MossFormer2: Combining Transformer and RNN-Free Recurrent Network for Enhanced Time-Domain Monaural Speech Separation

Shengkui Zhao, Yukun Ma, Chongjia Ni, Chong Zhang, Hao Wang
Trung Hieu Nguyen, Kun Zhou, Jia Qi Yip, Dianwen Ng, Bin Ma

Introduction

Our previously proposed monaural speech separation model

- ***MossFormer*** has achieved promising performance in monaural speech separation: SI-SDRi of 22.8dB and 21.2dB (WSJ0-2/3mix).
- However, ***MossFormer*** is inefficient for modeling finer-scale recurrent patterns presented in speech signals due to the fact that
  - it predominantly adopts a self-attention-based separation module in the masking net, and
  - the self-attention module tends to emphasize long-range, coarser-scale dependencies while being less effectively in modelling finer-scale recurrent patterns.

In this work

- We introduce a novel hybrid model, ***MossFormer2***, that provides the capabilities to model both long-range, coarse-scale dependencies and fine-scale recurrent patterns.
- ***MossFormer2*** integrates a recurrent module into the MossFormer framework, where the recurrent module is based on a feedforward sequential memory network (FSMN), which is an "***RNN-free***" recurrent network due to the ability to capture recurrent patterns without using recurrent connections.
- ***MossFormer2*** encourages parallel processing as the recurrent module relies only on linear projections and convolutions.

Our results:

- The ***MossFormer2*** hybrid model demonstrates remarkable enhancements over MossFormer and surpasses other state-of-the-art methods in ***WSJ0-2/3mix***, ***Libri2Mix***, and ***WHAM!/WHAMR!*** benchmarks.
- ***MossFormer2*** achieves SI-SDRi of **24.1dB** and **22.2dB** on ***WSJ0-2/3mix + DM***, and **21.7dB** on the ***Libri2Mix*** dataset.

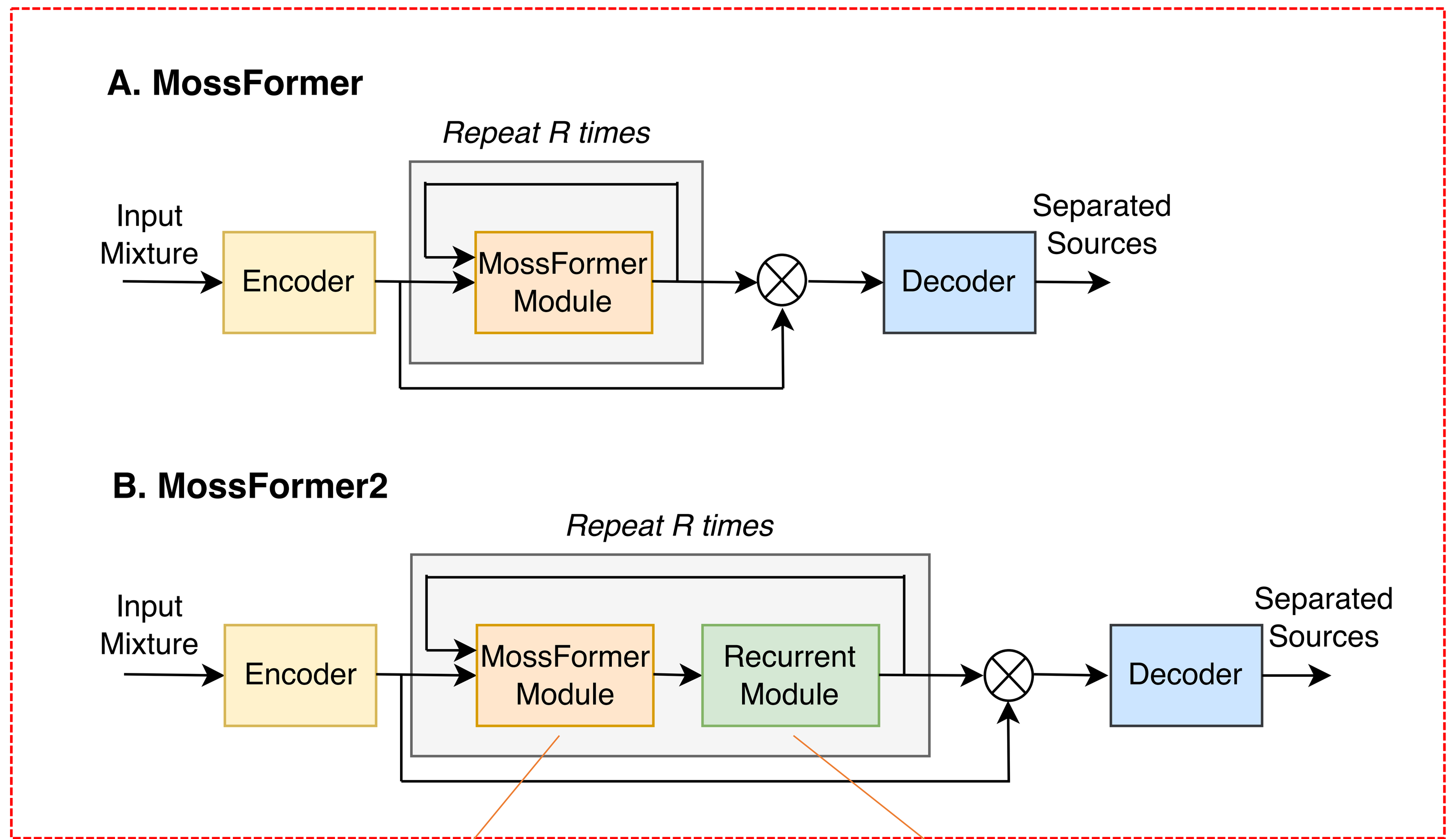
Our Approach

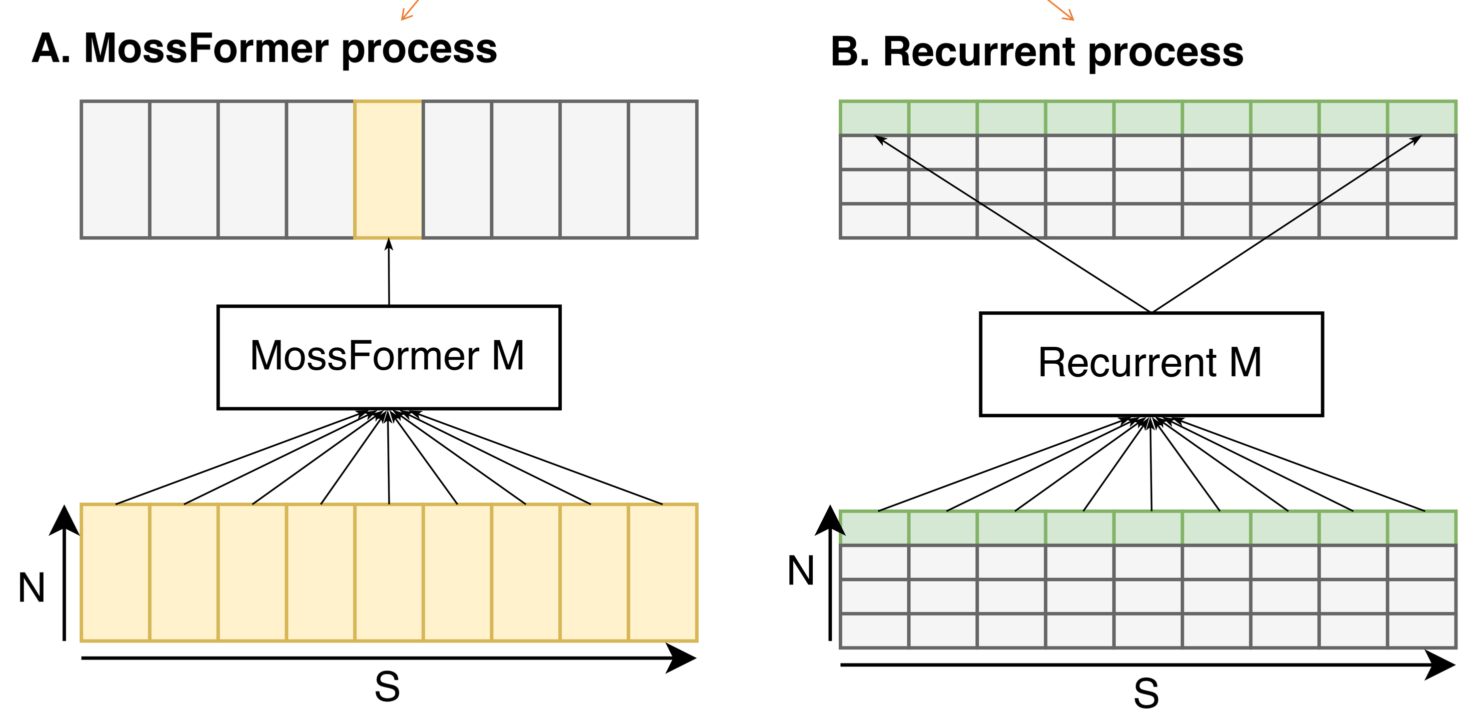
Problem Formulation

- Given a speech mixture $x = \sum_{i=1}^{C} s_i$, we aim to estimate C individual speech sources $s_i \in R^{1xT}$, $i = 1,2,\ldots,C$ based on a deep learning model.

From MossFormer to MossFormer2:

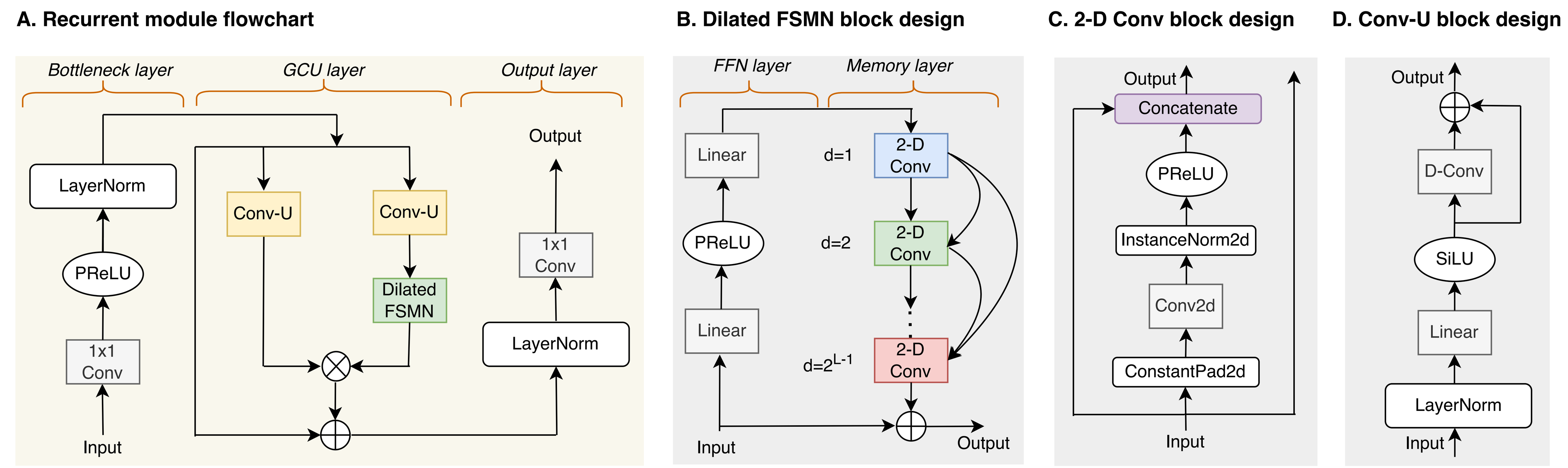
- The MossFormer module remains consistent across both ***MossFormer*** and ***MossFormer2***.
- ***MossFormer2*** forms a novel hybrid architecture by integrating a recurrent module into the MossFormer framework.
- The core concept of the MossFormer framework is applying joint local-global self-attention strategy to the entire sequence.
- Not relying on recurrence, the self-attention primarily captures long-range, coarse-scale dependencies.
- The dedicated recurrent module models intricate temporal dependencies within speech signals.
- We hypothesize that distinct embedding levels retain distinct recurrent patterns, thus the recurrent module conducts recurrent learning on each embedding dimension.
- Leveraging the combined strengths of self-attention and recurrent modelling, ***MossFormer2*** facilitates the capture of both broad dependencies and localized recurrent patterns.





RNN-Free Recurrent Module

- Unlike LSTM and GRU, our proposed recurrent module is based on FSMN without using recurrent connections.
- The proposed recurrent module is composed of a bottleneck layer, a GCU layer, and an output layer:



- The bottleneck layer is to decrease the embedding dimensionality while retaining crucial features.
- The GCU architecture is employed for sequential processing, inspiring from the gating mechanism of GLU.
- To facilitate model training, we add a skip connection to link the GCU layer's input to its output.
- The output layer restores the embedding dimensionality from the output of the GCU layer.
- We propose dilations for FSMN to achieve broader receptive fields and dense connections to enhance the information flow and facilitate the gradient propagation.

Experimental Results

- **Dataset (8 kHz)**: a) ***WSJ0-2/3mix***: clean, train: 20000 utts, dev: 5000 utts, test: 3000 utts. b) ***WHAM!*** and ***WHAMR!***: noisy and reverberant versions of WSJ0-2mix. c) **+DM**: dynamic mixing for ***WSJ0-2/3mix, WHAM!*** and ***WHAMR!***. d) ***Libri2Mix***: clean, 106 hours of training, 5.5 hours of dev and 5.5 hours of eval data.
- **Experimental setup**: We used SpeechBrain toolkit, Adam optimizer, initial learning rate (15e-5), maximum epochs (200), batch size (1), and SI-SDR training loss.
- **Experimental results**:

Table 1. Comparison for MossFormer and MossFormer2 on the WSJ0-2mix dataset. RTF denotes the real-time factor on test set.

Model	Para.(M)	R	N	K	N'	L	SI-SDRi	RTF
MossFormer (S)	25.3	25	384	16	-	-	22.5	0.025
MossFormer	42.1	24	512	16	-	-	22.8	0.038
MossFormer2 (S)	37.8	25	384	16	256	2	23.2	0.036
MossFormer2	55.7	24	512	16	256	2	**24.1**	0.053

Table 2. Performance comparison of MossFormer2 with the other state-of-the-art speech separation models on the WSJ0-2/3mix and Libri2Mix benchmark datasets.

Model	Para.(M)	SI-SDRi		
		WSJ0-2mix	WSJ0-3mix	Libri2Mix
Conv-TasNet [5]	5.1	15.3	12.7	14.7
DPRNN [6]	2.6	18.8	14.7	-
VSUNOS [7]	7.5	20.1	16.9	-
DPTNet [8]	2.6	20.2	-	-
Wavesplit [9]	29	22.2	17.8	19.5
SepFormer [10]	25.7	22.3	19.5	19.2
QDPN [14]	200.0	23.6	-	-
Separate And Diffuse [16]	-	23.9	20.9	-
SFSRNet [15]	59.0	24.0	-	20.4
MossFormer	42.1	22.8	21.2	19.7
MossFormer2	55.7	**24.1**	**22.2**	**21.7**

Table 3. Performance comparison of MossFormer2 with the other state-of-the-art speech separation models on the WHAM! and WHAMR! benchmark datasets.

Model	Para.(M)	SI-SDRi	
		WHAM!	WHAMR!
Conv-TasNet [5]	5.1	12.7	8.3
DPRNN [6]	2.6	13.9	10.3
VSUNOS [7]	7.5	15.2	12.2
Wavesplit [9]	29	16.0	13.2
SepFormer [10]	25.7	16.4	14.0
QDPN [14]	200.0	-	14.4
MossFormer	42.1	17.3	16.3
MossFormer2	55.7	**18.1**	**17.0**

Table 4. Ablation studies for MossFormer2 on the dilated FSMN, the GCU layer, and the bottleneck and output layers.

Model	SI-SDRi
MossFormer2	**24.1**
Without dilation in FSMN	23.9
Without dense connections in Dilated FSMN	24.0
Replace Conv_U with Linear in the GCU layer	23.8
Remove convolutional units (Conv_U) from the GCU layer	23.5
Remove bottleneck and output layers from the recurrent module	23.9

- **Discussion**: ***MossFormer2*** shows superior performance over ***MossFormer*** and the other state-of-the-art models, such as ***Separate And Diffuse***, ***QDPN***, and ***SFSRNet*** on diverse benchmarks.
- Our ablation studies highlights the impact of each proposed technique and demonstrates that adding the RNN-free recurrent module further contributes to separation performance.