INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Contrastive Learning based Deep Latent Masking for Music Source Separation

Jihyun Kim[1], *Hong-Goo Kang*[1]

[1]Department of Electrical and Electronic Engineering, Yonsei University, South Korea

jihyun93815@dsp.yonsei.ac.kr, hgkang@yonsei.ac.kr

Abstract

Recent studies on music source separation have extended their applicability to generic audio signals. Real-time applications for music source separation are necessary to provide services such as custom equalizers or to improve the sound of live streaming with diverse effects. However, most prior methods are unsuitable for real-time applications due to their high computational complexity, large memory usage, or long latency. To overcome these problems, we propose a Wave-U-Net type of music source separation network that utilizes high-dimensional masking for the deep latent domain features. We also introduce a contrastive learning technique to estimate the salient latent space embedding of each target source using a masking-based approach. The performance of our proposed model is evaluated on the MUSDB18HQ dataset in comparison with several baselines. The experiments confirm that our proposed model is capable of real-time processing and outperforms existing models.

Index Terms: Music source separation, Contrastive learning, Real-time processing

1. Introduction

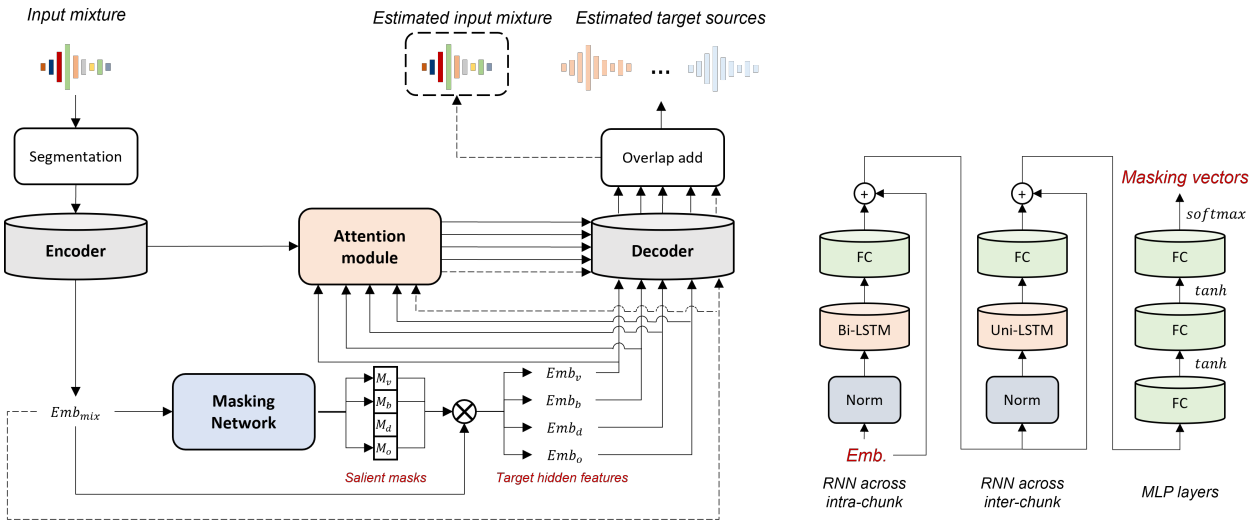
Music source separation refers to the task of extracting individual music sources from a mixture of signals. There are various deep learning techniques that can be used to perform music source separation [1–8]. Time-frequency (T-F) masking [9–14] is a widely used technique for music source separation, but it has some limitations. One of the main challenges is that it relies on complex phase information, which can be difficult to estimate accurately. This can lead to artifacts in the separated sources and limit the quality of the results. Moreover, T-F masking typically requires a large number of parameters, which can make the model difficult to train and slow to run in real-time applications.

A natural way to overcome this issue is to process the signal directly in the time domain [3, 5, 15–19]. Time-domain source separation models can be categorized into two types of approaches: basis signal estimation networks and U-Net style encoder/decoder networks. TasNet [15] is a basis signal estimation network that estimates the principal components of the input and masking values for each mixture. However, since it operates on very short input chunks, it is difficult to utilize long-time-interval information, which plays a crucial role in performing separation. To process long sequences efficiently, dual-path recurrent neural networks (DPRNNs) [17] split the input sequence into stacks of overlapped chunks, which are then processed in parallel using RNN layers. However, this structure requires very high computational complexity due to processing the short input signal continuously without temporal compression. DEMUCS [5] uses a Wave-U-Net framework that consists of multiple layers of down-sampling and up-sampling blocks with strided convolutional architectures. Moreover, it utilizes recurrent processing with bi-directional long-short term memory (bLSTM) networks in the bottleneck layer to process the temporal domain efficiently. However, this technique requires a large number of model parameters and the long input sequences. Since processing long input sequences causes latency problems and requires a large amount of memory, this method is not suitable for real-time applications.

Recent research on neural speech enhancement/separation network has been conducted, with the aim of providing real-time functionalities by reducing the number of parameters, computational complexity, and latency of models. DPRNNs [17] have been proposed as a simple way to model the long sequential input, by splitting the input into short segments and applying inter-chunk RNN and intra-chunk RNN. By combining a DPRNN with a U-Net based speech enhancement network, dual-path CRN (DPCRN) [20] processes short input segments more efficiently with small model size and low computational cost. In [21], another Wave-U-Net-based speech enhancement network was developed which modifies dual-path CRN into a time-domain approach and utilizes various attention mechanisms for efficient processing of the input sequence. This method achieved state-of-the-art performance on real-time speech enhancement tasks with low computational complexity. Although this method achieves good performance on speech enhancement, the same structure can't show good performance in music source separation tasks due to the complicated processing performed on the deepest latent layer. Moreover, modeling long sequential inputs is critical to processing the audio signals rather than speech signals, as it has more diverse properties.

In this paper, we propose a novel time-domain neural music source separation model that combines the advantages of the basis signal estimation network and the Wave-U-Net style encoder/decoder network. The baseline architecture of our model is similar to that used in [21]. We modify the entire architecture or hyper-parameters to render it more appropriate for music source separation. Compared with other state-of-the-art models for music source separation [22], we significantly reduce the number of model parameters and complexity in the deepest latent layer. Specifically, we include a light-weight masking-based separation network to achieve these reduction while maintaining high separation performance. To separate each target source's salient features effectively, we apply contrastive learning [23] on the deepest latent layer. In addition, we introduce an attention mechanism on the skip-connection layers of the U-Net-based network to efficiently utilize deep latent maskings, significantly improving source separation performance.

10.21437/Interspeech.2023-1723



(a) *Overall architecture of the proposed model* (b) *Masking network on bottleneck block*

Figure 1: *Details of the proposed Deep Latent Masking network. We apply a masking network to the bottleneck module and an attention module to each of the skip-connection layers. Dotted lines in (a) are only used for training. The masking network consists of the intra- and inter-chunk proceesing and MLP layers.*

2. Related Works

Most recent neural audio source separation networks consist of a U-Net style encoder, decoder, and an additional bottleneck module. The main differences between these networks reside in the detailed structure of each module, skip connections, and the ways in which input signals are handled.

2.1. U-Net based music source separation networks

DEMUCS, one of the state-of-the-art models for music source separation, has been developed based on Wave-U-Net architecture across multiple versions. [5], an early version of DEMUCS, uses 1D convolution layers with gated linear unit activation functions for the down-sampling blocks in the encoder and up-sampling blocks in the decoder. The consecutive down-sampling blocks enable efficient processing in the bottleneck module by down-sampling long signals multiple times. Moreover, to leverage the high-dimensional down-sampled time-domain characteristics of the input sequence in the latent domain, the bottleneck module used two bi-directional recurrent neural network (RNN) layers. This method achieved high performance by increasing the channel size of the bottleneck module, significantly enhancing the resolution of multi-target music source separation. However, this method requires a large number of model parameters and a long input sequence length (i.e., latency issue) due to the use of several sampling blocks with relatively large sampling factors.

In recent versions of DEMUCS [22, 24], the signal is processed in both the time- and frequency domains to achieve high-quality separation performance. Using a time-domain encoder and a frequency-domain encoder, as well as a time-domain decoder and a frequency-domain decoder, to process multi-domain signals at once, they successfully extended the U-net structure to a multi-domain analysis. Although these approaches achieved high separation performance, they still face latency issues and require a large number of model parameters due to the use of multi-domain input signals with long sequences. However, since improving separation performance is still a challenge, recent research has not focused on addressing the issues of latency and high computational complexity.

2.2. Dual-path RNN on bottleneck for real-time processing

For speech enhancement, [21] introduces time-domain DPCRN which has DPRNN [17] in the bottleneck module of the Wave-U-Net-based network to solve the problems of latency and computational complexity. After dividing a long input sequence into overlapped shorter chunks, each chunk is passed through sub-sampling blocks. Then, the obtained latent embeddings are used for intra- and inter-chunk processing in the bottleneck module. Intra-chunk processing captures local information across the time-axis, whereas inter-chunk processing captures global information at the same time step of each chunk.

A temporal and channel attention mechanism was also designed in [21] to extract salient information from the input features with light weight. The model can achieve real-time operation with low latency because it uses input sequences of the same length as the split chunks used during training. Although this method achieved state-of-the-art performance on speech enhancement, it did not perform as well on music source separation tasks because audio signals contain more diverse characteristics. This is because music signals are generally more complex and contain a wider range of frequencies and timbres than speech signals, which requires a larger number of model parameters to effectively separate the different sources in the mixture. Thus, the channel size of the bottleneck block becomes an important factor in determining separation performance. However, setting a very large channel size is not an optimal solution because the channel size is associated with the number of the model parameters.

3. Proposed Method

3.1. Overall structure

The objective of our proposed method is to reduce the number of model parameters and computational complexity while

maintaining high separation performance in real-time processing manner. It is not optimal to adopt the U-Net architecture for music source separation because it uses a small number of channels in the bottleneck module. In contrast to speech enhancement tasks, music source separation requires more complex processing as the characteristics of music signals are more diverse. To maintain low complexity while improving separation performance, we introduce a masking network to the bottleneck module, which computes the salient feature for each target source. Since our proposed masking network disentangles sound sources in the latent space, separation performance can be improved significantly. Furthermore, even when the channel of the bottleneck module is not large enough, the addition of the masking network can still improve separation performance.

In a U-net-based architecture, the output of each encoder block is passed to a corresponding layer in the decoder block via a skip-connection to compensate for any information losses caused by passing encoder and decoder sampling blocks. This is useful when there is only one target speaker to enhance, as in the case of normal speech enhancement. However, unlike conventional U-Net-based separation networks, our proposed network obtains target mask vectors. This technique requires performing separation in the bottleneck module, and then the decoder reconstructs each target source separately in the decoder network rather than all at once, as shown in Figure 1a. Since the target signals are generated independently of each other, the information transferred through the skip-connection needs to be refined appropriately for each target. We also use an attention mechanism in the skip-connections to provide coherent information, which is useful for separating each target source. This attention mechanism helps our model to capture and transfer more compact information for generating multiple target sources.

3.2. Masking networks on the bottleneck block

We obtain the masking vectors in the bottleneck module, as shown in Figure 1a. Our bottleneck module consists of RNN-based chunk processing layers which utilizes sequential information. To obtain masked salient features, we also include MLP layers [25] at the back of the chunk processing layers. The MLP layers comprise of three fully connected (FC) layers, and use hyperbolic tangent as the activation function for the first and second layers and softmax for the last layer. Through the softmax operation, we directly separate the target hidden features from the hidden feature of the mixture in the latent domain. As with real-world source separation, when all of the extracted features are summed together, it becomes identical to the hidden features of the input mixture.

As shown in Figure 1b, the masking network that consist of chunk processing and MLP layers computes the same number of masks as the number of target sources. After obtaining each target masking, the input hidden features are multiplied element-wise with the estimated masking values to obtain a salient feature for each target source. Each of the estimated target features is then passed through the decoder independently to reconstruct each target audio signal. During training, both the mixture and estimated target features (the output of the encoder) are passed through the decoder. To make the salient maskings more meaningful, the mixtures are reconstructed from the output of the encoder using the salient features without any masking.

3.3. Attention module

To transmit the target-related information reliably, we apply an attention module to the skip-connections. Here, we use the target feature (estimated in the bottleneck module using a masking network) as the key and the mixture features from each layer of the encoder as the query and value. An attention map is calculated from the processing of the key and query by a 1×1 convolution layer, which captures information that is more relevant to the target signal. Finally, the attention feature is obtained by multiplying the attention weights with the value. The estimated attention features are then transmitted to each layer of the decoder block to generate target-related information.

This attention module is essential for our proposed model because only a single decoder is used to generate multiple target sources by utilizing the masked hidden features. Since the decoder independently generates the target signal (in contrast to the conventional U-net approach), it effectively transfers target-related information from the encoder to the decoder during the generation process. In Section 5, we demonstrate that the addition of the attention module significantly improves separation performance compared to only using simple skip-connections.

4. Training

To train our model, we use multi-domain loss (MDL) and combination loss, as proposed in [26]. This is defined as follows:

$$\mathcal{L}_{CL} = \frac{1}{M} \sum_{m=1}^{M} \mathcal{L}_{MDL}^{m}, \tag{1}$$

$$\mathcal{L}_{MDL} = \mathcal{L}_{mSTFT} + \mathcal{L}_{MSE}, \tag{2}$$

where $M = \sum_{i=1}^{J-1} \binom{J}{i}$, and J denotes the number of target signals. Combination loss reflects all the possible combinatorial losses of estimated signals and prevents the estimated target signal from leaking to other signals.

We use multi-resolution STFT loss [27] for the frequency domain criterion and MSE loss for the time domain criterion to achieve multi-domain loss. The FFT sizes of multi-resolution STFT are set to (512, 1024, 2048).

We also use a contrastive loss $\mathcal{L}_{contrastive}$ [23] for the latent embeddings in the bottleneck module. $\mathcal{L}_{contrastive}$ utilizes the latent embeddings in the bottleneck module to reliably estimate masking values and is defined as follows:

$$\mathcal{L}_{Contrastive} = -\sum_{i \in I} \log \frac{\sum_{b \in P(i)} \exp z_i \cdot z_b / \tau}{\sum_{a \in N(i)} \exp z_i \cdot z_a / \tau}, \tag{3}$$

where z denotes the hidden feature (*Emb* in Figure 1a). In addition, $i \in I \equiv 1, 2, \cdots, N \times K$, $P(i)$ is a set of positive samples that belong to the remaining chunks of the same target, and $N(i)$ is a set of negative samples. The positive and negative pairs are obtained as follows: The input audio signal $\mathbf{W} \in \mathrm{IR}^{2 \times L}$ is split into K chunks of equal size $\mathbf{S}_k \in \mathrm{IR}^{2 \times \mathrm{T}}, k = 1, ..., K$, producing K chunks per target source. For the purpose of contrastive learning, each target embedding can be treated as having $K - 1$ positive samples and $N \times (K-1)$ negative samples, where N is the number of target sources. We utilize a simple convolutional network to transform the latent embeddings into a new domain, which make it easier to apply the contrastive loss. This network is only used during training. The total training loss is as follows:

$$\mathcal{L}_{separator} = \beta \cdot \mathcal{L}_{CL} + (1 - \beta) \cdot \mathcal{L}_{contrastive}, \tag{4}$$

where we set $\beta = 0.99$ in our experiments.

Table 1: *Comparison of SDR performance with other state-of-the-art models. Models marked with an asterisk (*) indicate their ability to operate in real-time. (MN = Masking Network, AM = Attention Module, CT = ConTrastive learning)*

Architecture	Params	avg.	vocals	drums	bass	other
DEMUCS v2 [5]	133.8 M	6.28	6.84	6.86	7.01	4.42
Hybrid DEMUCS [24]	83.6 M	7.68	8.13	8.24	**8.76**	5.59
HT DEMUCS [22]	41.4 M	7.52	7.93	7.94	8.48	5.72
Band-split RNN [28]	-	**8.24**	**10.01**	**9.01**	7.22	**6.70**
Conv-TasNet* [16]	5 M	5.73	6.81	6.08	5.66	4.37
Meta-TasNet* [18]	45.4 M	5.52	6.40	5.91	5.58	4.19
Ours* (baseline)	5.1 M	4.56	4.72	4.94	5.28	3.29
+MN*	5.5 M	5.30	5.69	5.83	6.02	3.67
+AM*	5.7 M	6.15	6.57	6.76	6.94	4.31
+CT*	**5.7 M**	**6.47**	**6.91**	**7.05**	**7.29**	**4.62**

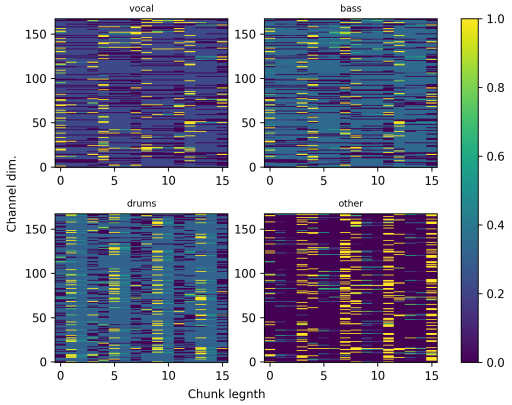


Figure 2: *Salient maskings for each target on the same chunk from the bottleneck block.*

5. Experiments

5.1. Data and evaluation

We evaluated the performance of our proposed model using the MUSDB18 dataset [29]. The dataset consists of 100 tracks for the training set and 50 tracks for the test set, all sampled at a rate of 44.1 kHz. For training, we randomly selected 75 tracks from the training set, while the remaining tracks were used for validation purposes. To evaluate the performance of our model, we utilize signal-to-distortion ratios (SDRs) calculated using the *bss_eval* metrics [30], as defined by the SiSEC18 [31].

5.2. Experiment settings and results

Table 1 shows the performance of our proposed method and other state-of-the-art source separation models. For fair comparison, we obtained the results without using any data augmentation for training. We also performed ablation studies to analyze the impact of each component of our model (i.e., masking networks, attention mechanism in skip-connections and contrastive loss) Here, it is evident that each of the techniques we applied significantly improves SDRs performance. Although our proposed model does not perform well compared to the Band-split RNN and hybrid-domain DEMUCS, our model demonstrates the best performance among the real-time processable models. In particular, our model achieves higher performance than DEMUCS v2, which requires a long input sequence and a large number of model parameters. All the reference models except TasNet-based models require long input signal, which is not suitable for real-time processing due to the long latency. Different from these models, we set the length of the input signal to be very short to avoid the latency issues.

Table 2: *Comparison of results based on different components of the model. All models include the MN, AM and CT proposed in our approach.*

dur. (samples)	depth	dim.	params.	RTF (cpu)	SDR (avg.)
512	3	128	1.4 M	0.76	4.92
512	3	256	5.7 M	3.36	5.94
512	4	256	5.7 M	1.99	5.81
1,024	4	128	1.4 M	0.81	5.39
1,024	4	256	5.7 M	1.91	**6.47**
1,024	5	256	5.7 M	0.88	6.18

Our model requires an input length of approximately 23.2 ms ($1,024$ samples at 44.1 kHz) on inference stage.

Table 2 shows the impact of hyper-parameters of our proposed model (i.e., segment duration, depth, and dimension). We compared the results in terms of SDRs and Real Time Factors (RTFs) which are computed on a single core of Intel(R) core(TM) i9-10900X CPU @ 3.70GHz. The separation networks that use long input sequence and large bottleneck channel size show better performance than short input and small channel size. Increasing the depth of the separation network, which refers to the number of sampling blocks, can reduce computational complexity for a fixed input sequence length and bottleneck channel size, as it shortens the temporal length of the bottleneck. However, if the depth becomes too deep, there may be a significant loss of temporal information, resulting in a decrease in performance. We performed various experiments by changing the input length, depth, embedding dimension, and summarize the results in Table 2.

5.3. Analysis

To conduct a more detailed analysis of our model, we plotted salient maskings for each target source obtained within a chunk (Fig. 2). We observed that the masking values of each source cover different sections of the embedding. Specifically, when the masking value of one target is high in a certain latent bin, the masking value of the other target is low in the same bin. This result indicates that the masking values correspond to each source in the mixture, effectively representing specific target-related information for separation. Since the mask values are continuous rather than binary, our masking network can effectively utilize the dimensionality of the bottleneck. This means that our model achieves efficient source separation without the need to increase the size of the bottleneck channel.

6. Conclusion

In this paper, we present a novel approach to music source separation based on the Wave-U-Net architecture. Our proposed method achieves real-time processing capabilities and maintains a small model size while improving the separation performance. In our proposed network, we estimate the latent domain embedding of each target source within the bottleneck module using a masking-based approach. We use contrastive learning on the bottleneck to obtain salient maskings. We also employ a cross-attention module to the encoder-decoder skip-connections to transmit information between them more effectively. Our model can operate in real-time with reduced computational complexity and a smaller number of model parameters. The experimental results demonstrate that our model achieves the optimum balance between music source separation performance and efficiency.

Acknowledgments. This research was supported by the Yonsei Signature Research Cluster Program of 2022 (2022-22-0002)

7. References

[1] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2136–2147, 2015.

[2] F. Lluís, J. Pons, and X. Serra, "End-to-End Music Source Separation: Is it Possible in the Waveform Domain?" in *Proc. Interspeech 2019*, 2019, pp. 4619–4623.

[3] 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 International Society for Music Information Retrieval Conference, ISMIR 2018, Paris, France, September 23-27, 2018*, E. Gómez, X. Hu, E. Humphrey, and E. Benetos, Eds., 2018, pp. 334–340. [Online]. Available: http://ismir2018.ircam.fr/doc/pdfs/205_Paper.pdf

[4] 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, 2019.

[5] A. Défossez, N. Usunier, L. Bottou, and F. Bach, "Music source separation in the waveform domain," *arXiv preprint arXiv:1911.13254*, 2019.

[6] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Singing-voice separation from monaural recordings using deep recurrent neural networks." in *ISMIR*, 2014, pp. 477–482.

[7] A. J. Simpson, G. Roma, and M. D. Plumbley, "Deep karaoke: Extracting vocals from musical mixtures using a convolutional deep neural network," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2015, pp. 429–436.

[8] S. Venkataramani, J. Casebeer, and P. Smaragdis, "End-to-end source separation with adaptive front-ends," in *2018 52nd Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2018, pp. 684–688.

[9] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep u-net convolutional networks," 2017.

[10] Y. Luo, Z. Chen, J. R. Hershey, J. Le Roux, and N. Mesgarani, "Deep clustering and conventional networks for music separation: Stronger together," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 61–65.

[11] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 246–250.

[12] D. S. Williamson and D. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 25, no. 7, pp. 1492–1501, 2017.

[13] S. I. Mimilakis, K. Drossos, J. F. Santos, G. Schuller, T. Virtanen, and Y. Bengio, "Monaural singing voice separation with skip-filtering connections and recurrent inference of time-frequency mask," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 721–725.

[14] W. Mack and E. A. Habets, "Deep filtering: Signal extraction and reconstruction using complex time-frequency filters," *IEEE Signal Processing Letters*, vol. 27, pp. 61–65, 2019.

[15] Y. Luo and N. Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 696–700.

[16] ——, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[17] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 46–50.

[18] D. Samuel, A. Ganeshan, and J. Naradowsky, "Meta-learning extractors for music source separation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 816–820.

[19] Y. Hu, Y. Chen, W. Yang, L. He, and H. Huang, "Hierarchic temporal convolutional network with cross-domain encoder for music source separation," *IEEE Signal Processing Letters*, vol. 29, pp. 1517–1521, 2022.

[20] X. Le, H. Chen, K. Chen, and J. Lu, "DPCRN: Dual-Path Convolution Recurrent Network for Single Channel Speech Enhancement," in *Proc. Interspeech 2021*, 2021, pp. 2811–2815.

[21] J. Lee and H.-G. Kang, "Real-time neural speech enhancement based on temporal refinement network and channel-wise gating methods," *Digital Signal Processing*, vol. 133, p. 103879, 2023.

[22] S. Rouard, F. Massa, and A. Défossez, "Hybrid transformers for music source separation," *arXiv preprint arXiv:2211.08553*, 2022.

[23] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[24] A. Défossez, "Hybrid spectrogram and waveform source separation," *arXiv preprint arXiv:2111.03600*, 2021.

[25] K. Li, X. Hu, and Y. Luo, "On the Use of Deep Mask Estimation Module for Neural Source Separation Systems," in *Proc. Interspeech 2022*, 2022, pp. 5328–5332.

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

[27] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6199–6203.

[28] Y. Luo and J. Yu, "Music source separation with band-split rnn," *arXiv preprint arXiv:2209.15174*, 2022.

[29] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "Musdb18-a corpus for music separation," 2017.

[30] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[31] F.-R. Stöter, A. Liutkus, and N. Ito, "The 2018 signal separation evaluation campaign," in *Latent Variable Analysis and Signal Separation: 14th International Conference, LVA/ICA 2018, Guildford, UK, July 2–5, 2018, Proceedings 14*. Springer, 2018, pp. 293–305.