasks. Source separation can be used to remove part of the spectrum in order to enhance the clarity of the features of interest for the task at hand. For instance, in the task of automatic chord estimation from audio, some authors have chosen to remove part of the spectrum related to percussive sounds to improve chord accuracy [6].

<sup>&</sup>lt;sup>1</sup>Center for New Music and Audio Technologies, University of California, Berkeley, Berkeley, California USA

<sup>&</sup>lt;sup>2</sup>École Normale Supérieure Paris-Saclay, Université Paris-Saclay, Paris, France

<sup>&</sup>lt;sup>3</sup>Laboratoire des Signaux et Systèmes, Centrale Supélec, CNRS, Université Paris-Saclay, Paris, France

Similarly some authors have investigated how beat tracking can be improved by using source separation as a preprocessing step for difficult songs with highly predominant expressive vocals [7]. Source separation can also be applied to separate different components of the signal and work on them separately to reduce the complexity of the task. For instance, in the context of tempo detection [8], the authors decompose the signal into sources to reduce rhythmic complexity, under the idea that some layers may be more rhythmically regular than the overall mix. Another example can be found in [9] where harmonic/percussive and solo/accompaniment source separation techniques are investigated as a pre-processing step to improve predominant instrument recognition in jazz music.

1

2

3

5

6

8

9

10

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

43

44

45

46

47

48

49

50

51

52

53

54

55

56

Source separation is an important addition to improve orchestration as it allows for the orchestration of more complex sounds in a way that makes sense in an orchestral context. Orchestral music often uses multiple layers, so by separating the layers of a target and individually orchestrating them, the resulting solution is improved.

For example, consider a target that has a continuous drone 58 sound throughout and a melody playing above this drone. Without source separation, Orchidea would detect each note of the melody as a new onset and cut the target in time at each new note. However, this segmentation in time would also effect the drone, cutting it in time and forcing it to have a new onset each time the melody changes notes. The resulting orchestration would have multiple onsets for the drone, even though only one onset is needed at the beginning of the drone. When source separation is applied to this target, the drone and melody would be orchestrated separately. Therefore, the drone could be orchestrated by a single onset that lasts for the duration of the drone. At 68 the same time, the melody could have as many new onsets as needed without affecting the drone. The orchestration would be greatly improved by removing the unnecessary onsets in the orchestration of the drone. Although this example is a dynamic orchestration, and we perform only static orchestrations, this paper represents a first step toward achieving dynamic orchestration with source separation.

The codebase for this paper can be found at: https://github.com/dzluke/smc2021, and you can listen to a selection of targets and orchestrations here: https://dzluke.github.io/smc2021/.

## 2. METHODOLOGY

We compare the effectiveness of different source separation methods for the task of orchestration by applying the separation methods to various targets, orchestrating the separated output of the method, and finally comparing this to the orchestration of the target if no separation was performed (see Fig. 1). Following the previous example, consider a target sound that is a combination of a low droning sound and high-pitched whistle. With separation, these two sounds would be disentangled into two separate subtargets and each would be individually orchestrated. The separated orchestrations would then be combined to create

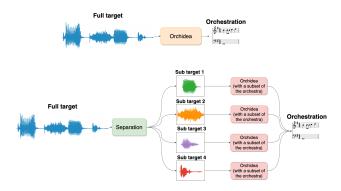


Figure 1. Diagram comparing the process of orchestration with and without separation. At the top, the full target orchestration is created simply by orchestrating the target. At the bottom, the separation process is performed, and the resulting sub-targets are individually orchestrated with subsets of the orchestra before being recombined to obtain the full orchestration.

the solution. This would then be compared to the orchestration if no separation had been performed.

#### 3. EXPERIMENTS

We test the effectiveness of five difference source separation methods, both supervised and unsupervised, on 300 custom made targets that are combinations of sounds from different databases. During testing, a target is input to a source separation method, which outputs four sub-targets. Then each sub-target is independently orchestrated with a randomly assigned subset of the full orchestra. These orchestrations are then combined to play simultaneously, creating a final orchestrated solution. Then the distance between the target and solution is calculated, giving us a metric to compare the various separation methods. The random split of the orchestra is the same for each separation method of the same target. See Fig. 2 for a diagram of this process.

# 3.1 Data

81

We created our own targets as combinations of four source sounds. The sources come from the NIGENS [10] and BBC [11] databases, and freesound.org [12]. We selected a total of 90 samples from these databases, choosing sounds that fit the following criteria:

- 1. Static sounds in which the harmonic average across time is a fitting representation of the sound
- 2. Sounds in which there is at least some pitched content and not only noise

The sounds chosen include alarms, bells, engine noises, soundscapes, synthesizer chords, and sound effects. Each target that we used for testing was a combination of four randomly chosen source sounds from the group of 90 sounds. When a target is created, the longest of the four sources is chosen to start playing at the beginning of the target. The other three sources are then randomly assigned

different times to begin playing such that they will start between the beginning of the target and half-way through the longest source sound.

## 3.2 Separation Methods

We consider source separation models trained to tackle two 5 distinct separation problems. The first is music source sep-6 aration (Open-Unmix, Demucs, ConvTasnet). For these 7 architectures, we used models that were pre-trained on the MUSDB18 dataset [13]. Thus, these models output 9 4 stereo tracks that correspond to broad instrumental cat-10 egories defined in the SiSEC 2018 Mus evaluation campaign: vocals, drums, bass and other [14]. Although the 12 audio sources we consider in this work are often non-13 musical, we hope that these source separation methods can 14 still improve the orchestration by separating other types of 15 sound in a musically meaningful way. The second type of 16 model we tested are ones trained for universal sound sepa-17 ration (TDCN++, NMF), which [15] defines as separating 18 mixtures of arbitrary sounds of different types. 19

# 3.2.1 Music source separation

20

21

22

23

24

25

26

27

28

30

31

32

33

34

35

36

37

38

39

41

42

43

44

45

46

47

48

49

50

51

52

53

54

55

56

Open-Unmix [16] is an open source implementation of a deep music source separation model, based on [17]. It is actually composed of multiple models that are trained for each instrumental target. Each of these models is trained on the specific target using the same architecture, based on a three layer bidirectional LSTM network. Open-Unmix operates on the Short Time Fourier Transform (STFT) of the input mixture, which is first standardized. It predicts the target by multiplying the output of the LSTM network with the magnitude spectrogram of the input, essentially applying a mask. Open-Unmix also uses Wiener filtering as a last processing step.

The **Demucs** [18] architecture is a deep learning model for musical source separation. It takes as input the stereo mixture, and outputs 4 stereo waveforms corresponding to 92 the four categories: vocals, bass, drums and other. De-93 mucs is based on a convolutional encoder, a bidirectional 94 LSTM, and a convolutional decoder. The encoder and the 95 decoder are linked with skip U-Net connections. Both the 96 encoder and the decoder are made of 6 blocks, each made 97 of a combination of a convolutional layer with kernel size 98 8, followed by ReLU activation. Finally there is another 99 convolution with kernel size 1, followed by gated linear 100 units (GLU). The U-network structure allows the decoder 101 to directly access the input signal and to transfer its phase, 102 improving results.

We also used the Conv-Tasnet [19] implementation <sup>103</sup> of [18], who adapted the architecture, originally designed <sup>104</sup> for speech separation at 8 kHz, for the task of music source <sup>105</sup> separation. Conv-Tasnet is a model based on a convolutional encoder-decoder, with a time-dilated convolutional <sup>106</sup> network separation module, estimating masks based on the encoder output. This is why we refer to it as **TDCN**, in accordance with [15]. It uses the audio waveform as input <sup>109</sup> and performs a series of 1-D convolutions at different resolutions with residual connections in order to generate a <sup>110</sup> pool of masks that are then applied on the embedding of <sup>111</sup>

the initial waveform to output the separated targets. A final 1-D convolutional layer is used as a decoder to recover the waveforms of each sub target. This type of architecture replaces RNNs in modeling sequential data, as series of dilated convolutions are a good way to capture temporal patterns.

### 3.2.2 Universal sound separation

62

Non-negative matrix factorization [20,21] is an unsupervised learning method, which takes an input non-negative matrix, in this case a spectrogram S, in  $\mathbb{R}^{N\times M}$  and decomposes it into a basis matrix A in  $\mathbb{R}^{N\times J}$  and a component matrix X in  $\mathbb{R}^{J\times M}$  such that  $S\approx A\times X$ . We use the coordinate descent algorithm from the scikit-learn Python package based on [20] with J set to 4, which selects for 4 features in the dataset. We then output 4 mono waveforms containing each isolated feature  $F_i$ , where  $F_i=A_i\times X_i$  for i=1,2,3, or 4.

TDCN++, a version of Conv-Tasnet, was improved by [15] in order to be used for general sound separation, and not only for speech. The main architecture remains the same as in [19], and the changes only affect the masking network. More specifically, feature-wise normalization replaces global normalization, residual connection ranges are extended to improve the flow of information, and learnable scale parameters are introduced to weight the outputs of each residual layer. Those modifications help the model to better extract features without drastically increasing its size. TDCN++ presents the interesting property of having almost the same architecture as TDCN but trained with generic sounds. Thus, it should naturally be better suited for our task, and the comparison between those two models in particular can give good insights about the impact of the data used for training.

## 3.3 Testing

In order to compare the effectiveness of the separation methods, a full target orchestration and a ground truth orchestration are done for each target. The full target orchestration is the orchestration of the target without any separation performed. The ground truth orchestration takes the four sources that make up the target and orchestrates each one separately, then combines them. This ground truth represents the orchestration that could take place if the separation method was perfect and separated a target exactly into its constituent sources.

The procedure we use for testing is as follows:

- The target is created as a combination of four randomly chosen sources, which are offset to begin playing at different times
- 2. The full target, without any separation performed, is orchestrated using the entire orchestra
- 3. Each separation method is performed on the target, each splitting the target into four sub-targets
- 4. The full orchestra is randomly split into four equal sized subsets

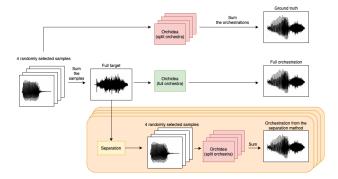


Figure 2. Diagram showing the testing process. Note that in both orchestrations that use a split orchestra, the split is the same.

1

2

5

6

8

9

10

11

12

13

14

15

16

17

18

19

20

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

- 5. For each method, the four sub-targets are separately orchestrated, each using a different subset of the orchestra, and then each orchestration is combined to play simultaneously
- The ground truth orchestration is created by orchestrating the four sources, using the same orchestra splits from step 4, and then combined to play simultaneously
- The distances between target and orchestration are calculated for the full target orchestration, the separated orchestration, and the ground truth orchestration

The orchestrations performed are static orchestrations, meaning a single onset of notes is created for each target, no matter if the target itself has multiple onsets. The OrchideaSOL database of orchestral samples is used with Orchidea to create the orchestrations [22]. This database contains recordings of extended playing techniques, which better fits the often noisy or inharmonic nature of our targets. The full orchestra, of which non-overlapping subsets were selected to orchestrate the subtargets, contains the following instruments: 8 violins, 4 violas, 3 cellos, 1 bass, 2 oboes, 2 flutes, 2 clarinets in Bb, 2 bassoons, 2 trumpets, 2 trombones, and 2 French horns.

As described in Section 3.1, each target is comprised of four sources that are offset in time to begin playing at different times. Figure 3 illustrates the effect this offset has on the full target, separated, and ground truth orchestrations. The full target orchestration places all onsets at the beginning of the orchestration, since it is unable to differentiate between different sources, it cannot place the different orchestrations at the appropriate offsets. However, the separated orchestrations are able to correctly place the onsets of the sub-target orchestrations so far as the method has correctly separated out the sources. In the ground truth orchestration, the sources are offset by the same amount as they are in the creation of the target, leading to the ground truth having each source orchestration beginning at the correct time. For this reason, despite only performing static orchestrations, the resulting separated and ground truth orchestrations are offset in time compared to the full target orchestration.

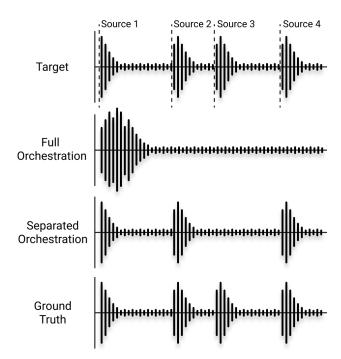


Figure 3. Example waveforms showing onsets of the four sources in the target, and possible onsets in the full, separated, and ground truth orchestrations. In this example, the separation method could not distinguish between sources 2 and 3 and identified them as a single sub-target. Therefore, the separated orchestration has only 3 onsets, one of them being an orchestration that combines sources 2 and 3. Note that this is a simplified representation, in the actual targets sources overlap in time and are not as discrete as this diagram shows.

# 3.4 Evaluation

We compare the effectiveness of different separation methods by comparing how well they work for orchestration. The output of a method is orchestrated, and these orchestrations are compared. A quantitative evaluation is performed through the use of a distance metric that measures the spectral distance between target and solution.

The distance metric cuts the target and solution into successive frames that are 4,000 samples in length, which is approximately 90 milliseconds at a sampling rate of 44.1 kHz. The spectral distance, as defined in Eqn. 1, is calculated between corresponding frames of the target and solution. The spectral distance metric is proposed in [5] as part of the cost function used in Orchidea during the optimization, and used in [23] to compare accuracies of orchestrations. The equation takes in the full FFT of the target x and full FFT of the solution  $\tilde{x}$ . Then for each bin k of the FFT, it calculates the absolute difference between the values. The differing values of  $\lambda_1$  and  $\lambda_2$  allow the metric to penalize the solution in different ways. For our purposes, we used  $\lambda_1=0.5$  and  $\lambda_2=10$ , which penalizes a solution that overshoots the harmonic energy of the target.

$$d(x,\tilde{x}) = \lambda_1 \sum_{k} \delta_{k1}(x_k - \tilde{x}_k) + \lambda_2 \sum_{k} \delta_{k2}|x_k - \tilde{x}_k|$$
 (1)

	Average distance
Demucs	25.11
Open-Unmix	24.48
TDCN	26.44
NMF	22.44
TDCN++	24.38
Full target	24.96
Ground truth	24.19

Table 1. Average across 300 targets showing the distance between target and orchestration for various methods. "Full target" means no separation was performed. Lower values are better.

where  $\delta_{k1}=1$  if  $x_k\geq \tilde{x}_k,0$  otherwise; and  $\delta_{k2}=1$  if  $x_k<\tilde{x}_k,0$  otherwise.

#### 3 3.5 Results

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

The average distances for full target, separated, and ground truth orchestrations are displayed in Table 1. These re-5 sults are the average distance between target and orchestration for 300 different targets. We found that NMF performed the best out of all separation methods, and even 8 outperformed the ground truth orchestrations, on average. TDCN++, Open-Unmix, and NMF performed better than 10 the full target orchestrations, meaning that separating tar-11 gets using these methods leads to a better orchestration 12 than if no separation is applied. 13

# 4. DISCUSSION

In this paper, we addressed the problem of using source 73 separation as a pre-processing step for computer-assisted 74 orchestration. The results show that performing source 75 separation followed by a static orchestration of each individual sub-target improved the spectral distance between the original target and its orchestrated version in three out 77 of five of the methods tested. This shows that source 87 separation is a useful pre-processing stage for computer 78 assisted orchestration.

Furthermore, the one unsupervised method we tested, 80 NMF, outperformed all the supervised methods. This suggests that methods like NMF can still be useful for specific problems when available data is limited. It is logical that unsupervised methods will perform well for problems in which there is limited data, as they require less data than supervised methods. NMF also outperformed the ground truth orchestrations, which could be due to the relative simplicity of NMF's output subtargets making subsequent orchestration easier for Orchidea, but more experimentation would need to be done in order to confirm this. It is important to notice that for the supervised methods, we used pre-trained versions of the neural models. All of these models, except for TDCN++, were trained on musical data. However, we tested them on a wide range of targets, many of which were not strictly musical. This could explain why Demucs, Open-Unmix, and TDCN performed worse than NMF, and why TDCN++, which was trained on 94

arbitrary sounds, outperformed the other supervised methods. Therefore, we may infer from our results that models trained on generic sounds generate a better separation for orchestration when compared to models trained only on musical sounds.

Since unsupervised methods perform well, it is reasonable to think that supervised methods trained on a dataset of arbitrary sounds could yield better overall performance for orchestration.

#### 4.1 Future Work

42

43

44

45

46

47

58

71

As we stated previously, we believe that one of the reasons the supervised methods did not perform as well is because of the data they were trained on. Therefore, a natural next step is to train these models with sounds that are more appropriate for computer-assisted orchestration.

Another aspect we would like to work on is the assignment of the orchestras. Currently, the subset of the orchestra that is assigned to each sub-target is randomly selected. However, it is possible that this random assignment could reduce the quality of the final orchestration. If, for example, one of the sub-targets contains mostly low pitch content, but the orchestra assigned to it contains mostly violins, violas, flutes, and oboes, the resulting orchestration will suffer. To solve this problem, a step could be added that jointly-optimizes the orchestras assigned to each subtarget, ultimately choosing an assignment that leads to the best orchestration of each sub-target.

As we stated in Section 1, we chose static orchestration instead of dynamic orchestration to avoid entangling the problems of time segmentation and source separation. Future work could include applying source separation to dynamic targets and performing dynamic orchestration.

Finally, one or multiple of the separation methods we tested should be implemented in Orchidea in order to improve the orchestrations Orchidea is capable of creating.

## 5. REFERENCES

- [1] J. Fernandez and F. Vico, "Ai methods in algorithmic composition: A comprehensive survey," *Journal of Artificial Intelligence Research*, vol. 48, p. 513–582, Nov 2013. [Online]. Available: http://dx.doi.org/10.1613/jair.3908
- [2] C. Ariza, "Navigating the landscape of computer aided algorithmic composition systems: a definition, seven descriptors, and a lexicon of systems and research," in *ICMC*, 2005.
- [3] Y. Maresz, "On computer-assisted orchestration," *Contemporary Music Review*, vol. 32, 02 2013.
- [4] C.-E. Cella and P. Esling, "Open-source modular toolbox for computer-aided orchestration," in *Timbre conference*, 2018.
- [5] C.-E. Cella, "Orchidea: a comprehensive framework for target-based assisted orchestration," *submitted to Journal of New Music Research, under review*, 2020.

- [6] J. Reed, Y. Ueda, S. M. Siniscalchi, Y. Uchiyama, 51 [19] Y. Luo and N. Mesgarani, "Tasnet: 1 S. Sagayama, and C.-H. Lee, "Minimum classification 52 2 error training to improve isolated chord recognition." 3 in ISMIR, 2009, pp. 609-614.
- [7] J. R. Zapata and E. Gómez, "Improving beat tracking in 55 [20] 5 the presence of highly predominant vocals using source 56 6 separation techniques: Preliminary study," in Proc. 9th 57 7 Int. Symposium on Computer Music Modeling and Re- 58 8 trieval, London. Citeseer, 2012, pp. 583-590.
- [8] P. Chordia and A. Rae, "Using source separation to im-10 prove tempo detection." in *ISMIR*, 2009, pp. 183–188. 11
- [9] J. S. Gómez, J. Abeßer, and E. Cano, "Jazz solo in-12 strument classification with convolutional neural net-13 works, source separation, and transfer learning." in IS-14 MIR, 2018, pp. 577-584. 15
- 16 [10] I. Trowitzsch, J. Taghia, Y. Kashef, and K. Obermayer, "Nigens general sound events database," Feb. 2019. 17 [Online]. Available: https://doi.org/10.5281/zenodo. 18 2535878 19
- 20 [11] BBC, "BBC sound effects archive," accessed: 2021-03-03. 21
- 22 [12] F. Font, G. Roma, and X. Serra, "Freesound technical demo," in ACM International Conference on Mul-23 timedia (MM'13), ACM. Barcelona, Spain: ACM, 24 21/10/2013 2013, pp. 411-412. 25
- 26 [13] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "Musdb18-a corpus for music separation," 27 28 2017.
- 14] F.-R. Stöter, A. Liutkus, and N. Ito, "The 2018 signal 29 separation evaluation campaign," 2018. 30
- [5] I. Kavalerov, S. Wisdom, H. Erdogan, B. Patton, K. W. 31 32 Wilson, J. L. Roux, and J. R. Hershey, "Universal sound separation," CoRR, vol. abs/1905.03330, 2019. 33 [Online]. Available: http://arxiv.org/abs/1905.03330 34
- F.-R. Stöter, S. Uhlich, A. Liutkus, and Y. Mitsufuji, 35 I "Open-Unmix - A Reference Implementation for 36 Music Source Separation," Journal of Open Source 37 Software, vol. 4, no. 41, p. 1667, Sep. 2019. [Online]. 38 Available: https://hal.inria.fr/hal-02293689 39
- 40 [17] S. Uhlich, M. Porcu, F. Giron, M. Enenkl, T. Kemp, N. Takahashi, and Y. Mitsufuji, "Improving music 41 source separation based on deep neural networks 42 through data augmentation and network blending," in 43 2017 IEEE International Conference on Acoustics, 44 Speech and Signal Processing (ICASSP), 2017, pp. 45 261-265. 46
- 47 [18] A. Défossez, N. Usunier, L. Bottou, and F. R. Bach, "Music source separation in the waveform domain," 48 CoRR, vol. abs/1911.13254, 2019. [Online]. Available: 49 http://arxiv.org/abs/1911.13254 50

- ideal time-frequency masking for speech separation,' CoRR, vol. abs/1809.07454, 2018. [Online]. Available: http://arxiv.org/abs/1809.07454
- A. Cichocki and A.-H. Phan, "Fast local algorithms for large scale nonnegative matrix and tensor factorizations," IEICE Transactions, vol. 92-A, pp. 708-721, 03 2009.
- 59 [21] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the beta-divergence," 2011.
- C. E. Cella, D. Ghisi, V. Lostanlen, F. Lévy, J. Fineberg, and Y. Maresz, "Orchideasol: a dataset of extended instrumental techniques for computer-aided orchestration," 2020.
- C. E. Cella, L. Dzwonczyk, A. Saldarriaga-Fuertes, H. Liu, and H.-C. Crayencour, "A Study on Neural Models for Target-Based Computer-Assisted Musical Orchestration," in 2020 Joint Conference on AI Music Creativity (CSMC + MuMe), ser. Proceedings of the 2020 Joint Conference on AI Music Creativity, Stockholm, Sweden, Oct. 2020. [Online]. Available: https://hal.archives-ouvertes.fr/hal-03059955

70