END-TO-END SOUND SOURCE SEPARATION CONDITIONED ON INSTRUMENT LABELS

Olga Slizovskaia[⋆†] *Leo Kim*[⋆‡] *Gloria Haro*[†] *Emilia Gomez*[†]

[†] Pompeu Fabra University
[‡] University of Waterloo

ABSTRACT

Can we perform an end-to-end music source separation with a variable number of sources using a deep learning model? This paper presents an extension of the Wave-U-Net [1] model which allows end-to-end monaural source separation with a non-fixed number of sources. Furthermore, we propose multiplicative conditioning with instrument labels at the bottleneck of the Wave-U-Net and show its effect on the separation results. This approach can be further extended to other types of conditioning such as audio-visual source separation and score-informed source separation.

Index Terms— Sound Source Separation, End-to-End Deep Learning, Wave-U-Net

1. INTRODUCTION

The goal of music source separation is to extract the mixture of audio sources into their individually separated source tracks. Undoubtedly, this is a challenging problem to solve and many attempts have been made to estimate the source signals as closely as possible from the observation of the mixture signals. The most common cases may vary with respect to the target task (such as singing voice [2, 3] or multi-instrument source separation [4, 5, 6]), use of additional information (blind [5, 3] or informed source separation [4, 7, 6]), and the amount of channels used for reconstruction (monaural [5] or multi-channel [4, 7] source separation).

There are many challenging aspects related to audio source separation. Most importantly, accurate separation with minimal distortion is desired. Supplementary information such as the number of sources present in the mix, musical notes in the form of MIDI or sheet music can be helpful but not widely available in most cases. However, information such as the source instrument labels can be easily found from video recordings of musical performances readily available on the web. Therefore, it sounds reasonable to learn to integrate the instrument label information into the source separation pipeline. At the same time, many sophisticated score- and timbre-informed methods have been proposed in the literature already [2]. We admire the idea of simplifying those frameworks, which became possible only recently with the advent of end-to-end deep neural networks.

In this paper, we study how to separate musical recordings of small ensembles (from duets to quintets) into individual audio tracks. We propose an extension of the Wave-U-Net [1], an end-to-end convolutional encoder-decoder model with skip connections, which supports a non-fixed number of sources and takes advantage of instrument labels in assisting source separation.

2. RELATED WORK

Traditionally, people have attempted to solve audio source separation through matrix-factorization algorithms. Independent Component Analysis (ICA) [8] and Non-negative Matrix Factorization (NMF) [9] are two common techniques used for source separation.

With the recent achievements in machine learning, researchers have started to adopt deep neural network paradigms to address the source separation problem. Since CNNs have been proven to be successful in image processing, raw audio data is often converted to 2D spectrogram images for analysis. The image data is then fed to a convolutional autoencoder which generates a set of masks that can be used to recover sound sources using inverse Short Time Fourier Transform (STFT) [3, 5, 10].

In this paper, we aim to continue researching on deep learning methods for the source separation problem. Furthermore, we focus on improving the results by experimenting with less conventional approaches. Primarily, we work directly with raw waveforms as opposed to time-frequency image representation. This approach is an active research area [1, 11] and gives us an additional advantage of preserving the phase information unlike other CNNs which only use the magnitude of STFT [5, 10]. Secondly, we want to enhance our results through conditioning with instrument label information. This type of guidance has been shown to have a good impact on the source separation performance. Thus, in [12], the authors use visual guidance for improving source separation quality. Additionally, in a concurrent work [13], the authors explore a similar idea of class-conditioning over the joint embedded space, but unlike us, they use an auxiliary network to model parameters of a GMM for the final source sep-

⋆ Equal contribution. Work done during the Deep Learning Camp Jeju 2018.

aration, and they take spectrograms as an input of the model.

Wave-U-Net model [1] is an adaptation of the U-Net [14], a convolutional encoder-decoder network developed for image segmentation. The U-Net approach has been adapted already for singing voice separation in [3], however this model applies 2D convolutions and works with spectrograms. Instead of doing a 2D convolution, Wave-U-Net performs series of 1D convolutions, downsampling and upsampling with skip connections on a raw waveform signal. This approach was presented at SiSEC evaluation campaign [15] and demonstrated competitive performance.

The input to this network is a single channel audio mix, and the desired output is the separated K channels of individual audio sources, where K is the number of sources present in the audio mix. An interesting aspect of the Wave-U-Net is that it avoids implicit zero paddings in the downsampling layers, and it performs linear interpolation as opposed to de-convolution. This means that our dimension size is not preserved, and our output results will actually become a lot shorter compared to our input. However, by doing this we can better preserve temporal continuity and avoid audio artifacts in the results.

3. EXPANDED WAVE-U-NET

3.1. Multi-Source Extention

The challenge with the original Wave-U-Net model is that it can only support a predefined number of input sources (2 and 4 sources in the original settings), limiting its application to only the specific group of instruments that it was trained on. We aim to build a more flexible model that can support a dynamic number of input sources and, therefore, be more suitable for separating classical music recordings. In classical music, the number of instruments playing in an ensemble may vary a lot but the instruments themselves are often known in advance. Here we don't tackle the problem of separating different parts played by the same instrument (like violin1 vs violin2) but rather try to separate a sound track played by the same instrument (violin1+violin2 vs viola). Therefore, we can fix a maximum number of output sources to the number of all different instruments which are present in the dataset. This is still not a true dynamic model since the number of sources has to be specified in advance. Thus, in order to have a more general model we fix the number of sources to a reasonable large number.

For the sources that are not available in the mix, the model is trained with silent audio as a substitute. Therefore, the model outputs all possible sources and it is forced to associate each output with a certain instrument and output silence for the sources that are not present in the mix. Note that at the training time we implicitly specify which source should be aligned with a particular instrument, but it is not needed at the inference time. We can instead use an energy threshold for extracting the sources of interest. We will refer to this model as *Exp-Wave-U-Net*.

3.2. Label Conditioning

In order to enhance the source separation results, we propose a conditioned label-informed Wave-U-Net model (*CExp-Wave-U-Net*). In particular, we use a binary vector whose size is the maximum number of sources considered. Each position of the vector is associated with a certain instrument: 1 indicates that the instrument is being played and 0 indicates either a non present instrument or a silent instrument (non-playing).

Conditioning is a term used to describe the process of fusing information of a different medium in the context of another medium. In case of Wave-U-Net, there are three locations where the use of conditioning is appropriate and corresponds to different fusion strategies:

- for early fusion, the conditioning can be applied to the top layer of the encoder, before downsampling;
- for middle fusion, we can integrate label information at the bottleneck of the Wave-U-Net;
- for late fusion, we can aggregate labels with audio output of the last decoder layer (after upsampling).

Moreover, there is a possibility of using several conditioning mechanisms (as described in [16]) such as

- concatenation-based conditioning;
- conditional biasing (additive bias);
- conditional scaling (multiplicative bias).

In this paper, we experiment with multiplicative conditioning using instrument labels at the bottleneck of the Wave-U-Net model. Therefore, the overall idea is to cancel out the unwanted sources at the most compressed part of Wave-U-Net while emphasizing the sources of interest. Even though the early fusion approach can be more abundant as it allows to integrate more information from the very beginning, we use multiplicative middle fusion as it provides a reasonable trade-off between expressiveness and memory and computational costs. At the same time, we leave additive bias and concatenation-based conditioning for further investigation.

4. IMPLEMENTATION DETAILS AND RESULTS

4.1. Dataset

As described in Sec. 3, the model takes the input in a form of a mix of the output sources where each source is either an instrumental track or a silent audio track for instruments not present in the mix. Instrument labels can be included optionally. We took advantage of the University of Rochester Musical Performance Dataset (URMP) [17] which consists of 44 pieces (11 duets, 12 trios, 14 quartets and 7 quintets) played by 13 different instruments (see Figure 1). We used 33 pieces for training and validation, and 11 pieces for testing.

4.2. Baseline

For the evaluation, we compare two proposed models with a Timbre-Informed NMF method from [7]. In this method, the authors first learn a timbre model for each note of each instrument, and apply this trained templates as the basis functions in NMF factorization procedure. Note that the timbre templates are trained with RWC [18], a dataset which consists of recordings of individual notes for different instruments. Unlike our approach, Timbre-Informed NMF requires specifying the timbre models for each piece at the inference time. We used learned timbre models for all instruments except for saxophone.

4.3. Implementation Details

Our implementation is available online[1] and is based on the original Wave-U-Net code[2]. We improved both input and training pipelines compared to the original work. The input pipeline is implemented as a TensorFlow Dataset and now supports parallel distributed reading. The training pipeline is re-implemented via a high-level TensorFlow Estimator API and supports both local and distributed training. Our implementation also supports half-precision floating-point format, which allows us to increase both training speed and batch size without loss of quality.

We train the model on a single Google Cloud TPU instance for 200k steps which takes approximately 23 hours. The best results are achieved using Adam optimizer with an initial learning rate of 1e-4. The aforementioned modifications together with the use of TPU allowed us to speed up training process by 24.8 times (35.3 times for the half-precision case) compared to a single GPU training.

4.4. Results

We perform quantitative evaluation of the model performance using standard metrics for blind source separation: *Source to Distortion Ratio* (SDR), *Source to Inference Ratio* (SIR), and *Source to Artifacts Ratio* (SAR) [19].

Table 1 shows average values of the metrics over all pieces and instruments in the dataset. We can see that there is no single winner but each method seems to be better with respect to one of the metrics. For example, InformedNMF baseline outperforms both deep models in terms of SDR while it is inferior to Exp-Wave-U-Net in terms of SAR and to CExp-Wave-U-Net in terms of SIR. Note that we can't directly compare our results with Wave-U-Net because it would require to train from 3 to 11 different models while for Exp-Wave-U-Net we just train a single model for all instruments and number of sources.

Next, we analyze the separation performance in depth for each instrument. Figure 1 summarizes the results for each

[1] https://github.com/Veleslavia/vimss

[2] https://github.com/f90/Wave-U-Net

Method	SDR	SIR	SAR
InformedNMF [7]	**-0.16**	1.42	9.31
Exp-Wave-U-Net	-4.12	-3.06	**12.18**
CExp-Wave-U-Net	-1.37	**2.16**	6.36

Table 1. URMP [17] dataset: SDR, SIR and SAR for different methods averaged over the testing set. Best values are shown in bold. Exp-Wave-U-Net states for an extension of Wave-U-Net with multiple output sources, CExp-Wave-U-Net states for a version of Exp-Wave-U-Net conditioned by labels of the instruments.

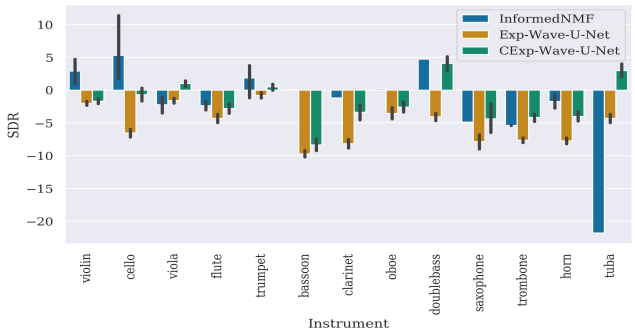
Model	nSources	SDR	SIR	SAR
InformedNMF[7]	2	3.08	4.98	10.55
	3	0.07	1.69	9.01
	4	-3.84	-2.62	8.65
Exp-Wave-U-Net	2	-0.42	1.75	10.98
	3	-3.85	-2.74	11.97
	4	-5.90	-5.33	12.87
CExp-Wave-U-Net	2	-0.16	4.62	7.48
	3	-0.68	2.88	5.91
	4	-2.56	0.44	6.35

Table 2. SDR, SIR and SAR for different methods averaged with respect to the number of sources in the mix.

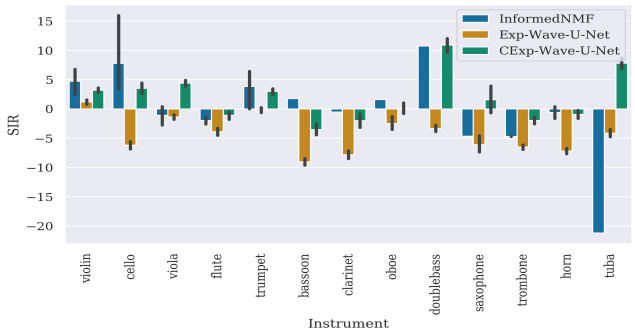
model and metric. We can see that the baseline approach (InformedNMF) performs reasonable well in terms of SDR and SIR for all instruments except for trombone and tuba. Exp-Wave-U-Net performs worse in SDR and SIR for all instruments but consistently outperforms the baseline and CExp-Wave-U-Net in SAR except for violin, trombone and saxophone. CExp-Wave-U-Net performs as good as the rest two in SDR and SIR (and achieves best results for tuba, double-bass, saxophobe and viola) but consistently worse in SAR.

At last, we report the separation results averaged with respect to the number of sources in the input mix in Figure 2. It is interesting to note that the performance of all methods goes down as the number of sources increases. Hovewer, it is more interesting that the performance of CExp-Wave-U-Net does not drop as much as in case of InformedNMF and Exp-Wave-U-Net. In absolute values (see Table 2), SDR for CExp-Wave-U-Net decreases from -0.16 dB to -2.56 dB while for the model without conditioning those values are -0.42 dB to -5.90 dB, and from 3.08 dB to -3.84 dB for the NMF baseline. The alike behaviour persists for SIR. From these results, we could anticipate that the conditioned model is more suitable for multi-instrument source separation.

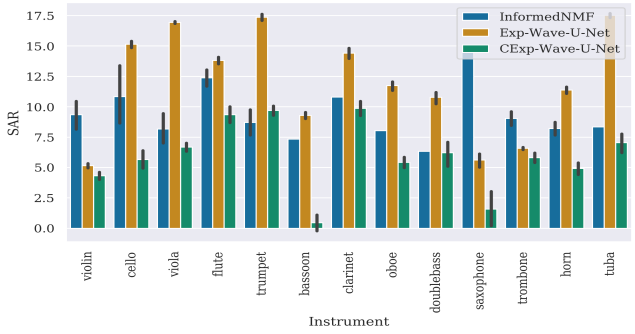
We would like to mention that despite their widespread use, the standard metrics are unable to estimate how well the model can discard unwanted sources (they are undefined if



(a) SDR (dB)



(b) SIR (dB)



(c) SAR (dB)

Fig. 1. Results in terms of SDR, SIR, and SAR for each instrument in the testing set of URMP [17] dataset.

the ground truth is silence). Nonetheless, we would like to provide samples of separated sources which should be discarded[3]. We notice that both conditioned and unconditioned versions of Exp-Wave-U-Net systematically output quieter sources for the absent instruments than InformedNFM, initialized by all possible timbre templates.

Some qualitative results for original[4] and expanded[5] Wave-U-Net can be also found online.

[3] https://goo.gl/e18F41
[4] https://youtu.be/mGfhgLt1Ds4
[5] https://youtu.be/mVqIMXoSDqE

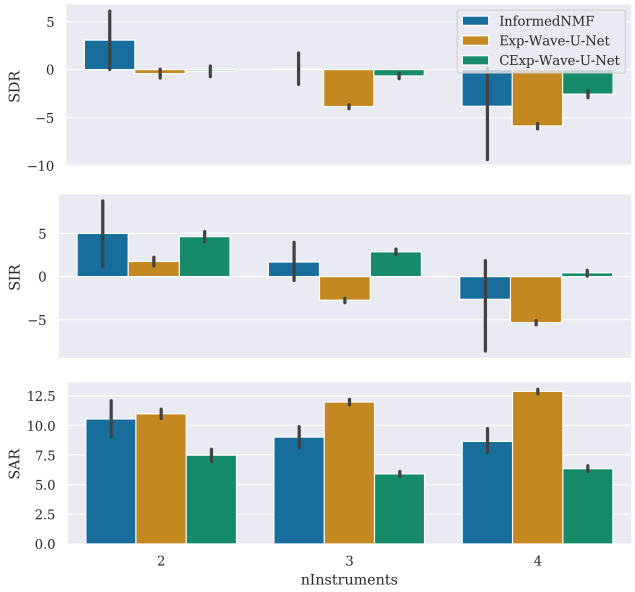


Fig. 2. Results in terms of SDR, SIR, and SAR averaged and reported by the number of instruments in the testing set of URMP [17] dataset.

5. CONCLUSION

In this paper we have proposed and explored two extensions of the Wave-U-Net architecture in the context of source separation of ensemble recordings with unknown number of input sources. We have shown that both Exp-Wave-U-Net and CExp-Wave-U-Net perform fairly competitive to the InformedNMF model despite being trained just on 33 audio mixes. We observed that CExp-Wave-U-Net outperforms the baseline approach when the number of input sources is bigger than 2. Moreover, we observed that Exp-Wave-U-Net produces a quieter output for the non-present instruments.

We plan to further experiment with different fusion models for conditioning and incorporate visual information available within URMP dataset. The visual guidance seems to be a prominent direction of research because in this case not only we do not need to have manually annotated instrument labels but can also get an additional information of the playing and non-playing state of each instrument by analyzing the corresponding video stream. This can be especially useful to resolve disambiguation and inference between two instruments of the same kind.

Acknowledgements. We would like to thank Terry Um and Eric Jang for their support during the camp, and Marius Miron for providing the code for the baseline. This work has received funding from the Maria de Maeztu Programme (MDM-2015-0502), ERC Innovation Programme (grant 770376, TROMPA), and MINECO/FEDER UE project (TIN2015-70410-C2-1-R).

6. REFERENCES

[1] D. Stoller, S. Ewert, S. Dixon, et al., "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," 19th International Society for Music Information Retrieval Conference (ISMIR), 2018.

[2] Zafar Rafii, Antoine Liutkus, Fabian-Robert Stoter, Stylianos Ioannis Mimilakis, Derry FitzGerald, and Bryan Pardo, "An overview of lead and accompaniment separation in music," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 26, no. 8, pp. 1307–1335, Aug. 2018.

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

[4] M. Miron, J. J. Carabias-Orti, J. J. Bosch, E. Gómez, and J. Janer, "Score-informed source separation for multichannel orchestral recordings," *Journal of Electrical and Computer Engineering*, vol. 2016, 2016.

[5] Pritish Chandna, Marius Miron, Jordi Janer, and Emilia Gómez, "Monoaural audio source separation using deep convolutional neural networks," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2017, pp. 258–266.

[6] Yushen Han and Christopher Raphael, "Informed source separation of orchestra and soloist.," in *ISMIR*, 2010, pp. 315–320.

[7] J. J. Carabias-Orti, M. Cobos, P. Vera-Candeas, and F. J. Rodríguez-Serrano, "Nonnegative signal factorization with learnt instrument models for sound source separation in close-microphone recordings," *EURASIP Journal on Advances in Signal Processing*, vol. 2013, no. 1, pp. 184, 2013.

[8] Aapo Hyvärinen and Erkki Oja, "Independent component analysis: algorithms and applications," *Neural networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[9] Tuomas Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE transactions on audio, speech, and language processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[10] Stefan Uhlich, Marcello Porcu, Franck Giron, Michael Enenkl, Thomas Kemp, Naoya Takahashi, and Yuki Mitsufuji, "Improving music source separation based on deep neural networks through data augmentation and network blending," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 261–265.

[11] Francesc Lluis, Jordi Pons, and Xavier Serra, "End-to-end music source separation: is it possible in the waveform domain?," *arXiv preprint arXiv:1810.12187*, 2018.

[12] Sanjeel Parekh, Slim Essid, Alexey Ozerov, Ngoc QK Duong, Patrick Pérez, and Gaël Richard, "Guiding audio source separation by video object information," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2017 IEEE Workshop on*. IEEE, 2017, pp. 61–65.

[13] Prem Seetharaman, Gordon Wichern, Shrikant Venkataramani, and Jonathan Le Roux, "Class-conditional embeddings for music source separation," *arXiv preprint arXiv:1811.03076*, 2018.

[14] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[15] Fabian-Robert Stöter, Antoine Liutkus, and Nobutaka Ito, "The 2018 signal separation evaluation campaign," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2018, pp. 293–305.

[16] Vincent Dumoulin, Ethan Perez, Nathan Schucher, Florian Strub, Harm de Vries, Aaron Courville, and Yoshua Bengio, "Feature-wise transformations," *Distill*, 2018, https://distill.pub/2018/feature-wise-transformations.

[17] B. Li, X. Liu, K. Dinesh, Z. Duan, and G. Sharma, "Creating a musical performance dataset for multimodal music analysis: Challenges, insights, and applications," *IEEE Transactions on Multimedia*, vol. PP, 12 2016.

[18] M. Goto, "Development of the rwc music database," in *Proceedings of the 18th International Congress on Acoustics (ICA 2004)*, 2004, pp. 553–556.

[19] 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, July 2006.