IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 13 June 2022, accepted 10 July 2022, date of publication 21 July 2022, date of current version 1 August 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3193245

RESEARCH ARTICLE

Dual-Path Hybrid Attention Network for Monaural Speech Separation

WENBO QIU AND YING HU
School of Information Science and Engineering, Xinjiang University, Ürümqi 830046, China
Key Laboratory of Signal Detection and Processing, Xinjiang, Ürümqi 830046, China

Corresponding author: Ying Hu (huying@xju.edu.cn)

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant U1903213, and in part by the Tianshan Innovation Team Plan Project of Xinjiang under Grant 202101642.

ABSTRACT Recent advances in the time-domain speech separation methods, particularly those specialized in using attention mechanisms to model sequences, have significantly improved speech separation performance. In this paper, we address monaural (one microphone) speaker separation, mainly in the case of two concurrent speakers. We propose a dual-path hybrid attention network (DPHA-Net) for monaural speech separation based on time-domain. The critical component of DPHA-Net, the DPHA module, comprises multiple attentions and is designed to capture the short and long-term context information dependencies. DPHA module consists of the multi-head self-attention (MHSA), element-wise attention (EA), and adaptive feature fusion (AFF) units. We proposed an improved multi-stage aggregation training strategy during the training. That strategy has proven very effective for audio separation in this paper. The results of experiments on the benchmark WSJ0-2mix, WHAM! and Libri2Mix datasets show that our proposed DPHA-Net can achieve the competitive performance. For the task of two speaker separation on the WSJ0-2mix dataset, our proposed DPHA-Net is superior to the state of the art with a margin of 0.3 dB absolute improvement on the SI-SNRi and a margin of 0.4 dB absolute improvement on the SDRi in the same condition.

INDEX TERMS Speech separation, hybrid attention, multi-stage aggregation training.

I. INTRODUCTION

Speech separation, commonly called the ‘‘cocktail party problem’’, is a fundamental task in signal processing with a wide range of applications [1], [2]. This task has been shown to be difficult for computers, especially when only a monaural recording of the mixed speech is available. Over the last decade, the performance of speech separation has been substantially improved by leveraging extensive training data and increasing computing resources.

In general, the speech separation methods based on deep learning can be divided into two categories: time-frequency (T-F) domain and time-domain methods. For the method based on the T-F domain, each mixture spectrogram calculated by short-time Fourier transform (STFT) is served as the input of the separation model to approximate the clean spectrogram of individual sources. A parallel stacked hourglass network [3] was also proposed to learn the features of a multi-band spectrogram, which is a fully convolutional multi-scale end-to-end network for music separation. Computational auditory scene analysis (CASA) [4] is a traditional approach for source separation, which is inspired by human auditory scene analysis (ASA) mechanisms [5]. Deep CASA [6], address talker-independent monaural speaker separation from the perspectives of the deep learning and CASA. Talker-independent speaker separation has to address the permutation problem of how the output layers are tied to the underlying speakers. DPCL avoids the permutation problem due to the permutation-invariant property of affinity matrices [7]. The deep clustering methods infer a set of source representations via clustering to separate two or three speaker speech waveforms [7], [8]. Along with either the original [9]–[11] or modified [12]–[14] phase of mixture audio, those estimated spectrogram of each source

The associate editor coordinating the review of this manuscript and approving it for publication was Manuel Rosa-Zurera.

are converted into waveform by inverse short-time Fourier transform (iSTFT).

For the method based on the time domain, the mixture waveform is directly modeled by the separation framework. Conv-TasNet [15], which is a fully convolutional time-domain audio separation network, directly modeling the mixture waveform using an encoder-separator-decoder framework. Similar to Conv-TasNet, two methods [16], [17] use U-Net architecture instead of dilated depth-wise convolution in the separator module. A two-step training procedure for source separation [18], which can work directly on the latent space and learn the ideal masks on a separate step, yields a consistent performance improvement under multiple sound separation tasks. In this paper, we focus on monaural speech separation based on the time domain.

A dual-path framework was firstly introduced in DPRNN [19], which splits the long sequential input into smaller chunks and applies intra- and inter-chunk operations iteratively. Currently, several state-of-the-art time-domain methods [20]–[24] are also followed with the dual-path framework. The self-attention mechanism [26] is applied to model the global dependency after chunk operation and verified that it can capture context dependency [20], [22]–[25]. The dual attention mechanism is proposed to capture global dependencies in the spatial and channel dimensions [27] or temporal and spectral dimensions [28], [29]. Kim and Hahn proposed a two-stage based approach to boost the performance of speech enhancement [30]. Li *et al.* proposed a multi-stage architecture for the temporal action segmentation in videos. Each stage takes an initial prediction from the previous stage and refines it [31]. Gao *et al.* proposed a hierarchical constraint strategy to regularize the training, which could effectively improve the separation performance [32].

Inspired by the abovementioned methods, we propose a dual-path hybrid attention network (DPHA-Net) with a multi-stage aggregation training strategy (MAT) for monaural speech separation. The key component of DPHA-Net, dual-path hybrid attention (DPHA) module, comprises multiple attentions mechanism. The module, via different types of attention, captures the feature information dependency along various dimensions. The DPHA modules are stacked iteratively, alternating between intra- and inter-chunk operations, to obtain contextual information. We also design an adaptive feature fusion unit to fuse multiple feature maps. In this study, we also adopt the multi-stage training strategy but improve its feature selective aggregation ability in significant ways, resulting in what we call a multi-stage aggregation training mechanism.

The rest of this paper is organized as follows. Section II describes our proposed method, including the DPHA module and multi-stage aggregation training mechanism. Section III presents experimental setup. Section IV presents experimental results, comparisons and analysis. Conclusion and related issues are discussed in Section V.

II. ALGORITHM DESCRIPTION

In this section, we introduce an end-to-end time-domain network for two speaker separation. Specifically, we review the pipeline at first and then start with the complete DPHA-Net architecture and introduce the DPHA module and AFF unit. Finally, we introduce a multi-stage aggregation training strategy.

A. OVERALL PIPELINE

An algorithm dedicated to performing talker-independent speaker separation aims to extract individual speech signals of different speakers from a mixture. Considering each mixture speech consists of C sources, the waveform of mixture speech is formulated as:

$$x(t) = \textstyle\sum_{i=1}^{C} s_i(t) \tag{1}$$

where $\mathbf{x(t)} \in \mathbb{R}^{1\times T}$ denotes the waveform of mixture speech, $\mathbf{s_i(t)} \in \mathbb{R}^{1\times T}$ that of speaker i. T is the length of signal, C the number of speakers. In this work, we focus on the single-channel speech separation where $C = 2$. As shown in Figure 1, the separation architecture we adopted is the same pipeline as Conv-TasNet [15], performing end-to-end audio source separation using a mask-based architecture with adaptive encoder and decoder basis.

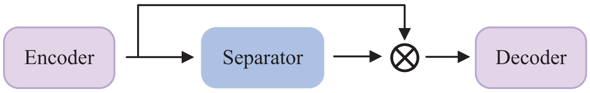

FIGURE 1. (Color Online). Diagram of End-to-End architecture.

B. NETWORK ARCHITECTURE

As shown in Figure 2, DPHA-Net consists of three parts: encoding and chunking, separation processing, and decoding and overlap-add. A time-domain mixture waveform, which comprises speech of $C = 2$ speakers with the length of T, is transformed into a 2-D representation via an encoder and further a 3-D representation via chunking processing. Subsequently, the 3-D representation is passed into stacked DPHA modules to perform local modeling (intra-chunk) and global modeling (inter-chunk) alternately and iteratively. The output of the last DPHA module, together with previous fused representations, are further fused to generate the final representation. Finally, this representation undergoes an overlap-add method to estimate a multiplicative function (mask) for each source. The output from the encoder is multiplied with the mask and transformed back to C estimated waveforms with a decoder. Figure 2 shows the flowchart of the network.

The encoder consists of a 1-D convolutional layer with E output channels and a rectified linear unit (ReLU) activation function. The mixture is transformed into a sequential input $\mathbf{x} \in \mathbb{R}^{E\times I}$ via an encoder, where E is the feature dimension, and I is the number of time steps. Via a bottleneck layer (i.e. 1×1 convolution layer) with a global layer normalization (GLN), $\mathbf{x}$ is transformed into $\mathbf{m}$ ($\mathbb{R}^{E\times I} \rightarrow \mathbb{R}^{N\times I}$). Then $\mathbf{m} \in \mathbb{R}^{N\times I}$ is split into overlapped S chunks with the length of J

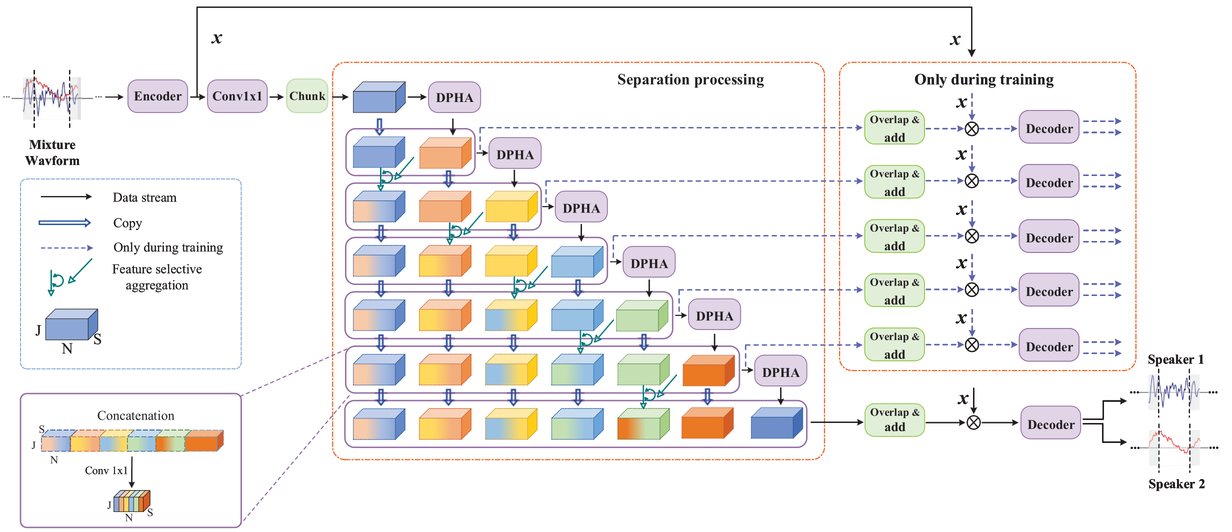

FIGURE 2. (Color Online). Diagram of DPHA-Net architecture. The purple dash lines represent the procedures that exist only during training. $Conv_{1\times1}$ represents 1 × 1 convolutional operation and here also represents the bottleneck layer. ⊗ the element-wise multiplication. The cubes represent feature maps $\mathbb{R}^{N\times J\times S}$. The purple line box represents layer aggregation operation.

and hop size $J/2$. Those chunks are concatenated into a 3-D tensor $\mathbf{X} \in \mathbb{R}^{N\times J\times S}$.

In the part of separation processing, $\mathbf{X}$ is fed into a series of DPHA modules. As shown in Figure 2, there are six stages ($L = 6$) in the part of separation processing. In each stage, the output of the DPHA module, together with previous selective aggregation representations, are further aggregated to generate the final representation. It undergoes a 2-D convolutional layer, and then overlap-add operation at the chunk level and frame level, respectively to generate sequences masks $M_c \in \mathbb{R}^{E\times I}, c = 1, \ldots, C$ for C sources.

$$Y_c = x \otimes M_c \tag{2}$$

C output embedding $\{Y_c, c = 1, \ldots, C\}$ of each stage are decoded by a decoder which contains a 1-D transposed convolutional layer. The decoded representation with the size of $C \times T$ represents C estimated waveforms with the length of T.

C. DPHA MODULE

As shown in Figure 3, the DPHA module consists of two subblocks, one for intra-chunk modeling and the other for inter-chunk modeling. Each subblock comprises three units: multi-head self-attention (MHSA), element-wise attention (EA), and adaptive feature fusion (AFF) units, and two operations: LN [33] and permutation. In each subblock, we employ a MHSA unit followed by an EA unit and AFF unit. Multi-head self-attention can model the relationship among each group feature and capture the long-term dependencies in different time steps [26].

Before being fed into the MHSA unit, 3-D representation $\mathbf{X}$ is divided into a series of 2-D slices $\mathbf{U} \in \mathbb{R}^{O\times N}$, where $O = J$ for intra-chunk modeling and $O = S$ for inter-chunk modeling. Each 2-D slice is passed into three linear layers and generate a query matrix $\mathbf{Q}$, a key matrix $\mathbf{K}$, and value matrix $\mathbf{V}$, respectively. Where $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{O\times F}$, F is the hidden layer dimension. The MSHA is calculated as the follows:

$$A_h = Softmax(\frac{Q_h^\top K_h}{\sqrt{F/H}})V_h \tag{3}$$

$$A = Concat[A_1, \ldots, A_H] \cdot W \tag{4}$$

$$MHSA(X) = P(X, A(GLN(X))) \tag{5}$$

where $\mathbf{Q_h}, \mathbf{K_h}, \mathbf{V_h} \in \mathbb{R}^{O\times \frac{F}{H}}$. h indices the attention head. H is the number of attention heads, $\mathbf{W} \in \mathbb{R}^{F\times F}$ the linear transformation matrix, $Concat[\cdot]$ the concatenation operation and A_h the result of self-attention operation. A is the result of multi-head self-attention, and then passed into an FC layer with PReLU activation for further transformation. $P(\cdot)$ the concatenation operation along channel dimension and followed with 1×1 convolution. $GLN(\cdot)$ denotes the global layer normalization operation [15] that the features $\mathbf{X} \in \mathbb{R}^{N\times I}$ are normalized over both the channel and time dimensions.

$$GLN(X) = \frac{X-\mu}{\sigma}\gamma + \beta,$$
$$\mu = \frac{1}{NI}\sum_{NI} F, \quad \sigma = \sqrt{\frac{1}{NI}\sum_{NI}(X-\mu)^2 + \epsilon} \tag{6}$$

where $\gamma, \beta \in \mathbb{R}^{N\times 1}$ are trainable parameters, and ϵ is a small constant for numerical stability. The $GLN(\cdot)$ significantly improves the convergence of our proposed models probably because of the interdependence of the gradient statistics between the channels [15].

The original DPRNN [19] applies a simple bidirectional long short-term memory (BLSTM) layer. While Gated-DPRNN [21] consists of two BLSTM layers where the Hadamard product of their outputs is concatenated with the input and then passed into a linear projection layer for dimension reduction. Inspired from these, the EA unit in DPHA-Net

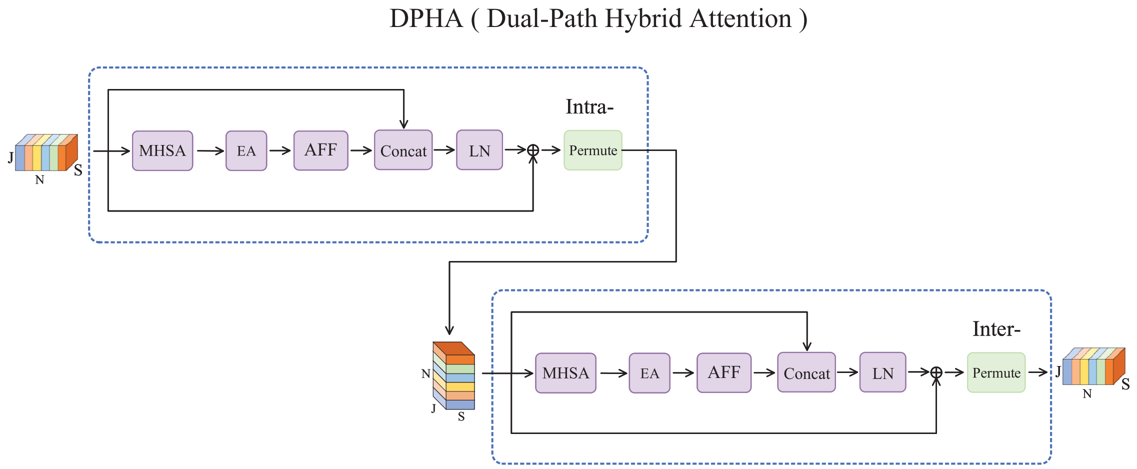

FIGURE 3. (Color Online). Diagram of the Dual-Path Hybrid Attention (DPHA) Module. ⊕ represents the element-wise addition.The input cubes represent feature maps $\mathbb{R}^{N\times J\times S}$.The middle and output feature representations after permutation are $\mathbb{R}^{N\times S\times J}$ and $\mathbb{R}^{N\times J\times S}$, respectively.

our proposed consists of two gated recurrent unit (GRU) layers, but one path of GRU is followed by a sigmoid function. The EA unit is used to capture the context dependencies among different time steps, so it named as element-wise attention. Two GRU and sigmoid function are denoted as B^1, B^2 and $\sigma(\cdot)$, respectively. The result of the EA unit $\mathbf{Z}$ is calculated as:

$$Z = P(B^1(X) \odot \sigma(B^2(X)), X) \tag{7}$$

where $\odot$ denotes the element-wise product operation, $P(\cdot)$ the concatenation operation along channel dimension and followed with 1 × 1 convolution.

D. ADAPTIVE FEATURE FUSION UNIT

In order to enhance the capability of feature extraction from the relevant time frames and channels, we design an AFF unit that can guide the network to pay suitable attention to the temporal and channel-wise characteristics. As shown in Figure 4, the part of feature extraction consists of two squeeze-and-excitation (SE) operations [34]. Via the SE operation, each branch focuses on discerning specific local region features rather than being spread evenly over the whole feature map, which leads to better robustness. The part of adaptive fusion consists of three convolution operations to adaptively obtain the learnable weights matrices. $\mathbf{Z_1}$ and $\mathbf{Z_3}$ are the output from channel-wise attention and temporal attention operation, respectively. The original feature representation $\mathbf{Z_2}$, $\mathbf{Z_1}$ and $\mathbf{Z_3}$ are passed into learnable 1 × 1 convolutional layers, respectively. At last, a summation operation of three branches completes the adaptive feature fusion of triple path. The output feature after AFF unit is calculated as following:

$$\hat{Z} = \textstyle\sum_{i=1}^{3} Conv1_i(Z_i) \tag{8}$$

where $Conv1_i$ denotes 1 × 1 convolution.

E. MULTI-STAGE AGGREGATION TRAINING STRATEGY

In previous speech separation methods [15], [16], [18]–[20], [23], [24], the last stage obtained the estimated signals to accomplish the separation task. In [21] and [22], the multi-scale loss was used to calculate the result of different separation stages during training, which requires reconstructing the estimated audio after each stage.

In the dense connectedly network, the output of a layer will never be modified once it is produced. Since shallow features will be repeatedly processed by their following layers, exploiting them in deep layers might be inefficient or redundant. Sparse feature reactivation (SFR) was proposed to reduce the redundancy in dense connections and keep the feature map always “fresh” at each dense layer [35]. Deep Aggregation [36] was proposed to fuse information across layers better and merge the feature iteratively and hierarchically that it makes the networks obtain better separation performance with fewer parameters.

Inspired by these work, we proposed a multi-stage aggregation training strategy, which consists of a multi-stage training strategy (MSTS) and a feature-selective aggregation mechanism (FSAM). Different from the SFR, the FSAM only consists of the group convolution stage but without the sparsification stage. The FSAM increases the utility of feature reusing. Giving the input feature $\mathbf{X_0}$, the output of feature aggregation operation $\tilde{\mathbf{X}}_l$ is calculated as following:

$$X_l^{in} = \begin{cases} LA\left(X_0\right), & l = 1 \\ LA\left(X_0, X_1\right), & l = 2 \\ LA\left(\tilde{X}_2, \cdots, \tilde{X}_{l-1}, X_{l-2}, X_{l-1}\right), & 3 \leq l \leq L \end{cases} \tag{9}$$

$$X_l = D_l\left(X_l^{in}\right) \tag{10}$$

$$\tilde{X}_l = \begin{cases} X_0, & if \quad l = 1 \\ R_l\left(X_{l-2}, X_{l-1}\right), & 2 \leq l \leq L \end{cases} \tag{11}$$

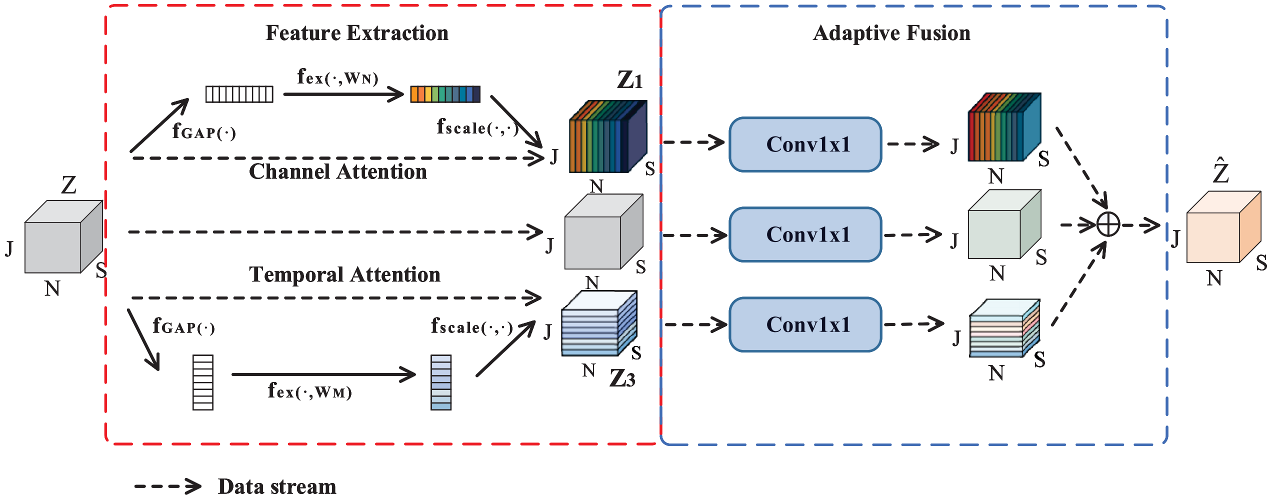

FIGURE 4. (Color Online). Diagram of the Adaptive Feature Fusion (AFF) Unit. The red dash box and blue one represent the feature extraction of triple-path and adaptive feature fusion, respectively. $f_{GAP}(\cdot)$ and $f_{ex}(\cdot, W)$ denote the global average pooling operation and gating mechanism with a sigmoid activation operation. $f_{scale}(\cdot)$ denotes element-wise multiplication between the feature map and scalar $f_{ex}(\cdot, W)$. The input and output feature representations are $\mathbf{Z} \in \mathbb{R}^{N \times J \times S}$ and $\hat{\mathbf{Z}} \in \mathbb{R}^{N \times J \times S}$, respectively.

where l indices the stage, $LA(\cdot)$ the layer aggregation operation, $D_l(\cdot)$ the function defined by the l-th DPHA module, and $R_l(\cdot)$ the feature selective function which consists of a 1×1 group convolution [37] followed by a LN and $ReLU$. In each $R_l(\cdot)$, the input is divided into $N/4$ groups over the channel dimension. The aggregation operation, $LA(\cdot)$, is formulated as follows:

$$LA(X_1, \cdots, X_n) = ReLU(BN(Conv([X_1, \cdots, X_n]))) \quad (12)$$

where $[\cdot]$ denotes that the feature maps $\{\mathbf{X_i}, i \in [1, \ldots, n]\}$ are concatenated along the channel dimension. Then, the concatenated feature map is transformed via a convolution layer with both the kernel size and stride size of 1, followed by BN and ReLU activation functions. By this way, the aggregations of features of different stages can enrich temporal context information in deep features. The context information between channels is enhanced by combining the feature channels of different levels. The following experimental results validate our hypothesize: although the features produced by early layers seem unimportant at deep layers in dense networks, they may have potential after being reactivated.

Similar to [38], a weighted loss function with utterance-level permutation invariant training (uPIT) [39] is adopted. The stages more deeper, the obtained semantic information more closer to the separation task, thus we assign the larger weights to the outputs of deeper layers in the weighted scale-invariant source-to-noise ratio (SI-SNR) [15] loss function:

$$\begin{cases} s_{target} := \frac{\langle \hat{s}, s \rangle s}{\|s\|^2} \\ e_{noise} := \hat{s} - s_{target} \\ si - snr := 10 \, log_{10} \frac{\|s_{target}\|^2}{\|e_{noise}\|^2} \end{cases} \quad (13)$$

$$Loss = \frac{\sum_{i=1}^{I} i \cdot L_{si-snr}}{\sum_{i=1}^{I} i} \quad (14)$$

where $\mathbf{s} \in \mathbb{R}^{1 \times T}$ and $\hat{\mathbf{s}} \in \mathbb{R}^{1 \times T}$ are the original clean sources and estimated sources, respectively. $\|s\|^2$ denotes the signal power. $i/\sum_{i=1}^{I} i$ is the weight of loss at the i-th stage. The waveforms of clean utterances are used as training targets for calculating the loss of each stage. The intuition behind this is that the output of the deeper stage can approximate the separation task better. With this multi-stage aggregation training strategy, the feature selective aggregation mechanism can assist the network to adaptively select useful information.

III. EXPERIMENTAL SETUP

A. DATASET

We evaluated DPHA-Net on the two-speaker separation task using WSJ0-2mix [7], WHAM! [41] and Libri2Mix [40], respectively. The WSJ0-2mix dataset has 30 hours of training data and 10 hours of validation data. The dataset also has 5 hours of evaluation data which generated, in the same way, using utterances from 18 unseen speakers in the validation set si_dt_05 and evaluation set si_et_05. WHAM! added noise to WSJ0-2mix, which was recorded in different scenes. The Libri2Mix, derived from the LibriSpeech corpus [42] with random extracts were selected for different speakers and mixed with uniformly sampled Loudness Units relative to Full Scale (LUFS) [43] between -25 and -33 dB, contains two training sets (*train-100*, *train-360*), one dev set, and one test set. We used the training data of WSJ0-2mix and WHAM!, and *train-100* set of Libri2Mix for training. All signals are sampled at 8 kHz. We used the same test set to compare different methods. Table 1 reports the dataset statistics.

B. MODEL CONFIGURATIONS

We trained the models for 200 epochs on 4-second long segments with an Adam optimizer [44]. Batch size is set to 8, and the initial learning rate 0.001. The learning rate is decayed to 0.98 times if the accuracy of the validation set

TABLE 1. Statistics of WSJ0-2mix, WHAM! and Libri2Mix.

Dataset	Parameter	train	valid	test
WSJ0-2mix	#examples	20,000	5,000	3,000
&	#speakers	101	101	18
WHAM!	mean length	5.4s	5.5s	5.7s
Libri2Mix	#examples	13,900	3,000	3,000
	#speakers	251	40	40
	mean length	15.0s	13.2s	13.2s

TABLE 2. Hyperameters in DPHA-Net.

Symbol	Hyperparameter	Configuration
E	Number of filters in codec	128
W	Length of the filters (in samples)	4
B	Number of channels in bottleneck	64
N	Number of channels in DPHA blocks	64
H_i/H_o	Input/hidden dimensions in GRU layer	64/128
L	Number of DPHA blocks	6
H	Number of attention heads	4
J	Length of chunks	180

is not improved in 2 consecutive epochs. Gradient clipping by a maximum gradient norm of 5 is always applied for proper convergence. Early stopping is applied when no best validation model is found for eight consecutive epochs. The hyperparameters of DPHA-Net are listed in Table 2 and codes implementation are available online.[1]

C. TRAINING OBJECTIVE AND EVALUATION METRICS

In this paper, the objective of training the end-to-end system is the weighted SI-SNR, which is the Equation (14). We report the scale-invariant signal-to-noise ratio improvement (SI-SNRi) [15], [45], signal-to-distortion ratio improvement (SDRi) [46], perceptual evaluation of subjective quality (PESQ) [47] and short-time objective intelligibility (STOI) [48] as objective measures of separation accuracy. The SDRi and SI-SNRi are defined as:

$$SDRi(\hat{s}, s, x) = SDR(\hat{s}, s) - SDR(x, s) \tag{15}$$

$$SI - SNRi(\hat{s}, s, x) = SI - SNR(\hat{s}, s) - SI - SNR(x, s) \tag{16}$$

where $\mathbf{x}$ is mixed speech, $\hat{\mathbf{s}}$ is the estimated target speech and $\mathbf{s}$ is the reference target speech.

IV. EXPERIMENTAL RESULTS

In this section, we make a comparison among different methods of the two-speaker separation task on two datasets. For the WSJ0-2mix dataset, the results of SDRi, SI-SNRi, PESQ and STOI are listed, and for the Libri2Mix dataset, that of the SDRi and SI-SNRi are listed according with compared methods.

TABLE 3. Comparison with 10 methods in the two-speaker condition on the WSJ0-2mix dataset. '–' means the corresponding result not reported. '*' only SI-SNR and SDR (without improvement) are reported. Generally, SI-SNRi and SDRi are smaller than SI-SNR and SDR.

Metrics	# param	SI-SNRi(dB)	SDRi(dB)	STOI(%)	PESQ
mixture	–	0.00	0.00	56.10	1.99
Conv-TasNet [15]	5.1 M	15.3	15.6	–	3.24
Two-Step CTN [18]	8.6 M	16.1	–	–	–
Deep CASA [6]	12.8 M	17.7	18.0	93.2	3.51
A-FRCNN [17]	6.1 M	18.3	18.6	–	–
DPRNN [19]	2.6 M	18.8	19.0	92.64	3.49
Sudo rm -rf [16]	2.6 M	18.9	–	–	–
Gated-DPRNN [21]	7.5 M	20.1	20.4	92.52	3.52
DPTNet* [20]	2.6 M	20.2	20.6	–	–
GALR [23]	2.3 M	20.3	20.5	–	–
SepFormer [24]	26 M	20.4	20.5	–	–
DPHA-Net	6.1 M	20.7	20.9	98.14	3.70

TABLE 4. Comparison with 7 methods in the two-speaker condition on the Libri2Mix dataset.

Metrics	# param	SI-SNRi(dB)	SDRi(dB)
DANET [10]	9.1 M	8.4	–
Conv-TasNet [15]	5.1 M	12.61	13.02
SANet [8]	–	12.8	–
GCD-TasNet [49]	–	13.63	14.03
DPRNN [19]	2.6 M	14.57	14.99
DPTNet [20]	2.6 M	14.90	15.37
A-FRCNN [17]	6.1 M	16.7	17.2
DPHA-Net	6.1 M	16.51	16.96

A. COMPARISON BETWEEN DPHA-NET WITH PREVIOUS METHODS

We make a comparison between our proposed DPHA-Net with some state-of-the-art models reported in the literature recently.

Table 3 lists the results of the DPHA-Net and 10 compared methods on the WSJ0-2mix dataset: Conv-TasNet [15], Two-Step CTN [18], Deep CASA [6], A-FRCNN [17], DPRNN [19], Sudo rm -rf [16], Gated-DPRNN [21], DPTNet [20], GALR [23], SepFormer [24]. As can be seen, our proposed DPHA-Net is superior to the compared methods by a certain margin. In addition, SepFormer, whose performance is closest to DPHA-Net, has quite a few parameters compared with the DPHA-Net. As seen from Figure 5, the spectrogram of separated speech by the DPHA-Net is more closer to clean speech comparing with the Gated-DPRNN and DPRNN.

Table 4 lists the results of the DPHA-Net and 7 compared methods on the Libri2Mix dataset: DANET [10], Conv-TasNet [15], SANet [8], GCD-TasNet [49], DPRNN [19], DPTNet [20], A-FRCNN [17]. As we can seen, in general, the proposed DPHA-Net is superior to compared methods but slightly less than the A-FRCNN. Meanwhile, comparing with the results of the WSJ0-2mix dataset, the separation performance of the same method such as DPHA-Net, DPRNN and DPTNet on the Libri2Mix dataset is not as excellent as

[1] https://github.com/wbQIU-xju/DPHA

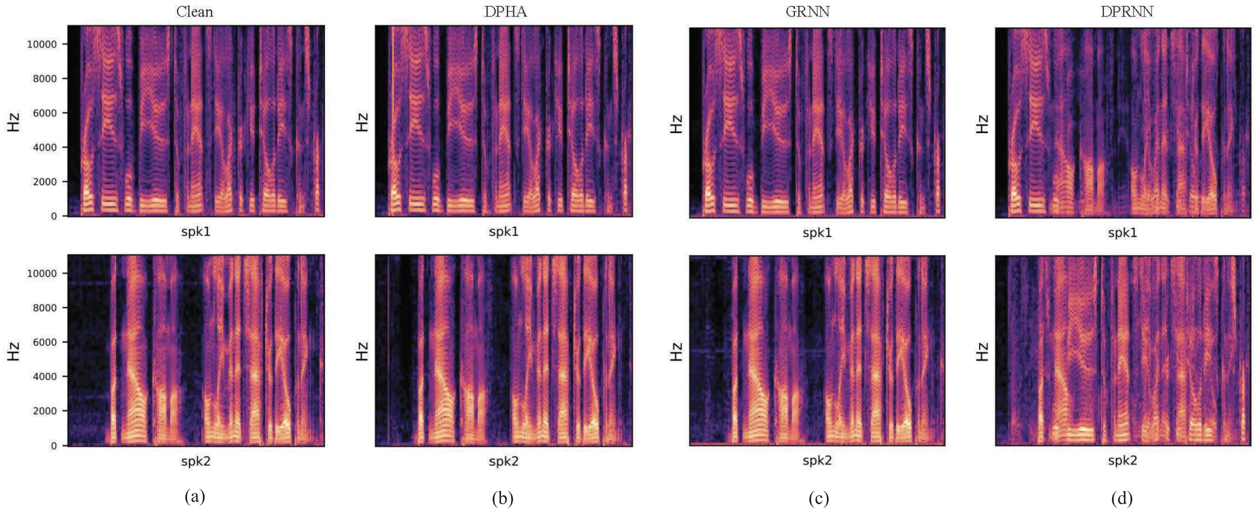

FIGURE 5. (Color Online). Two-speaker separation results of different models in magnitude STFT on WSJ0-2mix test dataset.Three models are DPHA-Net, Gated DPRNN and DPRNN. (a) Two-speaker clean speech magnitude spectrum. (b) The separated result of DPHA-Net. (c) The separated result of Gated-DPRNN (GRNN). (d) The separated result of DPRNN.

TABLE 5. Comparison between our proposed method with four methods using WHAM! dataset. '–' means the corresponding result not reported.

Metrics	# param	SI-SNRi(dB)	SDRi(dB)
Conv-TasNet [15]	5.1 M	12.7	–
DPRNN [19]	2.6 M	13.7	14.1
A-FRCNN [17]	6.1 M	14.5	14.8
Gated-DPRNN [21]	7.5 M	**15.17**	–
DPHA-Net	6.1 M	14.71	15.07

TABLE 6. Ablation results of two-speaker separation on WSJ0-2mix. The best scores are in bold.

Metrics	# param	SI-SNRi(dB)	SDRi(dB)	STOI(%)	PESQ
baseline	2.6 M	18.6	18.2	97.45	3.58
DPHA-Net	6.1 M	**20.7**	**20.9**	98.14	**3.70**
–AFF (i)	5.8 M	20.5	20.7	**98.19**	3.70
–FSAM(ii)	6.0 M	20.2	20.4	98.09	3.67
–MAT (iii)	5.9 M	17.75	17.99	97.07	3.42
–EA (iv)	3.6 M	19.97	20.17	97.71	3.63
–MHSA (v)	5.8 M	19.7	19.9	97.52	3.64
–MSTS (vi)	6.0 M	18.0	18.2	97.26	3.48

the WSJ0-2mix dataset. It shows that the Libri2Mix dataset is more challenging for the speech separation task [40]. Comparing with the A-FRCNN, the experimental results of DPHA-Net is better than that of the A-FRCNN on the WSJ0-2mix but slightly less than that of the A-FRCNN on the Libri2Mix dataset.

Next, we verified the performance of the DPHA-Net under the noisy condition on the WHAM! dataset. Table 5 lists the results of the DPHA-Net and 4 compared methods on the WHAM! dataset. As we can seen, the proposed DPHA-Net is superior to compared methods but slightly less than Gated-DPRNN. Meanwhile, on the WSJ0-2mix and Libri2Mix datasets, the separation performance of DPHA-Net is both excellent.

B. AN ABLATION STUDY FOR DPHA-NET

In order to understand the contribution of each of the various components in the DPHA-Net, we perform an ablation study in order to explore how to choice appropriate component to construct the DPHA-Net models for the speaker separation task. We choose the DPRNN[2] as the baseline method. The ablation experiments were performed on the WSJ0-2mix dataset. The ablation results are listed in Table 6 and drawn in Figure 6 in the form of bar chart. Several variants of DPHA-Net are compared in Table 6: (i) the DPHA-Net without the adaptive feature fusion unit ("–AFF"); (ii) the DPHA-Net model trained exploiting a weighted loss and densely connected operation but without feature selective aggregation mechanism("–FSAM"); (iii) the DPHA-Net trained exploitng a SI-SNR loss that is applied only at the final output of the model while without feature selective aggregation operations ("–MAT"); (iv) the DPHA-Net without element-wise attention ("–EA"); (v) the DPHA-Net without mutli-head self-attention unit ("–MHSA"); (vi) the DPHA-Net trained exploitng a SI-SNR loss that is applied only at the final output of the model ("–MSTS"). Comparing with the complete DPHA-Net, without MAT, the score of SI-SNRi decreases by 2.95 dB, that of SDRi by 2.91 dB and PESQ by 0.28.While without MSTS, the score of SI-SNRi decreases by 2.7 dB, that of SDRi by 2.7 dB and PESQ by 0.22. While without FSAM, the score of SI-SNRi decreases by 0.5 dB, that of SDRi by 0.5 dB. While without MSHA unit, the score of SI-SNRi decreases by 1.0 dB, that of SDRi by 1.0 dB and PESQ by 0.06.

As can be seen, each of the aforementioned components contributes to the performance gain of the DPHA-Net, while the multi-stage aggregation training strategy much more than

[2] https://github.com/ShiZiqiang/dual-path-RNNs-DPRNNs-based-speech-separation

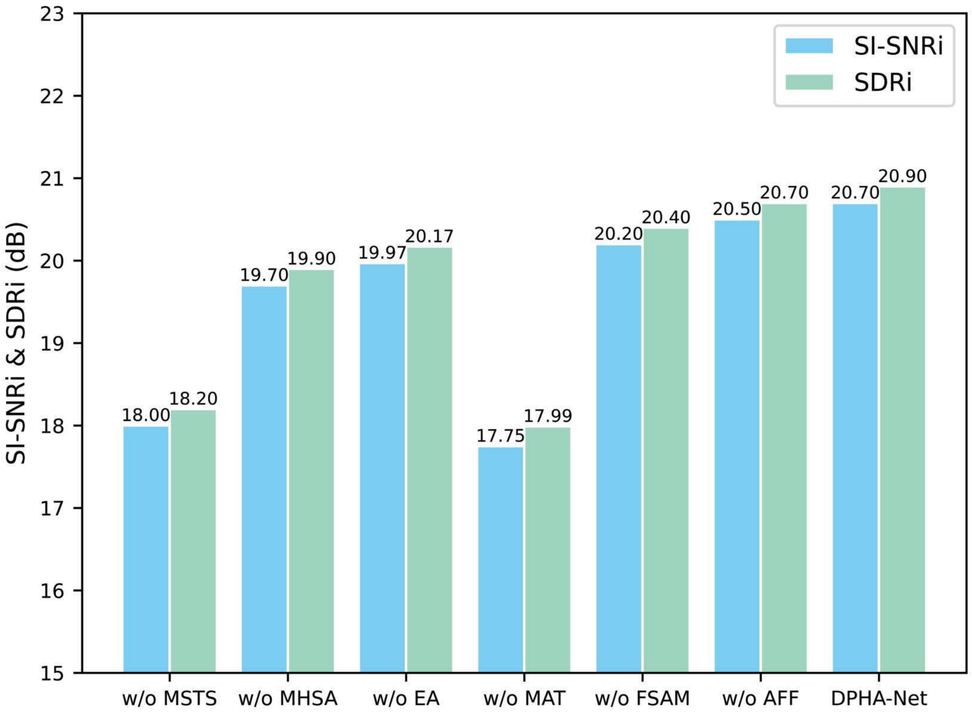

FIGURE 6. (Color Online). Bar charts of ablation results of two-speaker separation. "w/o" is the abbreviation of without, such as "w/o MSTS" denotes without the MSTS.

TABLE 7. Average SI-SNRi, SDRi, PESQ and STOI for DPHA-Net to different gender combinations on WSJ0-2mix. The best scores are in bold.

Gender	SI-SNRi(dB)	SDRi(dB)	STOI(%)	PESQ
Female-Female	18.83	19.07	96.80	3.57
Female-Male	**21.28**	**21.46**	**98.60**	**3.75**
Male-Male	20.50	20.69	98.08	3.69
Average	20.7	20.9	98.14	3.70

the others. By multi-stage aggregation training, the model is forced to approximate the more suitable separation network and also estimate the signal in a more accurate scale closer to the time-domain resolution leading to better results. This is also consistent with other studies [21], [22], [31]. As shown in Figure 6, the results of ablation experiments obviously verify that the MAT is crucial for the DPHA-Net.

The separation performance with respect to different gender combination and overall performance across all combinations are reported in Table 7. From this table, we can clearly see that our approach achieves much better SI-SNRi, SDRi, PESQ and STOI on female-male combinations than same gender conditions. The result also proved consistent with the observation in [6].

C. SUBJECTIVE AND OBJECTIVE QUALITY EVALUATION OF DPHA-NET

In addition to SDRi and SI-SNRi, we also compared the DPHA-Net with DPRNN on the subjectiveand objective quality. Since PESQ aims to predict the subjective quality of speech, human quality evaluation can be consideredas the ground truth [15]. Therefore, we conducted an experiment in which we asked 20 normal hearing subjects to listen and rate the quality of the separated speech sounds. We randomly chose 25 two-speaker mixture utterances from the test set of WSJ0-2mix dataset. The subjects were asked to rate the quality of the clean utterances, the separated utterances by DPHA-Net and the DPRNN, respectively on the scale of 0 to 5 (the scale is the same as PESQ). For example, a very clean utterance was first given as the reference for the highest possible score (i.e. 5). Then, according to the separated speech they heard, the subjects randomly gave scores between 0 and 5.

TABLE 8. Mean opinion score (MOS,N=20) and PESQ for the 25 selected utterances from the WSJ0-2mix test set.

Method	MOS	PESQ
DPHA-Net	4.06	3.79
DPRNN	3.99	3.55
Clean	4.05	4.5

Table 8 shows the results of human subjective quality test, where the MOS score of DPHA is slightly better than that of the DPRNN. Because the PESQ score is calculated corresponding to the clean unterance, so the PESQ score of the last row is maximum score, 4.5 while the MOS score is lower than that of DPHA-Net.

D. CROSS VALIDATION ON DIFFERENT DATASET FOR DPHA-NET

We also performed the experiments on the WSJ0-2mix and Libri2Mix datasets to validate the robustness of proposed DPHA-Net while the training and testing datasets

TABLE 9. Comparison of different train/dev set on different test set in the two-speaker condition on WSJ0-2mix and Libri2Mix. Where the number in the table is arranged as: SI-SNRi (dB) / SDRi (dB) / STOI (%) / PESQ).

Dataset	WSJ0-2mix train/dev	Libri2Mix train-100/dev
WSJ0-2mix test	20.9/20.7/98/3.7	16.9/17.2/96/3.4
Libri2Mix test	14.1/13.6/93/3.1	16.5/17.0/94/3.3

are inconsistent. In this experiment, we chose the complete DPHA-Net as the main model. Table 9 lists four group scores with different training and testing setup. Interestingly, we found two obviously results: (1) When the training and testing datasets are consistent, the results of experiments on the Libri2mix dataset both lower than that on the WSJ0-2mix dataset. (2) When the model is trained on WSJ0-2mix training and validation dataset, the performance difference between testing on the WSJ0-2mix and Libri2Mix test datasets is more larger than comparing with when trained on the Libri2Mix training and dev dataset. These results show that maybe the WSJ0 dataset is more comfortable for DPHA-Net. Overall, our proposed DPHA-Net model has certain robustness, matching our expectations. Furthermore, the result on Libri2Mix also proves the conclusion in [40] that the Libri2Mix dataset is more complicated than the WSJ0-2mix for the task of two-speaker separation.

V. CONCLUSION

This paper introduces our proposed DPHA-Net for monaural speech separation, which combines the advantages of attention mechanisms and multi-stage training strategy. For the task of two-speaker separation in the without noisy condition, the DPHA-Net is superior to 10 compared methods on the WSJ0-2mix dataset, and to 6 compared methods on the Libri2Mix dataset, respectively. The results of ablation experiments show the effectiveness of AFF, FSAM, EA, MHSA, MSTS and MAT. The experimental results show that the multi-stage aggregation training strategy with Feature Selective Aggregation Mechanism (FSAM) play an important role in DPHA-Net.

In this paper, we performed the experiments on the task of two-speaker separation, it did not consider the more natural setting such as the number of speakers is unknown and reverberated conditions. We will focus on speech separation in reverberated condition and the unknown number of speakers in the future.

REFERENCES

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *J. Acoust. Soc. Amer.*, vol. 25, no. 5, pp. 975–979, Sep. 1953.

[2] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 10, pp. 1702–1726, Oct. 2018.

[3] B. Bhattarai, Y. R. Pandeya, and J. Lee, "Parallel stacked hourglass network for music source separation," *IEEE Access*, vol. 8, pp. 206016–206027, 2020.

[4] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Hoboken, NJ, USA: Wiley, 2006.

[5] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA, USA: MIT Press, 1994.

[6] Y. Liu and D. Wang, "Divide and conquer: A deep CASA approach to talker-independent monaural speaker separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 12, pp. 2092–2102, Dec. 2019.

[7] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 31–35.

[8] F. Jiang and Z. Duan, "Speaker attractor network: Generalizing speech separation to unseen numbers of sources," *IEEE Signal Process. Lett.*, vol. 27, pp. 1859–1863, 2020.

[9] Y. Isik, J. Le Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," in *Proc. Interspeech.*, 2016, pp. 545–549.

[10] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 246–250.

[11] Z.-Q. Wang, J. L. Roux, and J. R. Hershey, "Alternative objective functions for deep clustering," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 686–690.

[12] Z.-Q. Wang, J. Le Roux, D. Wang, and J. Hershey, "End-to-end speech separation with unfolded iterative phase reconstruction," in *Proc. Interspeech*, Sep. 2018, pp. 2708–2711.

[13] Z.-Q. Wang, K. Tan, and D. Wang, "Deep learning based phase reconstruction for speaker separation: A trigonometric perspective," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 71–75.

[14] Y. Masuyama, K. Yatabe, K. Nagatomo, and Y. Oikawa, "Joint amplitude and phase refinement for monaural source separation," *IEEE Signal Process. Lett.*, vol. 27, pp. 1939–1943, 2020.

[15] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.

[16] E. Tzinis, Z. Wang, and P. Smaragdis, "Sudo RM-RF: Efficient networks for universal audio source separation," in *Proc. IEEE 30th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Sep. 2020, pp. 1–6.

[17] X. Hu, K. Li, W. Zhang, Y. Luo, J.-M. Lemercier, and T. Gerkmann, "Speech separation using an asynchronous fully recurrent convolutional neural network," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 3056–3060.

[18] E. Tzinis, S. Venkataramani, Z. Wang, C. Subakan, and P. Smaragdis, "Two-step sound source separation: Training on learned latent targets," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 31–35.

[19] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 46–50.

[20] J. Chen, Q. Mao, and D. Liu, "Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation," in *Proc. Interspeech*, Oct. 2020, pp. 2642–2646.

[21] E. Nachmani, Y. Adi, and L. Wolf, "Voice separation with an unknown number of multiple speakers," in *Proc. Int. Conf. Mach. Learn.*, vol. 119, 2020, pp. 7164–7175.

[22] K. Tan, B. Xu, A. Kumar, E. Nachmani, and Y. Adi, "SAGRNN: Self-attentive gated RNN for binaural speaker separation with interaural cue preservation," *IEEE Signal Process. Lett.*, vol. 28, pp. 26–30, 2020.

[23] M. W. Y. Lam, J. Wang, D. Su, and D. Yu, "Effective low-cost time-domain audio separation using globally attentive locally recurrent networks," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Jan. 2021, pp. 801–808.

[24] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 21–25.

[25] Y. Jin, C. Tang, Q. Liu, and Y. Wang, "Multi-head self-attention-based deep clustering for single-channel speech separation," *IEEE Access*, vol. 8, pp. 100013–100021, 2020.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[27] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3146–3154.

[28] C. Tang, C. Luo, Z. Zhao, W. Xie, and W. Zeng, "Joint time-frequency and time domain learning for speech enhancement," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Jul. 2020, pp. 3816–3822.

[29] H. Wang, Y. Zou, D. Chong, and W. Wang, "Environmental sound classification with parallel temporal-spectral attention," in *Proc. Interspeech*, Oct. 2020, pp. 821–825.

[30] J. Kim and M. Hahn, "Speech enhancement using a two-stage network for an efficient boosting strategy," *IEEE Signal Process. Lett.*, vol. 26, no. 5, pp. 770–774, May 2019.

[31] S.-J. Li, Y. AbuFarha, Y. Liu, M.-M. Cheng, and J. Gall, "MS-TCN++: Multi-stage temporal convolutional network for action segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Sep. 4, 2020, doi: 10.1109/TPAMI.2020.3021756.

[32] C. Gao, Y. Gu, and I. Marsic, "Time-domain mapping based single-channel speech separation with hierarchical constraint training," 2021, *arXiv:2110.10593*.

[33] J. Lei Ba, J. Ryan Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.

[34] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7132–7141.

[35] L. Yang, H. Jiang, R. Cai, Y. Wang, S. Song, G. Huang, and Q. Tian, "CondenseNet v2: Sparse feature reactivation for deep networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3569–3578.

[36] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2403–2412.

[37] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 1–9.

[38] S. Chen, Y. Wu, Z. Chen, T. Yoshioka, S. Liu, J. Li, and X. Yu, "Don't shoot butterfly with rifles: Multi-channel continuous speech separation with early exit transformer," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 6139–6143.

[39] M. Kolbaek, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 10, pp. 1901–1913, Oct. 2017.

[40] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "LibriMix: An open-source dataset for generalizable speech separation," 2020, *arXiv:2005.11262*.

[41] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. L. Roux, "WHAM! Extending speech separation to noisy environments," in *Proc. Interspeech*, Sep. 2019, pp. 1368–1372.

[42] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 5206–5210.

[43] ITU-R, *Algorithms to Measure Audio Programme Loudness and True-Peak Audio Level*, document Rec. ITU-R BS.1770-4, 2015.

[44] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[45] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR—Half-baked or well done?" in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2019, pp. 626–630.

[46] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.

[47] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, May 2001, pp. 749–752.

[48] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE/ACM Trans. Audio, Speech, Language Process*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

[49] T. Lan, Y. Qian, Y. Lyu, R. Mokhosi, W. Tai, and Q. Liu, "Improved speech separation with time-and-frequency cross-domain feature selection," in *Proc. Interspeech*, Aug. 2021, pp. 3525–3529.

WENBO QIU is currently pursuing the M.E. degree with the School of Information Science and Engineering, Xinjiang University, China, under the supervision of Ying Hu. His current research interest includes source separation, i.e., speaker separation, music separation, and target speaker extraction.

YING HU received the B.S. and M.S. degrees from Xinjiang University, China, in 1997 and 2002, respectively, and the Ph.D. degree in information and communication engineering from Xian Jiao Tong University, China, in 2016. She is currently an Associate Professor with the School of Information Science and Engineering, Xinjiang University. Her main research interests include source separation, sound localization, sound event detection, audio information retrieval, and speech emotion recognition.