future internet

MDPI

Article

Effective Monoaural Speech Separation through Convolutional Top-Down Multi-View Network

Aye Nyein Aung, Che-Wei Liao and Jeih-Weih Hung *

Department of Electrical Engineering, National Chi Nan University, Nantou 545, Taiwan; s111356509@mail1.ncnu.edu.tw (A.N.A.); s109323018@mail1.ncnu.edu.tw (C.-W.L.)
* Correspondence: jwhung@ncnu.edu.tw

Abstract: Speech separation, sometimes known as the "cocktail party problem", is the process of separating individual speech signals from an audio mixture that includes ambient noises and several speakers. The goal is to extract the target speech in this complicated sound scenario and either make it easier to understand or increase its quality so that it may be used in subsequent processing. Speech separation on overlapping audio data is important for many speech-processing tasks, including natural language processing, automatic speech recognition, and intelligent personal assistants. New speech separation algorithms are often built on a deep neural network (DNN) structure, which seeks to learn the complex relationship between the speech mixture and any specific speech source of interest. DNN-based speech separation algorithms outperform conventional statistics-based methods, although they typically need a lot of processing and/or a larger model size. This study presents a new end-to-end speech separation network called ESC-MASD-Net (effective speaker separation through convolutional multi-view attention and SuDoRM-RF network), which has relatively fewer model parameters compared with the state-of-the-art speech separation architectures. The network is partly inspired by the SuDoRM-RF++ network, which uses multiple time-resolution features with downsampling and resampling for effective speech separation. ESC-MASD-Net incorporates the multi-view attention and residual conformer modules into SuDoRM-RF++. Additionally, the U-Convolutional block in ESC-MASD-Net is refined with a conformer layer. Experiments conducted on the WHAM! dataset show that ESC-MASD-Net outperforms SuDoRM-RF++ significantly in the SI-SDRi metric. Furthermore, the use of the conformer layer has also improved the performance of ESC-MASD-Net.

Keywords: speech separation; conformer; multi-view attention; residual conformer network

Citation: Aung A.N.; Liao, C.-W.; Hung, J.-W. Effective Monoaural Speech Separation through Convolutional Top-Down Multi-View Network. *Future Internet* **2024**, *16*, 151. https://doi.org/10.3390/fi16050151

Academic Editors: Fatima Salahdine, Hassan El Alami and Mohammed Ridouani

Received: 31 March 2024
Revised: 25 April 2024
Accepted: 26 April 2024
Published: 28 April 2024

1. Introduction

Speech separation involves the processing of audio signals in order to separate multiple speech signals that overlap, with the objective of identifying and processing each speaker's speech separately. speech separation is very beneficial in diverse interference conditions, such as while recording conferences, making phone conversations, or mixing music. It allows us to isolate distinct speech outputs and solve the problem of speaker overlap. Effective speech separation methods usually involve a comprehensive understanding of signal processing, machine learning, and artificial intelligence.

Deep neural network (DNN) techniques have greatly improved state-of-the-art speech separation approaches. To enhance the functionality of speech separation systems, the state-of-the-art methods employ a variety of DNN designs and training techniques [1], including the following.

- Deep clustering-based methods [2,3] learn embeddings from mixed audio signals that can be used to distinguish between different speakers. In [2], the spectrogram embeddings are learned through a DNN to be discriminative for clustering, and the

work in [3] extends the deep clustering framework by end-to-end training with better regularization, longer temporal context, and a deeper structure.

- Time-Domain Audio Separation Network (TasNet) [4,5] circumvents the drawbacks of time-frequency representation by operating directly in the time domain to separate speech signals. TasNet employs a 1D convolutional encoder–decoder system, with the separation module as its fundamental component. This module is responsible for estimating source masks based on the encoder outputs. These masks are employed to measure the encoded mixed signal in order to separate different sources.
- The ability to capture longer contextual information has been greatly improved with the rise of self-attention mechanisms and transformer design [6], leading to even better speech separation quality. Sepformer [7] and DPTNet [8] are two examples of speech separation systems that adopt transformers. They are typically implemented in the dual-path architecture, which introduces a large number of parameters due to the fact that these systems model intra-chunk local features and inter-chunk global relationships separately.

Several studies have developed more efficient speech separation architectures to address the growing concern over large models with significant computing demands. For example, the dual-path recurrent neural network (DRPNN) [9] is a lightweight speech separation model that efficiently models long sequences in time for a single channel. DPRNN splits the sequence into smaller chunks and, thus, allows for better gradient flow and easier optimization. This structure also enables the network to capture both local and global dependencies in the sequence. Another study [10] uses a BLSTM-based binary masking module to create a comparable lightweight speech separation system. This masking module is designed to separate the sound source of interest (SOI) from the cumulative environmental interference. Furthermore, TDANet [11] simulates the human brain's top-down attention mechanism by performing speech separation using global attention and cascaded local attention layers, providing comparable performance to Sepformer but with significantly less computational complexity. In addition, SuDoRM-RF [12] and its derivative SuDoRM-RF++ [13] use successive depth-wise convolutional downsampling to extract multi-resolution temporal features, reducing the number of layers needed while effectively capturing long-term temporal relationships for speech separation.

This study presents a new network called ESC-MASD-Net to improve the performance of speech separation models. This network is based on the SuDoRM-RF++ architecture but with two additional modules: the residual conformer network (ResCon) and the multi-view attention (MA) network. These two modules are crucial elements of the MANNER network [14]. We also enhance the U-Convolutional blocks in ESC-MASD-Net by adding a conformer layer [15]. We evaluate our model using the WHAM! dataset [16], and the results indicate that ESC-MASD-Net outperforms SuDoRM-RF++ in terms of scale-invariant signal-to-distortion ratio improvement (SI-SDRi) scores. This improvement is due to the contribution of each additional component (ResCon, MA, and the conformer layer). These findings suggest that ESC-MASD-Net has the potential to be a solution for speech separation in real-world applications.

The main contributions of this study can be summarized as follows.

- We propose ESC-MASD-Net, a speech separation framework that achieves state-of-the-art performance while maintaining low computation load and compact model size.
- ESC-MASD-Net is built on the SuDoRM-RF++ structure, and it considers the input feature stream's channel, global, and local information by employing a multi-view attention (MA) block. Additionally, the framework uses a residual conformer (ResCon) block that tunes the input channel numbers flexibly for extracting information without any loss due to the residual connection arrangement. Evaluation experiments indicate that adding MA and ResCon blocks significantly improves SI-SDRi.
- We investigate the conformer layer structure presented in [15] and propose three different arrangements to adopt the conformer layer in the U-Convolution blocks in ESC-MASD-Net. The evaluation results exhibit that ESC-MASD-Net equipped with

the conformer layer achieves superior performance, and thus, the conformer layer is well additive to ESC-MASD-Net.

The rest of the paper is organized as follows. Section 2 provides a brief overview of the SuDoRM-RF++ framework, while Section 3 introduces the new ESC-MASD-Net, which incorporates two modules, ResCon and MA, to enhance SuDoRM-RF++. The use of the comformer layer in the U-Convolutional blocks of ESC-MASD-Net is explained in Section 4. Section 5 provides details on the experimental setup, while Section 6 presents the experimental results and discussions. Finally, Section 7 concludes the paper with some remarks.

2. SuDoRM-RF++

The SuDoRM-RF++ method is depicted in Figure 1a. SuDoRM-RF++ adheres to the ESD architecture, which comprises three main components: the Encoder, Separator, and Decoder. The encoder is responsible for converting a mixture $\mathbf{x} \in \mathbb{R}^T$, which contains audio signals from K speakers and interferences into a representation $\mathbf{v_x}$ that characterizes the signal. Subsequently, the separator produces K latent representations $\{\hat{\mathbf{v}}_k, k = 1, 2, \cdots, K\}$ for each speaker in the mixture. Finally, the decoder proceeds to reconstruct the separated K source signals in the time domain, with each signal being denoted as $\hat{s}^k \in \mathbb{R}^T$.

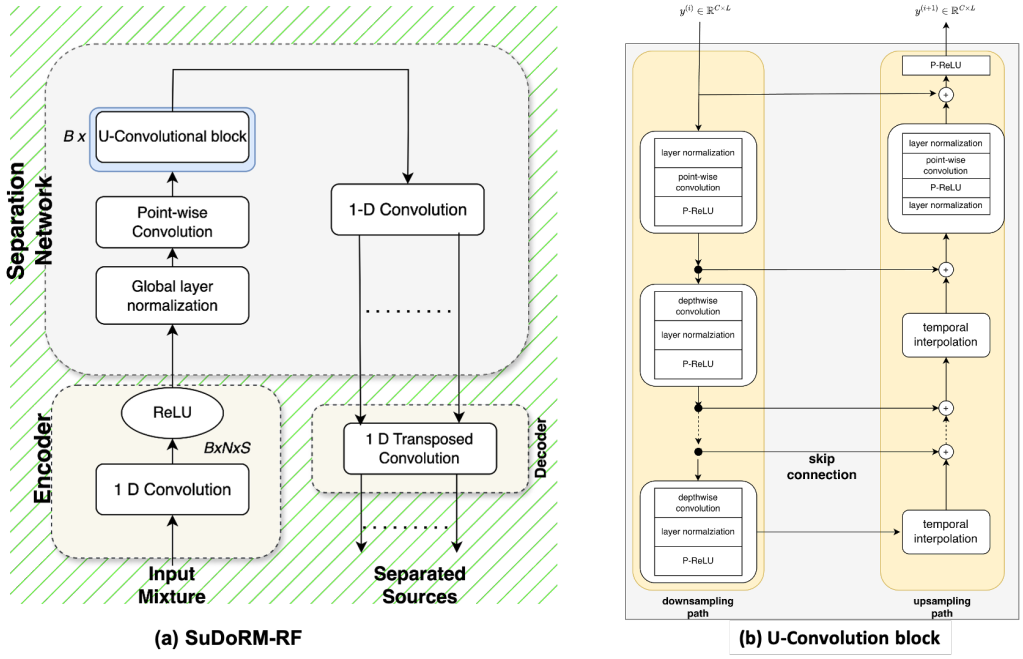

Figure 1. (**a**) The flowchart of SuDoRM-RF++. (**b**) The flowchart of a single U-Convolution block.

2.1. Encoder

The encoder $\mathcal{E}$ architecture is a one-dimensional convolutional network with $C_\mathcal{E}$ channels, kernel size $K_\mathcal{E}$, and stride $K_\mathcal{E}/2$. To ensure non-negativity, the output of the convolutional network is passed via a rectified linear unit (ReLU) layer. Hence, the encoded representation of the input mixture is expressed as:

$$\mathbf{v_x} = \mathcal{E}(\mathbf{x}) = \text{ReLU}(\text{Conv1D}(\mathbf{x})) \in \mathbb{R}^{C_\mathcal{E} \times L}. \tag{1}$$

2.2. Separator

The separator $\mathcal{S}$ initially applies a point-wise convolution and global layer normalization (GLN) on the encoded representation $\mathbf{v_x}$, expressed by:

$$\mathbf{y}_0 = \text{Conv1D}(\text{GLN}(\mathbf{v_x})) \in \mathbb{R}^{C \times L}. \tag{2}$$

The outcome $\mathbf{y}_0$ is subsequently processed using a sequence of U-Convolutional blocks. The U-Convolutional block serves as the signature operation of SuDoRM-RF++, employing temporal downsampling and resampling operations in succession to extract separation information from inputs of various resolutions. The U-Convolutional block structure is illustrated in Figure 1b, while a comprehensive explanation of its specific operations can be found in [13]. We omit this part due to the extensive description required and its exclusion from the scope of our study.

The output of the B consecutive U-Convolutional blocks, $\mathbf{y}^{(B)}$, is passed through a convolution network in order to extract separated latent representations for multiple speakers in concatenation $\widehat{\mathbf{v}} = [\widehat{\mathbf{v}}_1, \widehat{\mathbf{v}}_2, ..., \widehat{\mathbf{v}}_K]$, as follows:

$$\widehat{\mathbf{v}} = \text{Conv1D}\Big(\Big(\mathbf{y}^B\Big)\Big) \in \mathbb{R}^{KC\mathcal{E}\times L}. \tag{3}$$

2.3. Decoder

The decoder is a transposed convolution network that shares the same stride and kernel size as the encoder. With the latent representation $\widehat{\mathbf{v}}_k$ corresponding to the kth speaker, the decoder outputs the separated signal s^k:

$$\hat{s}_k = \text{Conv1D}^T(\widehat{\mathbf{v}}_k). \tag{4}$$

It is important to note that only one trainable convolution network is used to turn the latent representation back into the time-domain signals. This means that all speakers use the same decoder.

3. Presented Method: ESC-MASD-Net

With SuDoRM-RF++ as the archetype, this study proposes a new speech separation framework, ESC-MASD-Net; the flowchart is depicted in Figure 2. Comparing Figure 2 with Figure 1a, ESC-MASD-Net modifies SUDORM-RF++ in two distinct stages in the separator: first, it substitutes the point-wise convolution after the GLN process with a residual conformer (ResCon) block; second, it adds a multi-view attention (MA) block after the serial U-ConvBlocks. In the following, we introduce the operations of ResCon and MA blocks, which are two fundamental blocks of MANNER [14], a highly-effective speech enhancement framework.

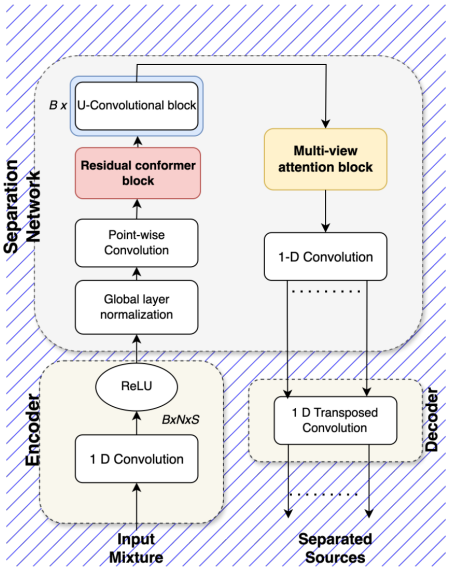

Figure 2. Flowchart of effective speaker separation through convolutional multi-view attention and SuDoRM-RF network (ESC-MASD-Net).

3.1. Residual Conformer Block

The residual conformer (ResCon) block, depicted in Figure 3, is a variation of the Conformer convolution block that increases the channel size to produce enriched channel representation. The primary branch of ResCon comprises a series of operations, including point-wise convolution with a channel growth factor of G_0, batch normalization, gated linear unit (GLU), depth-wise convolution, another batch normalization, Swish activation, and point-wise convolution with a channel growth factor of G_1. As a residual connection added to the main branch, the second branch consists of a point-wise convolution layer with channel growth factor G_1. The ReLU activation is applied to further process the output.

Below are some of the key features and advantages of the ResCon block.

1. The branch of the ResCon block expands channels for information capture and enhancement. Furthermore, this branch uses depthwise separable convolution, which combines depth-wise and point-wise convolution to drastically reduce overall processing requirements when compared to a standard convolution network.
2. The ResCon block uses a GLU (Gated Linear Unit) which operates as follows:

$$\mathbf{Y} = (\mathbf{X} * \mathbf{W} + \mathbf{b}) \otimes \sigma(\mathbf{X} * \mathbf{V} + \mathbf{c}), \tag{5}$$

   where $\mathbf{X}$ is the input, $\mathbf{W}$, $\mathbf{V}$, $\mathbf{b}$, and $\mathbf{c}$ are learnable affine transformation parameters, and σ is the sigmoid function. The GLU assists in learning which features of the input to pass through and which to block, allowing it to focus on relevant features for the given task. In addition, the number of channels can be adjusted using GLU's affine transformations.
3. The Swish function is represented as $y = x\sigma(\beta x)$, where β is a learnable parameter and σ is the sigmoid function. Swish has been demonstrated to outperform ReLU activation for deep neural network training. Unlike ReLU, Swish does not have the problem of terminating the learning process when the input is negative.
4. The ResCon block's residual connection branch preserves all input information, guaranteeing that the main branch only grasps information that will improve or sustain performance. Furthermore, the point-wise convolution applied to the residual connection aids in alternating the number of input channels with little computational effort.

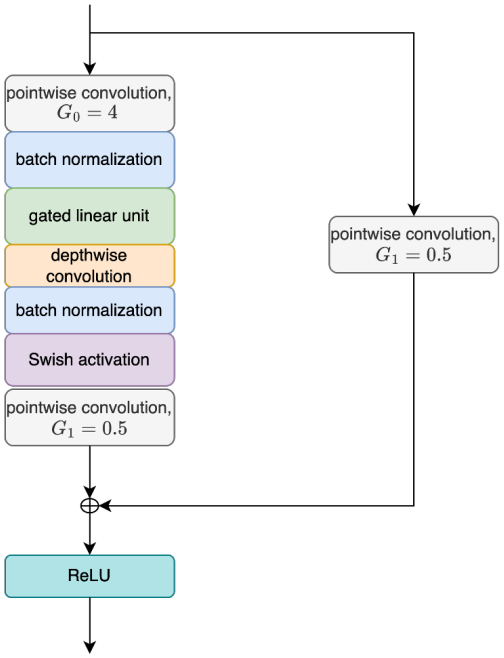

Figure 3. Flowchart of residual conformer block.

3.2. Multi-View Attention Block

The Multi-view Attention (MA) block, which consists of channel, global, and local attention, is integrated into the separation network's core architecture to fully represent the signal information provided by the output of the U-Conv Block. Channel attention, first of all, highlights representations from compressed channels. Meanwhile, dual-path-based global and local attention effectively reflects long sequential features. Figure 4 shows the architecture of the MA block.

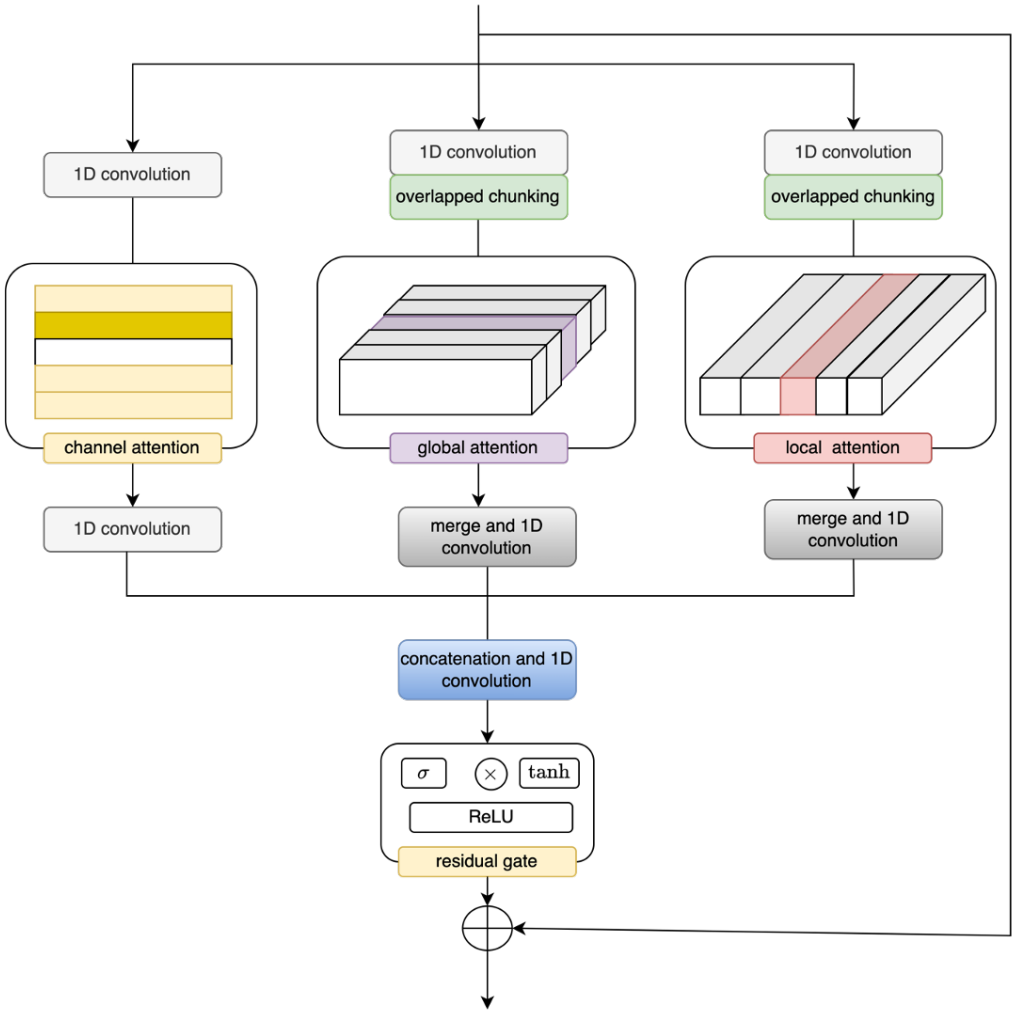

Figure 4. Flowchart of multi-view attention block.

Fundamentally, the MA block performs the following changes to the latent representation [14].

- Path Separation: The input travels through three routes in the MA block, each of which has a convolution layer that changes the channel size from C to $\lfloor C/3 \rfloor$. We use chunking with a 50% overlap ratio for both global and local attention paths, splitting $x \in \mathbb{R}^{\lfloor C/3 \rfloor \times L}$ into $x \in \mathbb{R}^{\lfloor C/3 \rfloor \times P \times S}$, where P and S stand for chunk size and number of chunks, respectively. This way, long sequential features are efficiently presented due to separating the global and local information. The three perspectives of attention are described as follows.

1. Channel Attention: Given the input $x_c \in \mathbb{R}^{\lfloor C/3 \rfloor \times L}$ for the channel attention path, the average and max pooling processes are used to aggregate the signal information for each channel $i = 1, 2, ..., \lfloor C/3 \rfloor$. The pooling outputs, x_c^{avg} and x_c^{max}, are then passed through a common densely connected network with one hidden layer having $\lfloor C/6 \rfloor$ nodes, followed by the sigmoid activation function σ to obtain the channel attention weight $\alpha_c \in \mathbb{R}^{\lfloor C/3 \rfloor \times 1}$:

$$\alpha_c = \sigma\left(W_1\left(W_0\left(x_c^{avg}\right)\right) + W_1(W_0(x_c^{max}))\right), \tag{6}$$

where $W_1 \in \mathbb{R}^{\lfloor C/3 \rfloor \times \lfloor C/6 \rfloor}$ and $W_0 \in \mathbb{R}^{\lfloor C/3 \rfloor \times \lfloor C/6 \rfloor}$ denote the weights of the used densely connected network. The channel attention weight is then broadcasted along the time dimension and multiplied with x_c to produce the channel attention output $x_c' \in \mathbb{R}^{\lfloor C/3 \rfloor \times L}$.

2. Global Attention: The global attention is based on the self-attention of Transformer, in which the chunk-wise representation $x_g \in \mathbb{R}^{\lfloor C/3 \rfloor \times P \times S}$ for the global attention input is taken into account. The corresponding output $x_g' \in \mathbb{R}^{\lfloor C/3 \rfloor \times P \times S}$ is determined by multi-head self-attention (MHA):

$$\alpha_g = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right), \quad x_g' = W(\alpha_g V), \tag{7}$$

where the linear transformation with each weight, $W_{q,k,v} \in \mathbb{R}^{1 \times S \times S}$, and $x_g \in \mathbb{R}^{\lfloor C/3 \rfloor \times P \times S}$, represent Q, K, and $V \in \mathbb{R}^{\lfloor C/3 \rfloor \times P \times S}$. Ultimately, a linear layer, $W \in \mathbb{R}^{1 \times S \times S}$, is applied to $\alpha_g V$ in order to extract the global attention output.

3. Local Attention: The local sequential features in each chunk are represented by local attention. On the chunked input $x_L \in \mathbb{R}^{P \times \lfloor C/3 \rfloor \times S}$, a depthwise convolution layer with a kernel size of $\lfloor S/2 \rfloor - 1$ is applied. Following the depthwise convolution layer, we concatenate the channel-wise average and max pooling to estimate the local attention weight $\alpha_l \in \mathbb{R}^{P \times 1 \times S}$ as follows:

$$\alpha_l = \sigma(F([x_l^{avg}; x_l^{max}])), \tag{8}$$

where F is the convolution layer that halves the channel size. At last, we express the final result of the local attention path as:

$$x_l' = x_l \times \alpha_l. \tag{9}$$

- Path Aggregation: Following the three attention paths, each output is concatenated and then passed through a convolution layer. A residual gate with sigmoid activation, hyperbolic tangent activation, and ReLU is used to process the resulting output further. Finally, a residual connection is made.

4. U-Convolutional Blocks Enhanced with Conformer Layer

The work in [15] proposes a mask estimation network topology with consecutive subsampling and supersampling layers close to the U-Convolutional block used in SuDoRM-RF++. However, conformer layers are applied at the output of the final subsampling layer. The flowchart of a conformer layer is depicted in Figure 5, showing that it is composed of a feed-forward module, a convolution module, a multihead self-attention (MHSA) module, and another feed-forward module. As stated in [15], the conformer layer operates as a dual-path network, sequentially processing both local and global contexts while being less computationally demanding. In this study, we suggest incorporating the conformer layer into the U-Convolutonal blocks of the separator for ESC-MASD-Net to investigate its potential additive nature. We propose the following three locations for the conformer layer:

(a) Right before the U-Convolutional blocks;
(b) At the bottom (right after mostly downsampled layer) of the first U-Convolutonal block.
(c) At the bottom of all U-Convolutional blocks.

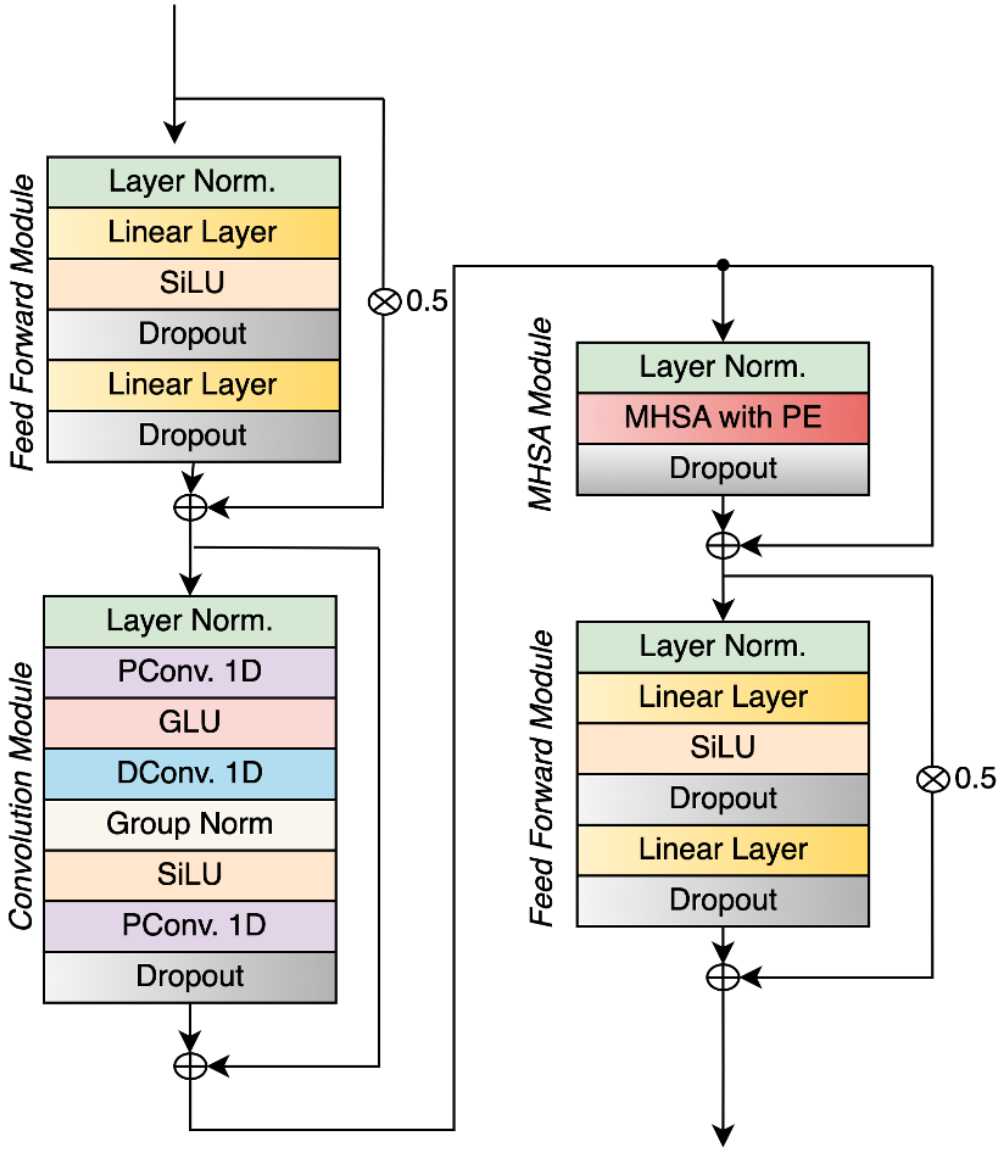

Figure 5. Flowchart of a conformer layer [15].

These three arrangements are depicted in Figures 6–8 for clarity. In these figures, the ordinary U-convolutional block and revised U-convolutional block (with conformer at the bottom of the ordinary U-Convolutional block) are depicted in Figures 9 and 10. Options (a) and (b) utilize just a single conformer layer, but option (c) employs B conformer layers. Nevertheless, the conformer layer employed in options (b) and (c) exhibits a significantly reduced number of parameters compared to the one utilized in option (a). This disparity arises from the fact that the data processed in options (b) and (c) undergo temporal downsampling, but the data in (a) does not.

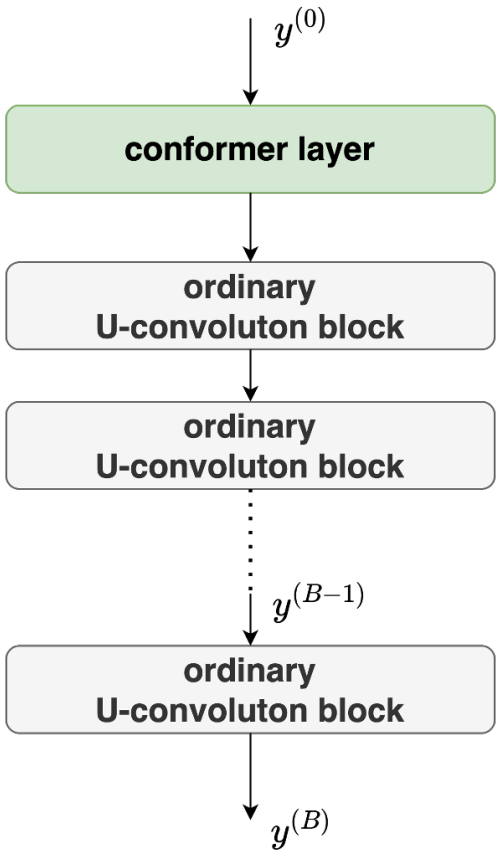

Figure 6. Option (a): the conformer layer is right before the B U-Convolutional blocks.

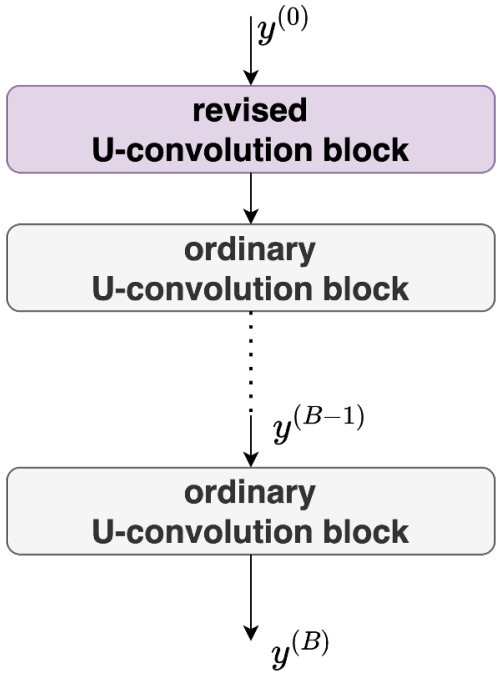

Figure 7. Option (b): the conformer layer is at the bottom of the **first** U-Convolutional block, and thus there is one revised U-Convolutional block in the beginning, concatenated with $(B-1)$ ordinary U-Convolutional blocks.

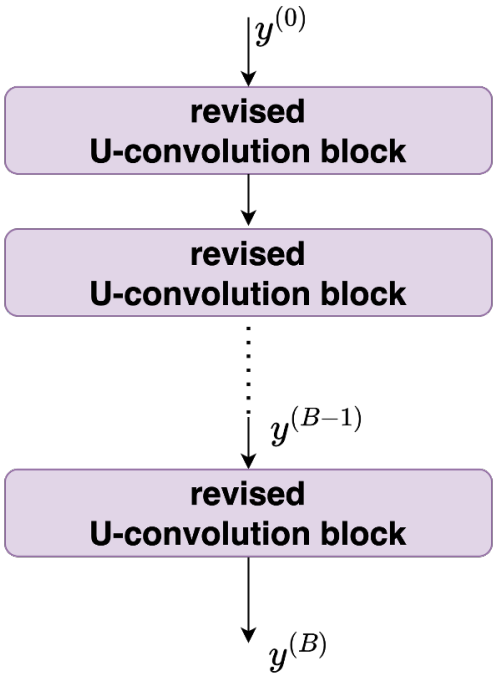

Figure 8. Option (c): the conformer layer is at the bottom of **all** U-Convolutional blocks, and thus there are B revised U-Convolutional blocks in concatenation.

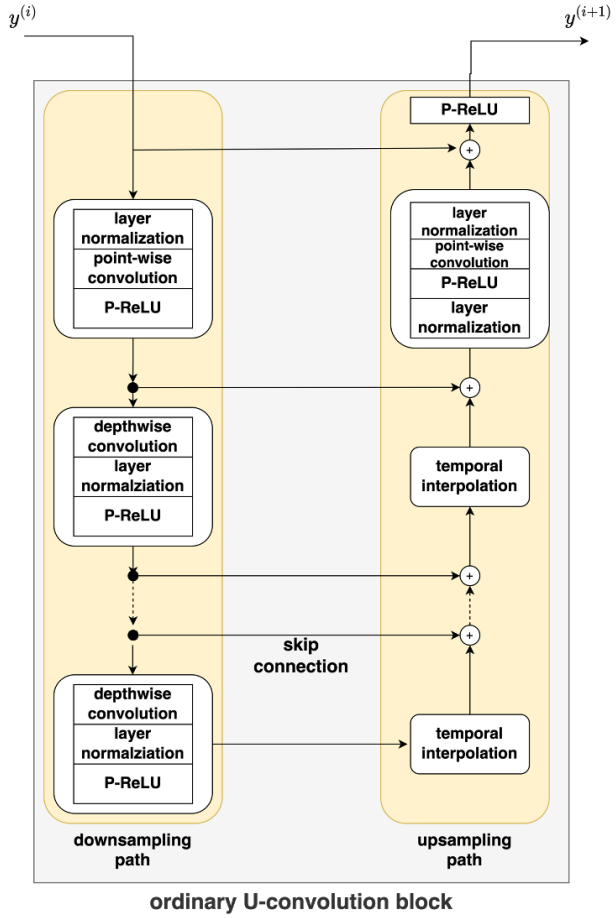

Figure 9. Flowchart of an ordinary U-Convolutional block.

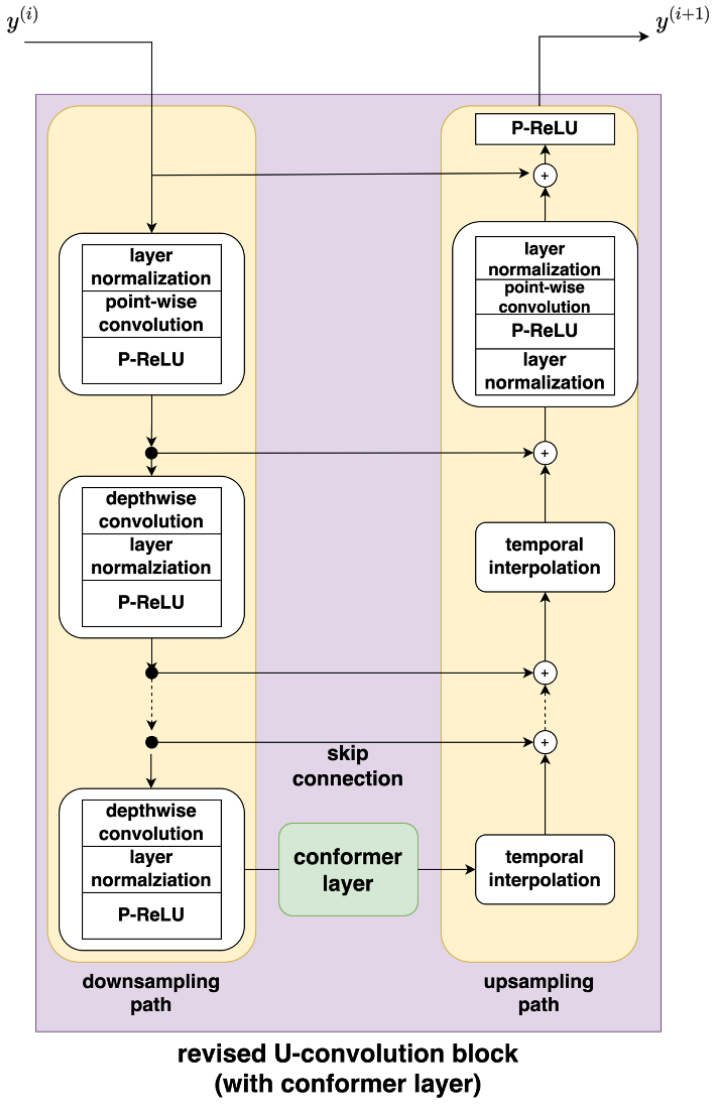

Figure 10. Flowchart of a revised U-Convolutional block, which adds a conformer layer to the bottom of an ordinary U-Convolutional block.

5. Experimental Setup

5.1. Data

We have evaluated various speech separation networks using the WHAM! dataset [16] and task. WHAM! is a noisy variant of WSJ0-2Mix [2], which is used to train models for separating speech in audio mixtures. However, the WSJ0-2Mix dataset is less practical because it cannot generalize to speeches from a wider range of speakers due to the overlap between the two speakers' audios being 100%. WHAM! improves upon this by blending speeches with background noise captured in cafes, restaurants, and bars. Furthermore, the noise samples utilized in this dataset have undergone processing to exclude any parts that contain intelligible speech. The noise in WHAM! poses a more significant barrier to mixed audios compared to WSJ0-2Mix, mostly because of its consistently sampled signal-to-noise ratio (SNR), which varies between -6 dB and 3 dB.

5.2. Data Generation and Augmentation

To create a speech mixture, the following method is used: Initially, a pair of speakers is chosen at random. Subsequently, short segments lasting 4 s are randomly selected from two audio source files. The source segments are finally combined and noise is applied at a

random signal-to-noise ratio (SNR). The training set consists of 20,000 mixtures, while the validation and test sets each contain 3000. The choices of utterances for training, validation, and test sets exactly follow the arrangement of the works SuDoRM-RF [12] and SuDoRM-RF++ [13]. All audio clips are downsampled to 8 kHz and normalized by subtracting their mean and dividing by their standard deviation.

5.3. Separation Network Configurations

For the single-channel separation task with training data under anechoic condition, the SuDoRM-RF++ architecture, as documented in [13], serves as the foundational model for ESC-MASD-Net. Hyperparameters for SuDoRM-RF++ and ESC-MASD-Net are configured to mirror those yielding optimal performance as reported in [13]. Within the encoder and decoder modules, a kernel size ($K_{\mathcal{E}}$) of 21 and a basis count ($C_{\mathcal{E}}$) of 512 are applied to input mixtures sampled at 8 kHz. Each U-ConvBlock configuration entails an input channel count (C) of 128, four successive resampling operations, and an expanded channel count of 512. Subsampling operations reduce the temporal dimension by a factor of 2, while depth-wise separable convolutions feature a kernel length of 5 and a stride of 2. The separation network incorporates 4 U-Conv Blocks ($B = 4$), aligning with the minimal configuration across diverse SuDoRM-RF versions.

Regarding the multi-view attention block, parameters include a channel count (CM) of 512, four attention heads, attention dimensions of 256, and an attention dropout rate of 0.1.

5.4. Training Objectives

The objective function used for training the speech separation networks is the scale-invariant source-to-distortion ratio (SI-SDR). This function calculates the difference between the estimated speech signal $\hat{s}$ and the clean source signal s. SI-SDR is developed to address the misuse of SDR and improve the accuracy of evaluation results.

The SI-SDR is calculated by:

$$\rho(\hat{s}, s) = 10 \log_{10} \left(\frac{\left\| \frac{\langle \hat{s}, s \rangle}{\langle s, s \rangle} s \right\|^2}{\left\| \hat{s} - \frac{\langle \hat{s}, s \rangle}{\langle s, s \rangle} s \right\|^2} \right), \tag{10}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product operation. The term $\frac{\langle \hat{s}, s \rangle}{\langle s, s \rangle} s$ denotes the orthogonal projection of the estimated speech signal $\hat{s}$ on the line spanned by the clean source signal s, while the term $\hat{s} - \frac{\langle \hat{s}, s \rangle}{\langle s, s \rangle} s$ denotes the (error) residual in $\hat{s}$ that is orthogonal to the clean source signal s. As the result, the SI-SDR is not affected by the absolute power (loudness) of $\hat{s}$ and s serves as a robust measure for speech separation.

In speech separation, the multiple clean source signals may be swapped, leading to a permutation problem. To overcome this issue, permutation invariant training (PIT) [17] is used to find the maximum SI-SDR value.

5.5. Evaluation Details

We report the scale-invariant signal-to-noise ratio improvement (SI-SDRi) as objective measures of separation accuracy. SI-SDR is defined in Equation (10). The improvement is defined as the gain between the SI-SDR of the separated signal and the SI-SDR of the mixture. According to [16], the average SI-SDR for the utterances in the 8 kHz test set of the WHAM! database is -4.5 dB. A higher SI-SDRi value indicates a better separation performance.

5.6. Programming

In order to evaluate the ESC-MASD-Net framework, we utilize publicly available scripts for SuDoRF++ [18], ResCon and MA blocks in MANNER [19], as well as the conformer layer [20]. In the near future, we will enhance the readability of the codes

associated with ESC-MASD-Net by reorganizing them and subsequently making them accessible online.

6. Experimental Results and Discussion

First, Table 1 displays the SI-SDRi scores obtained using SuDoRM-RF++ and the newly provided ESC-MASD-Net, which includes ResCon and MA blocks while retaining the original U-ConvBlock. Each of these networks is trained for 100 epochs. The table provides the following observations.

1. The given ESC-MASD-Net outperforms SuDoRM-RF++ by 1.33 in SI-SDRi, demonstrating the success of SC-MASD-Net in speech separation.
2. Regarding an ablation study, the SI-SDRi decreases by 0.11 when the MA block is removed from ESC MA-SDNet, whereas it decreases by 0.54 when the ResCon block is removed from ESC-SDNet. These findings suggest that ESC-MASD-Net may benefit more from ResCon than MA in speech separation.
3. ESC-MASD-Net with MA alone and ESC-MASD-Net with ResCon alone perform better than SuDoRM-RF++. As a result, MA and ResCon both have the potential to improve SuDoRM-RF++'s separation behavior.

Table 1. The signal-to-distortion ratio improvement (SI-SDRi) obtained by various speech separation frameworks.

Model	SI-SDRi
SuDoRM-RF++	12.38
ESC-MASD-Net	13.71
ESC-MASD-Net without MA block	13.60
ESC-MASD-Net without ResCon block	13.17

Next, we evaluate alternative extensions of ESC-MASD-Net by inserting conformer layers into different positions of its four U-Convolutional blocks. Table 2 shows the outcomes of this evaluation. Our observations are as follows.

1. The inclusion of the conformer layer to ESC-MASD-Net consistently increased the SI-SDRi score, regardless of where it was in the network. This demonstrates the utility of the conformer layer.
2. We obtained the best overall performance (SI-SDRi score of 13.95) when a single conformer layer was inserted directly before the first U-Convolutional block. Adding a conformer layer to the first U-Convolutional block (at the bottom, most downsampled portion) yielded an SI-SDRi score of 13.80. Adding a conformer layer to each of the four U-Convolutional blocks, on the other hand, resulted in an SI-SDRi score of 13.90, which was a 0.1 improvement over the previous example but required three additional conformer layers.

Furthermore, we conduct experiments to see if increasing the number of training epochs could enhance the performance of ESC-MASD-Net. The original 100 epochs setting is found to be nearly optimal for SuDoRM-RF++, but it may not be the best option for ESC-MASD-Net. We list the SI-SDRi results for ESC-MASD-Net and its two variants that incorporate a conformer layer with optimal validation performance epochs in Table 3. By increasing the number of epochs from 100 to over 150 for ESC-MASD-Net and its two variants, we observe an SI-SDRi improvement of at least 0.20. As a result, ESC-MASD-Net significantly outperforms its archetype SuDoRM-RF++, with an SI-SDRi improvement from 12.38 to 14.24.

Table 2. The signal-to-distortion ratio improvement (SI-SDRi) obtained by Effective Speaker separation through convolutional multi-view attention and SuDoRM-RF network (ESC-MASD-Net) with conformer at different locations.

Model		SI-SDRi
ESC-MASD-Net		13.71
ESC-MASD-Net with Conformer layer	(a) right before the U-Convblocks	13.95
	(b) at the bottom of the first U-Convblock	13.80
	(c) at the bottom of all four U-Convblocks	13.90

Table 3. The signal-to-distortion ratio improvement (SI-SDRi) obtained by effective speaker separation through convolutional multi-view attention and SuDoRM-RF network (ESC-MASD-Net) variants using different number of epochs.

Model		Epoch	SI-SDRi
SuDoRM-RF++		100	12.38
ESC-MASD-Net		100	13.71
		153	13.94
ESC-MASD-Net with a single conformer layer	(a) right before the U-Convblocks	100	13.95
		151	14.24
	(b) at the bottom of the first U-Convblock	100	13.80
		153	14.02

Finally, Table 4 includes the model size and achieved SI-SDRi results of some state-of-the-art speech separation methods performed on the Wham! dataset. From this table, we see that our best model achieved the SI-SDRi score of 14.24 using 3.6M parameters. We also observed that our best model performs on par or even outperforms several state-of-the-art methods with much higher computational resource needs.

Table 4. Model size and performance comparison of the presented effective speaker separation through convolutional multi-view attention and SuDoRM-RF network (ESC-MASD-Net) with the other state-of-the-art speech separation models on WHAM!

Model	Parameters (Millions)	SI-SDRi
Conv-TasNet [4]	5.1	12.7
DPRNN [9]	2.7	13.9
SuDoRM-RF2.5x [12]	6.4	14.1
Sepformer [7]	26	15.0
MossFormer [21]	42.1	17.3
MossFormer2 [22]	55.7	18.1
ESC-MASD-Net (Ours)	3.6	14.24

7. Conclusions

The purpose of this study is to improve the state-of-the-art speech separation framework, SuDoRM-RF++, by incorporating it with a number of different convolution-based modules. These modules include residual conformer network, multi-attention network, and conformer layers. Having demonstrated that each additional module offers greater

separation performance, we also show that these modules can work together to lead to even better outcomes than can be achieved by each module functioning independently. We will, in the future, research ways to lessen the complexity of the presented ESC-MASD-Net without compromising its performance.

Author Contributions: Conceptualization, A.N.A.; Methodology, A.N.A., C.-W.L. and J.-W.H.; Software, J.-W.H.; Validation, C.-W.L. and J.-W.H.; Investigation, A.N.A. and C.-W.L.; Data curation, A.N.A.; Writing—original draft, A.N.A. and J.-W.H.; Writing—review & editing, J.-W.H.; Visualization, C.-W.L.; Supervision, J.-W.H.; Project administration, J.-W.H.; Funding acquisition, J.-W.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data that supports the findings of this study requires WSJ0 corpus and the WHAM noise corpus.The WSJ0 corpus which is a base corpus required to construct the wsj0-2mix dataset is available for download at https://doi.org/10.35111/ewkm-cg47 upon request under a specific license agreement. The set of WHAM noise samples along with the scripts to build WHAM! dataset from wsj0-2mix dataset are openly available to download at http://wham.whisper.ai/.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Wang, Y.; Narayanan, A.; Wang, D. On training targets for supervised speech separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2014**, *22*, 1849–1858. [CrossRef] [PubMed]
2. Hershey, J.R.; Chen, Z.; Roux, J.L.; Watanabe, S. Deep clustering: Discriminative embeddings for segmentation and separation. *arXiv* **2016**, arXiv:1508.04306.
3. Isik, Y.; Roux, J.L.; Chen, Z.; Watanabe, S.; Hershey, J.R. Single-channel multi-speaker separation using deep clustering. *arXiv* **2016**, arXiv:1607.0217.
4. Luo, Y.; Mesgarani, N. Tasnet: Time-domain audio separation network for real-time, single-channel speech separation. *arXiv* **2018**, arXiv:1711.00541.
5. Luo, Y.; Mesgarani, N. Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2019**, *27*, 1256–1266. Available online: https://dl.acm.org/doi/10.1109/TASLP.2019.2915167 (accessed on 30 March 2024). [CrossRef] [PubMed]
6. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *arXiv* **2017**, arXiv:1706.03762.
7. Subakan, C.; Ravanelli, M.; Cornell, S.; Bronzi, M.; Zhong, J. Attention is all you need in speech separation. *arXiv* **2021**, arXiv:2010.13154.
8. Chen, J.; Mao, Q.; Liu, D. Dual-path Transformer network: Direct context-aware modeling for end-to-end monaural speech separation. In Proceedings of the Interspeech 2020, Shanghai, China, 25–29 October 2020; Available online: http://www.interspeech2020.org/uploadfile/pdf/Wed-2-4-6.pdf (accessed on 30 March 2024).
9. Luo, Y.; Chen, Z.; Yoshioka, T. Dual-path rnn: Efficient long sequence modeling for time-domain single- channel speech separation. *arXiv* **2020**, arXiv:1910.06379.
10. Maldonado, A.; Rascon, C.; Velez, I. Lightweight online separation of the sound source of interest through BLSTM-based binary masking. *arXiv* **2020**, arXiv:2002.11241.
11. Li, K.; Yang, R.; Hu, X. An efficient encoder-decoder architecture with top-down attention for speech separation. *arXiv* **2023**, arXiv:2209.15200.
12. Tzinis, E.; Wang, Z.; Smaragdis, P. Sudo rm -rf: Efficient networks for universal audio source separation. *arXiv* **2020**, arXiv:2007.06833.
13. Tzinis, E.; Wang, Z.; Jiang, X.; Smaragdis, P. Compute and memory efficient universal sound source separation. *J. Signal Process. Syst.* **2022**, *94*, 245–259. [CrossRef]
14. Park, H.J.; Kang, B.H.; Shin, W.; Kim, J.S.; Han, S.W. Manner: Multi-view attention network for noise erasure. *arXiv* **2022**, arXiv:2203.02181.
15. Ravenscroft, W.; Goetze, S.; Hain, T. On time domain conformer models for monaural speech separation in noisy reverberant acoustic environments. *arXiv* **2023**, arXiv:2310.06125.
16. Wichern, G.; Antognini, J.; Flynn, M.; Zhu, L.R.; McQuinn, E.; Crow, D.; Manilow, E.; Roux, J.L. Wham!: Extending speech separation to noisy environments. *arXiv* **2023**, arXiv:1907.01160.
17. Yu, D.; Kolbæk, M.; Tan, Z.; Jensen, J. Permutation invariant training of deep models for speak-er-independent multi-talker speech separation. *arXiv* **2017**, arXiv:1607.00325.
18. Available online: https://github.com/etzinis/sudo_rm_rf (accessed on 30 March 2024)
19. Available online: https://github.com/winddori2002/MANNER (accessed on 30 March 2024)

20. Available online: https://github.com/jwr1995/pubsep (accessed on 30 March 2024)
21. Zhao, S.; Ma, B. MossFormer: Pushing the Performance Limit of Monaural Speech Separation using Gated Single-Head Transformer with Convolution-Augmented Joint Self-Attentions. *arXiv* **2023**, arXiv:2302.11824.
22. Zhao, S.; Ma, Y.; Ni, C.; Zhang, C.; Wang, H.; Nguyen, T.H.; Zhou, K.; Yip, J.; Ng, D.; Ma, B. MossFormer2: Combining Transformer and RNN-Free Recurrent Network for Enhanced Time-Domain Monaural Speech Separation. *arXiv* **2024**, arXiv:2312.11825.