LIGHTWEIGHT END-TO-END DEEP LEARNING MODEL FOR MUSIC SOURCE SEPARATION

Yao-Ting Wang[1], Yi-Xing Lin[1], Kai-Wen Liang[1], Tzu-Chiang Tai[2], and Jia-Ching Wang[1]

[1]Department of Computer Science and Information Engineering, National Central University, Taiwan
[2]Department of Computer Science and Information Engineering, Providence University, Taiwan

ABSTRACT

In this work, we propose a lightweight end-to-end music source separation deep learning model. Deep learning models for audio source separation based on time-domain have been proposed for end-to-end processing. However, the proposed models are complex and difficult to use when the computing resources of the device are limited. Additionally, long delays may be expected since long-term inputs are required to obtain adequate results for separation, making the models unsuitable for applications that require low latency. In the proposed model, Atrous Spatial Pyramid Pooling is used to reduce the number of parameters, and the receptive field preserving decoder is utilized to enhance the result of separation while the input context length is limited. The experimental results show that the proposed method obtains better results than previous methods while using 10% or fewer parameters.

Index Terms— Deep learning, lightweight, music source separation

1. INTRODUCTION

Deep Neural Networks (DNNs) have made a hit in the multimedia signal processing field. Previous works adopted Short Time Fourier Transform (STFT) to transform the signal to spectrum and used the magnitudes to separate the sound sources [1]. It has also been combined with Deep Clustering (DPCL) and the Bidirectional Long Short Term Memory (BLSTM) Network, or the U-Net for music source separation [2].

To avoid the loss of information caused by ignoring the phase information, methods based on the time domain information for end-to-end processing have been proposed. Conv-TasNet proposed by Luo [3] performs very good in speaker independent speech separation, and the Wave-U-Net proposed by Stoller [4] also obtained good results in music source separation.

Although the separation technology has made certain progress, these models usually use large numbers of parameters and have difficulties to be implemented with limited computational resources. On the other hand, it generally requires long term input to gain more advances in those methods. Since long term input is associated with extremely high latency, it is less suitable for applications that require low latency.

The objective of this work is to magnify the separation results of various sound sources in real-time applications. Therefore, we focus on input with short context length, that is, in a low-latency state, to improve the existing deep learning model with the purpose of reducing the number of parameters thereby increasing the operational speed and further improving the whole system performance.

The rest of the paper is organized as follows. Section 2 represents the proposed model. Experimental results and discussions will be given in Section 3 and Section 4 concludes this work.

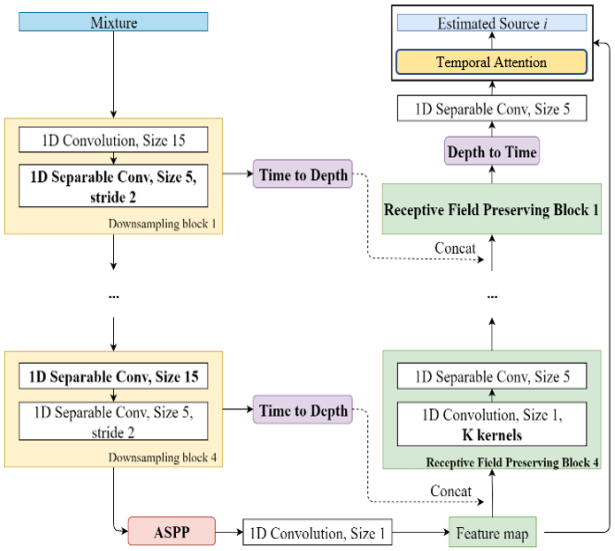

Figure 1. Overview of the proposed model with the ASPP and the RFPD.

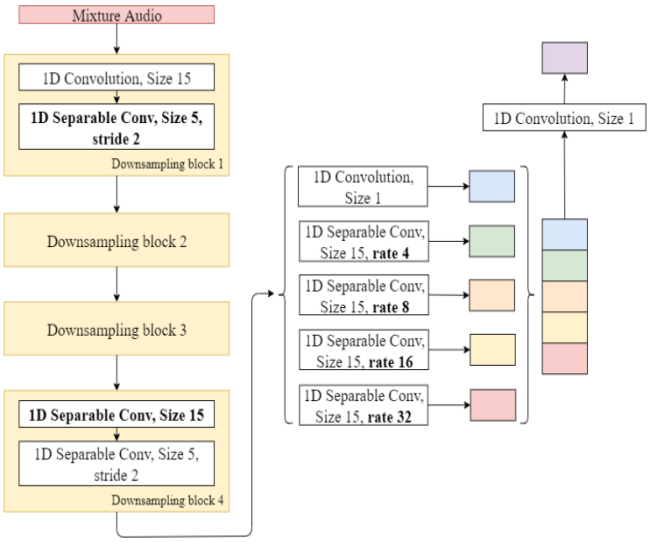

Figure 2. The encoder with the ASPP and separable convolution in the 4-th downsampling block.

2. MODEL DESCRIPTION

The proposed model is designed based on the architecture of Wave-U-Net [4] while replacing the encoder such as the DeepLabv3+ [5] for a lightweight model with the Atrous Spatial Pyramid Pooling (ASPP). We also develop a novel Receptive Field Preserving Decoder (RFPD) to improve the decoder. In addition, we refer to the attention mechanism related methods [8], [9] to build a temporal attention block, which can further improve the separation result. The block diagram of the proposed model is shown in Figure 1.

2.1 Encoder

The encoder architecture is shown in Figure 2. The input is a mixture of audio signals. The downsampling block consists of a 1-D standard with strides of 2 for downsampling as [10] and a 1-D depth separable convolution

The depth-wise convolution does not change the number of channels. According to the MobileNetV2 [11], if we want to replace the standard convolution with a depth separable convolution, the feature map must have enough channels for the information between the channels and the spaces to uncouple in the encoding phase, otherwise the information suffers a serious loss after the nonlinear transformation via the activation function.

The first three downsampling blocks with standard convolution and the standard convolution of the fourth downsampling block are replaced by the depth separable convolution, which can effectively reduce the parameter size of the overall model and the computational cost. To avoid the loss of information caused by the downsampling from the excessive coding, we perform only four downsamplings. However, the receptive field of the encoder becomes too small. To obtain a larger receptive field without introducing too many parameters, we refer to the DeepLabv3+[5] and added the ASPP in the last layer.

2.2 Receptive Field Preserving Decoder

We have adopted the Space to Depth (S2D) and the Depth to Space (D2S) that have been commonly used in the super-resolution methods [12] [13] to perform upsampling, referred as Time to Depth in time series. It improves the operational speed and is beneficial to platforms with limited hardware resources. Some image semantic segmentation methods try to decrease the number of decoder parameters thereby accelerating the computational processes.

However, the decoder's ability in audio source separation is as important as the processing time. Restored signals with better details from an effective decoder gain much benefits to the performance. Since more upsampling will increase the time resolution and decrease receptive field of the convolution kernels with the same size at the same time. The improvement from decoders would be ceiled in this way. Hence, we propose a decoder that has the ability to preserve the receptive field. During decoding, the depth to time operation is not performed, but time to depth is used to map the lower temporal-resolution to the high-level features. The high-level features which contain the information from the high temporal-resolution features can then be obtained In this way, the size of the receptive field can be preserved during the decoding process, which helps to compensate the information lost due to downsampling and restore the audio details.

3. EXPERIMENTS

To evaluate the performance of our models, we conduct experiments with two tasks and compare the results to the Wave-U-Net [4]. The first task is the separation of singing voice and music, and the second task is to separate vocals, bass, drums, and other instruments into categories defined by the SiSEC [14].

3.1 Datasets

75 tracks from the training set of the MUSDB database [14] have been randomly chosen for our training set. Another 25 tracks are used as the validation set for early stopping. The final performance has been evaluated on a MUSDB test set with 50 songs. We have also added the entire CCMixter

database [15] to the training set in the singing voice and music separation task.

We have adopted the data augmentation method from the Wave-U-Net [4], uniformly selected a factor from the interval of [0.7, 1.0], multiplied by the source signals and set the input mixture as the sum of the weighted source signals. All signals have been normalized to the interval of [-1, 1], and downsampled to 22,050 Hz.

3.2 Training procedure

During training stage, audio segments have been randomly chosen to form a batch. We calculate the mean squared error (MSE) over all source output samples in a batch as the loss. The ADAM optimizer has been used with a learning rate of 0.00001, decay rates of $\beta_1 = 0.9$ and $\beta_2 = 0.999$ and a batch size of 32. One epoch contains 2,000 iterations, and early stopping is performed if there is no improvement in the validation set within 30 epochs. The last model is fine-tuned, with the same batch size and a decreased learning rate of 0.000001, and the process is repeated until there is no improvement in the validation loss for 30 epochs. Finally, the model with the best validation loss is selected as the outcome.

Table 1. Comparison among the proposed 3 models in SDR (dB).

	(a)	(b)	(c)
Voc.	4.55	4.64	**4.92**
Acc.	10.74	10.77	**10.97**

Table 2. Comparison for the models, include the number of parameters, receptive field and target field.

Model	Context length	#parameters	Receptive field	Target field
Wave-U-Net[4]	147443	10.2M	6.69s	0.743s
Proposed	16384	0.85M	0.743s	0.743s

3.3 Experiment results

3.3.1. Evaluation metrics

We use the signal-to-distortion (SDR) metric [16] to evaluate the performance of our model. Following the procedures used for the SiSEC 2018[14], an audio track is partitioned into non-overlapping, one second audio segments. An averaged score over each audio track or the whole dataset is calculated by a segment-wise metrics to evaluate the model performance. Based on the discussion from the Wave-U-Net [4], this work only considers the median SDR scores.

3.3.2. Model comparison

In this approach, we have proposed multiple functions to gain improvements. Therefore, we evaluate the different combinations of the proposed techniques. The performance comparison is given in Table 1. In Table 1, Voc. denotes the vocals and Acc. represents the accompaniment. We compare the following five models: (a) adopted RFPD only; (b) RFPD and temporal attention; (c) use upsampling to replace RFPD in (b). It is noted that the encoders of these models are same, i.e., the ASPP. In Table 1, we have proven that the RFPD further improves the performance of separation. Attention is helpful even if we have only added it in the last layer.

Another major improvement in this paper is that in addition to the improvement of separation performance, the amount of parameters is also significantly reduced, which is achieved under the condition that the receptive field is not too large. In Table 2, we have only taken 0.743s of the audio signal as the input of our model, and the model that we compared with has an input that is 12 times larger. Nevertheless, we have obtained better performance than the compared model, as shown in Tables 3 and 4. The results indicate that the proposed model is more suitable for the implementation on different platforms, especially on mobile devices. In Table 4, the blocks of temporal attention have been added for vocals, bass and drums, the difference between the mixture signal and the summation of the mentioned instruments is represented by “other”.

Table 3. Comparison for singing voice and music separation in SDR (dB).

	Wave-U-Net[4]			Proposed
	Mono M3	Stereo M4	Stereo M5	Mono
Voc.	3.96	4.46	4.58	**4.92**
Acc.	7.53	10.69	10.66	**10.97**

Table 4. The SDR performance in separating vocals, bass, drums, and other instruments.

	Vocals	Other	Bass	Drums
Wave-U-Net[4]	3.0	2.03	2.91	4.15
Proposed	**3.68**	**2.50**	**3.09**	**4.45**

As shown in Figure 3, it appears that the results may be further improved if the audio is transferred to the spectrum form. Although the main features of the vocals are preserved, the influence of the accompaniment still exists. Therefore, we may consider combining the time-domain and frequency-domain information in the future.

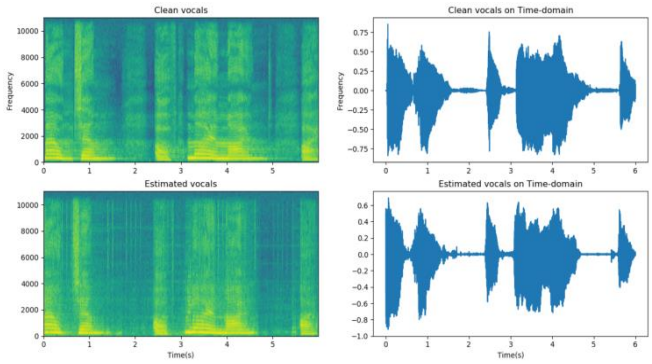

Figure 3. Comparison between the clean and the estimated singing voice. The first row is the clean voice and the second row is the estimated result.

4. CONCLUSION

The proposed model successfully takes advantages of the prevailing methods in image semantic segmentation towards reductions in the numbers of the coding layers and the model parameters. We have obtained better results than the previous methods and meanwhile required only less than 10% of the previous number of parameters is required. The mutual conversion of depth and time has also been used to improve operational speed thereby enabling the proposed method to function with much fewer hardware resources. Considering the common and different parts between image and sound, we have applied the time attention mechanism. Finally, we have also proposed a novel decoder with receptive field preservation which further enhances the separation result of audio sources.

5. REFERENCES

[1] Y. Luo, Z. Chen, J. R. Hershey, J. L. Roux, and N. Mesgarani, "Deep clustering and conventional networks for music separation: stronger together," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 61–65, 2017.

[2] A. Jansson, E. J. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde. "Singing voice separation with deep U-Net convolutional networks," *International Society for Music Information Retrieval Conference*, pp. 323–332, 2017.

[3] Y. Luo and N. Mesgarani, "TasNet: surpassing ideal time frequency masking for speech separation," *arXiv preprint* arXiv:1809.07454, 2018.

[4] D. Stoller, S. Ewert, and Simon Dixon, "WaveU-Net: a multi-scale neural network for end-to-end source separation," *International Society for Music Information Retrieval Conference*, vol. 19, pp. 334–340, 2018.

[5] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," *European Conference on Computer Vision*, 2018.

[6] A. Vaswani, N. Shazeer, N. Parmar, and J. Uszkoreit, "Attention is all you need," *arXiv Preprint*, arXiv:1706.03762, 2017.

[7] J. Hu, L. Shen, and G. Sun. "Squeeze-and-excitation networks," *arXiv preprint* , arXiv:1709.01507, 2017.

[8] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: convolutional block attention module," *The European Conference on Computer Vision*, 2018.

[9] F. Wu, A. Fan, A. Baevski, Y. N. Dauphin, and M. Auli, "Pay less attention with lightweight and dynamic convolutions," *International Conference on Learning Representations*, 2019.

[10] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," *Computer Vision and Pattern Recognition*, 2017.

[11] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," *Computer Vision and Pattern Recognition*, 2018.

[12] W. Shi, J. Caballero, F. Husz´ar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," *Computer Vision and Pattern Recognition*, 2016.

[13] M. S. Sajjadi, R. Vemulapalli, and M. Brown, "Frame-recurrent video super-resolution," *Computer Vision and Pattern Recognition*, 2018.

[14] Z. Rafii, A. Liutkus, F. R. Stter, S. I. Mimilakis, and R.Bittner, *The MUSDB18 corpus for music separation*, 2017.

[15] A. Liutkus, D. Fitzgerald, and Z. Rafii, "Scalable audio separation with light kernel additive modeling," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 76–80, 2015.

[16] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, pp. 1462–1469, 2006.