Separate and Reconstruct: Asymmetric Encoder-Decoder for Speech Separation

Ui-Hyeop Shin*, SangYoun Lee, Taehan Kim, Hyung-Min Park

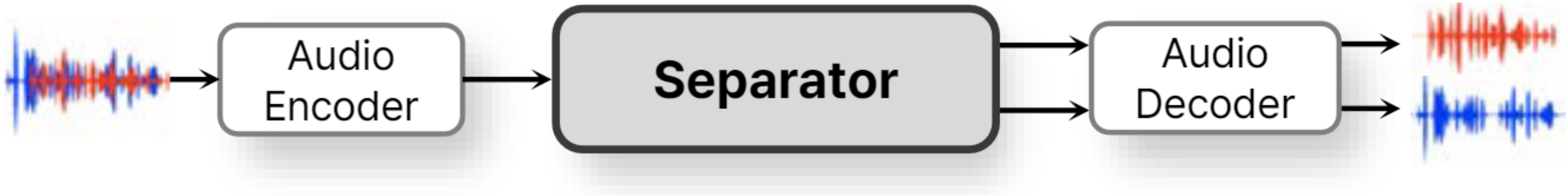
{dmlguq123,leesy0882,taehank,hpark}@sogang.ac.kr

*Presenter

Intelligent Information Processing Lab. (IIPLab), Sogang University, Seoul, Republic of Korea

Motivation: Separation-Reconstruction(SepRe)

- Time-domain audio separation network (TasNet)
  - in the latent space with convolutional encoder instaed of STFT

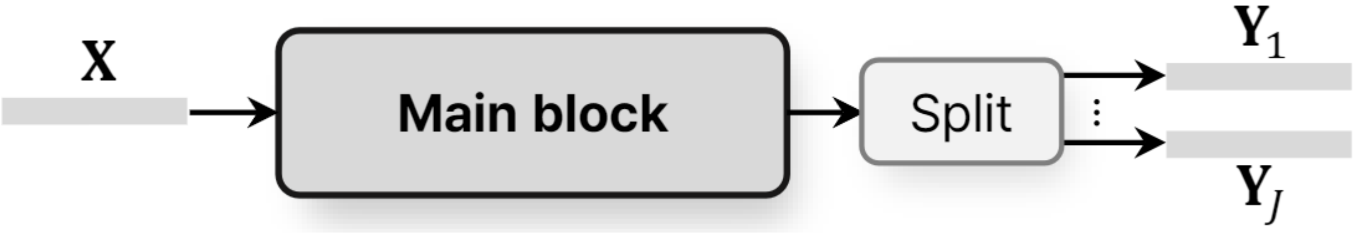


  - shortening the kernel length in the encoder → Effective!
  - requires modeling of long sequences

$T \approx 10^2 - 10^3$ → $T \approx 10^4 - 10^5$

Motivation: Separation-Reconstruction(SepRe)

- **Late split structure of TasNet**
  - TasNet: expand the feature size in terms of both length and channel
  - why?



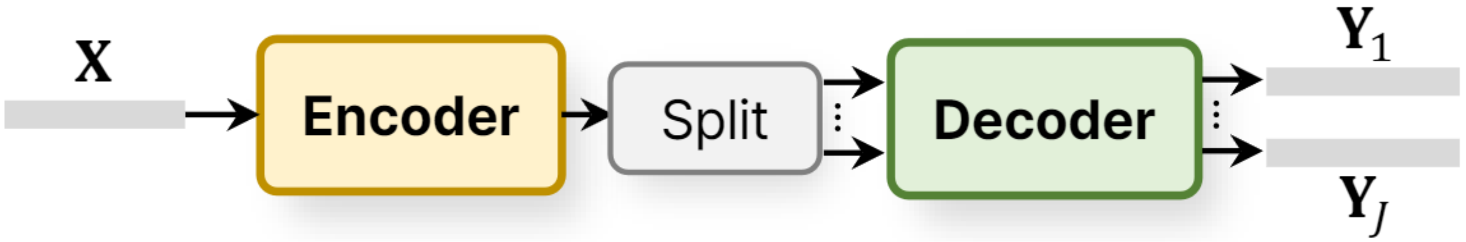
→ late split structure of TasNet!

- requires to encode all speaker information in a single feature sequence
- an information bottleneck.
- the separator must generate all separated features at once

→ increases the risk of local minima

Motivation: Separation-Reconstruction(SepRe)

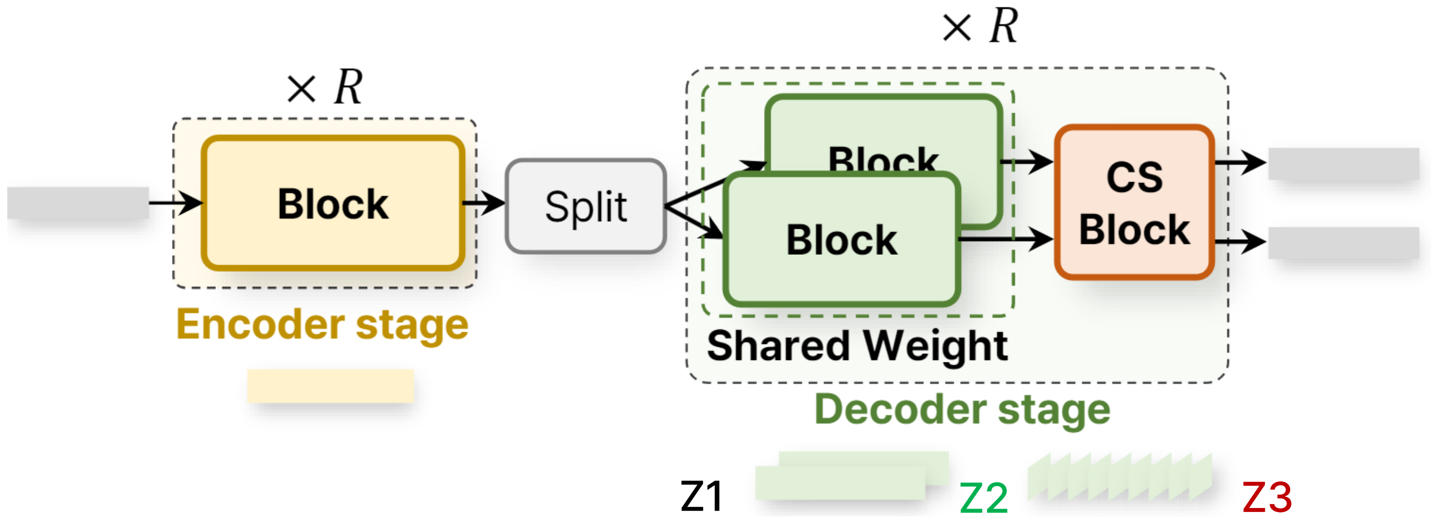
- Proposed early split structure with shared decoder (ESSD)
  - adds a speaker dimension to the feature sequence in advance to distinguish between speakers.

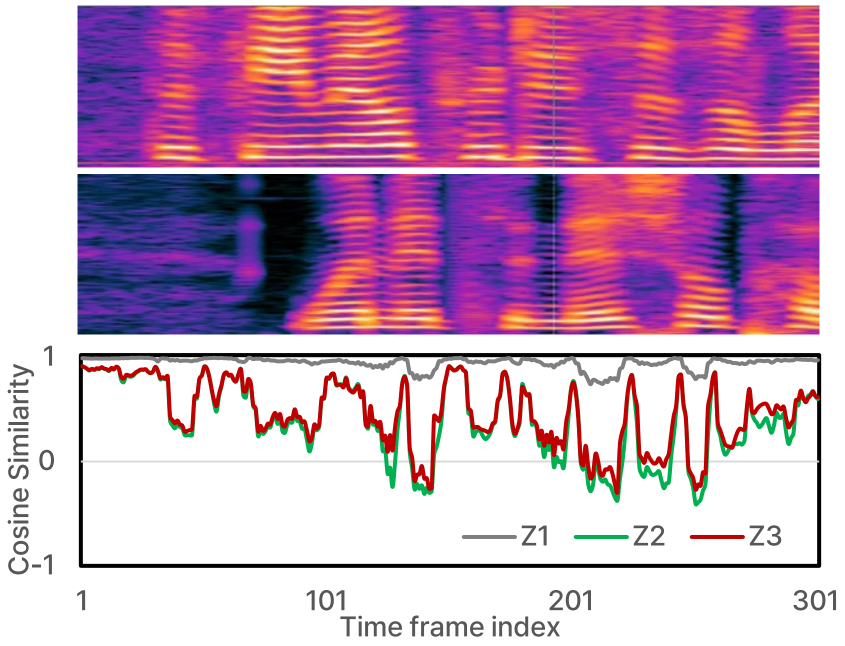


  - asymmetric strategy of separation encoder and shared decoder
  - after splitting, uses weight-sharing decoder to capture discriminative features
  - reduces the burden on the separator's encoder

Motivation: Separation-Reconstruction(SepRe)

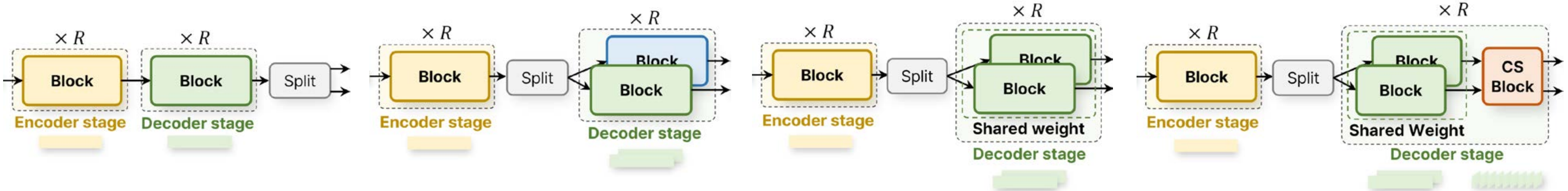
- Separation-Reconstruction(SepRe) Method





  - weight-sharing block
    - capture discriminative features
  - cross-speaker block
    - attend to each other for mistakenly clustered elements
  - discriminating and attending between sequences → reconstruction decoder

Motivation: Separation-Reconstruction(SepRe)



(a) a late split

(b) an early split + multiple decoder

(c) an early split + shared decoder (ESSD)

(d) an early split + shared decoder + CS (SepRe)

Case	MACs (G/s)	Param. (M)	SI-SNRi (dB)
late split+origin dec.	5.0/18.3	2.8/11.6	19.0/21.6
late split+large dec.	9.0/33.7	4.9/20.1	19.7/22.0
early split+multi dec.	7.9/29.5	4.5/18.4	19.8/22.1
early split+shared dec.	7.9/29.5	2.8/11.6	21.3/23.1
early split+shared dec.+CS	10.4/39.8	3.5/14.2	22.4/23.8

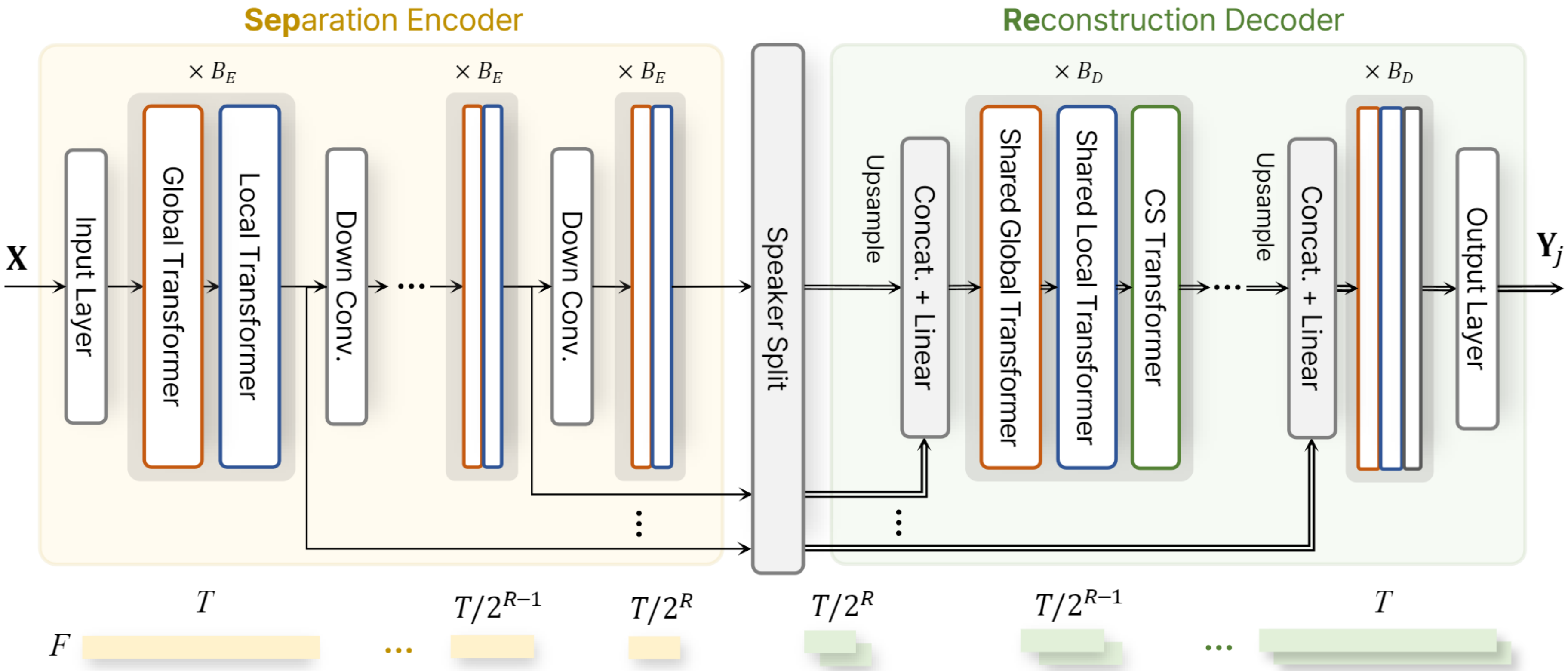
Case	ESSD	CS	ML	Param. (M)	SI-SNRi (dB)
1 (origin.)				5.1	15.3
2	✓			5.4	17.5
3	✓		✓	5.5	17.8
4 (SepRe)	✓	✓		5.7	19.2
5 (SepRe)	✓	✓	✓	5.7	19.5

(a) Conv-TasNet with SepRe method.

Case	ESSD	CS	ML	Param. (M)	SI-SNRi (dB)
1 (origin.)				26.0	20.4
2	✓			27.1	21.3
3	✓		✓	27.2	22.0
4 (SepRe)	✓	✓		28.0	21.6
5 (SepRe)	✓	✓	✓	28.0	22.7

(b) Sepformer with SepRe method.

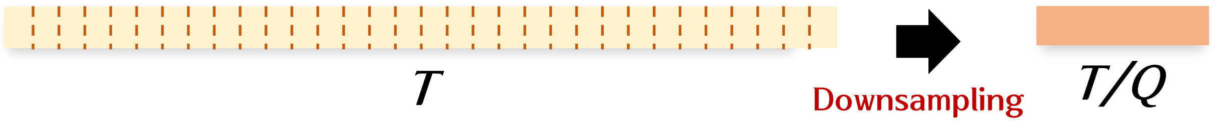
Architecture of SepReformer



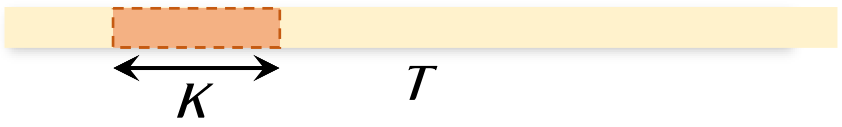
- SepRe method
- U-Net based on multi-scaled sequence
- Proposed Local and Global Processing Unit
- Speaker split: each features in a stage of the encoder are split to the number of spks

Global-Local Transformer

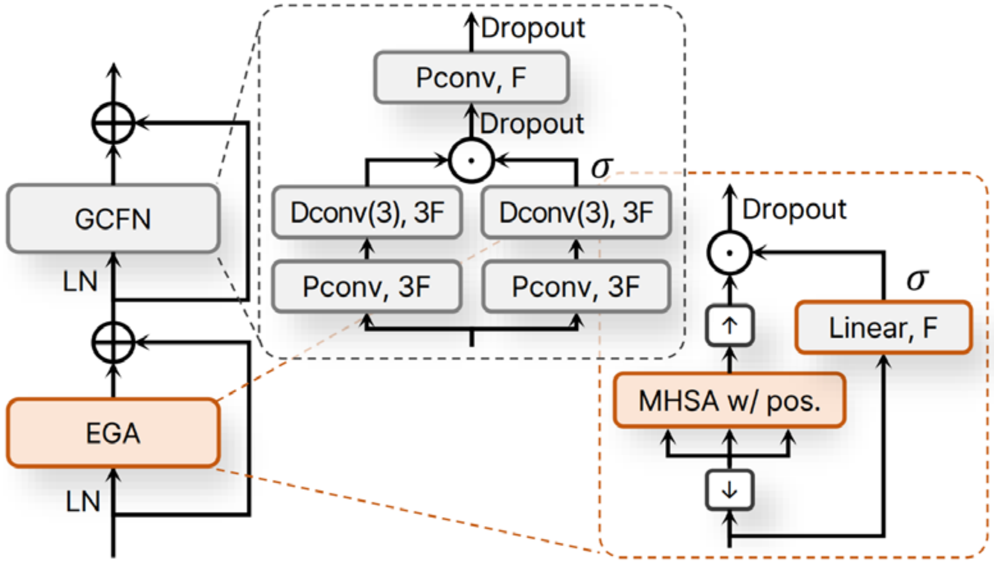
- **Global Transformer with EGA**
  - Efficient Global Attention

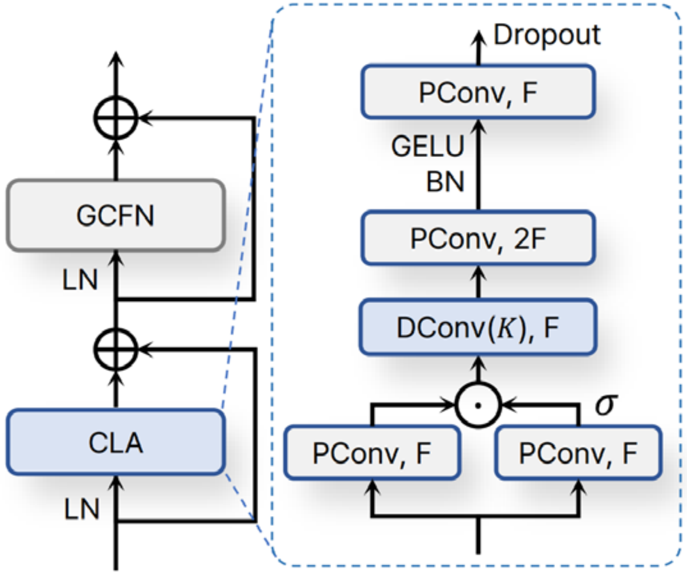


- **Local Transformer with CLA**
  - Convolutional Local Attention



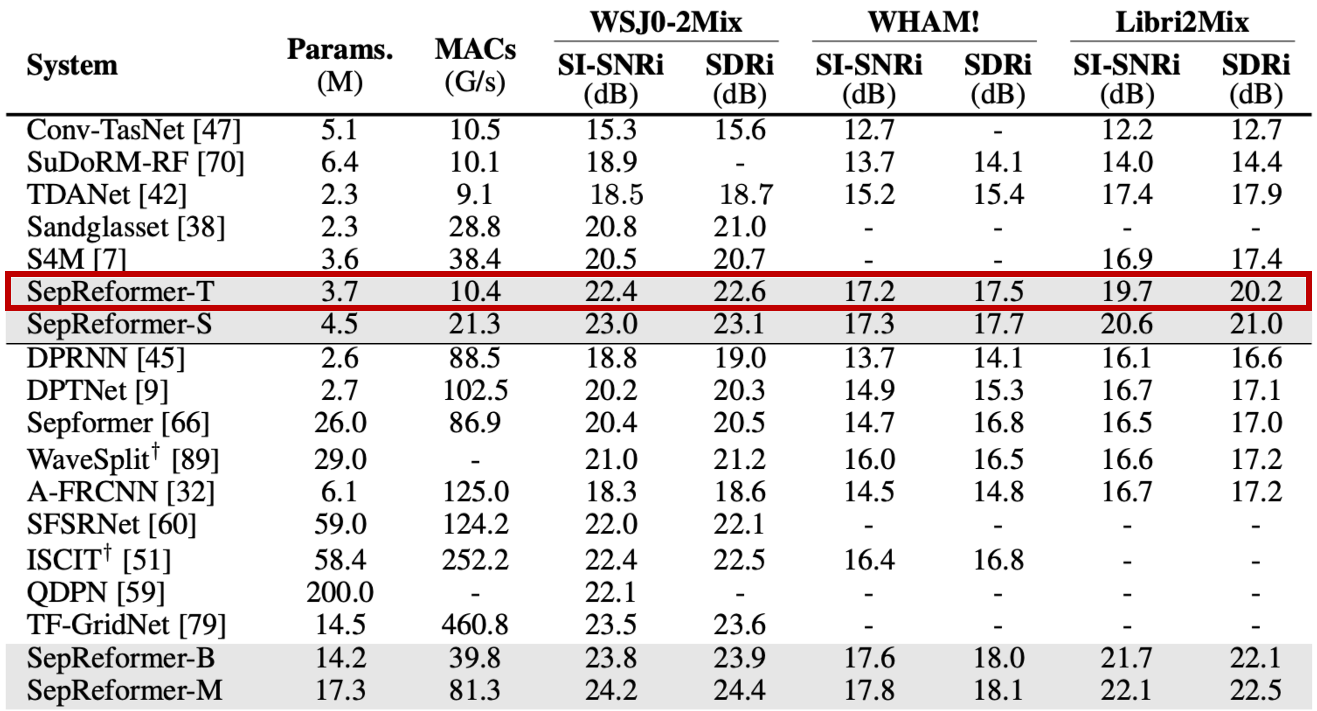
Separator	**Long sequence model**	**Param.** (M)	**MACs** (G/s)	**SI-SNRi** (dB)
Conv-TasNet	TCN	5.1	10.5	15.6
DPRNN	Dual-path + BLSTM	2.6	88.5	18.8
SuDoRM-RF	Multi-scale + Convolution	6.4	10.1	18.9
Sepformer	Dual-path + Transformer	26.0	86.9	20.4
TDANet	Multi-scale + Transformer	2.3	9.1	18.5
MossFormer(S)	GAU	10.8	44.0	20.9
S4M	Multi-scale + SSM	3.6	38.4	20.5
Ours	Global-Local Transformer	11.9	43.1	21.3
Ours + U-Net	Multi-scale + Global-Local Transformer	11.6	18.3	21.2





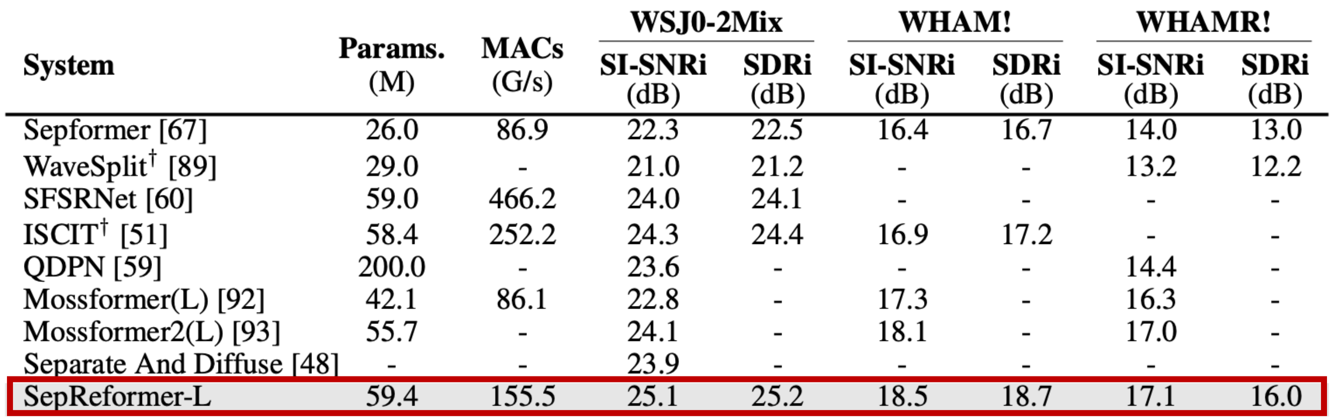
Overall Performance

- **For benchmark datasets**
  - WSJ0-2Mix / Libri2Mix
    - Clean mixture
  - WHAM!
    - Noisy mixture
  - WHAMR!
    - Noisy-reverberant mixture



System	Params. (M)	MACs (G/s)	WSJ0-2Mix		WHAM!		Libri2Mix	
			SI-SNRi (dB)	SDRi (dB)	SI-SNRi (dB)	SDRi (dB)	SI-SNRi (dB)	SDRi (dB)
Conv-TasNet [47]	5.1	10.5	15.3	15.6	12.7	-	12.2	12.7
SuDoRM-RF [70]	6.4	10.1	18.9	-	13.7	14.1	14.0	14.4
TDANet [42]	2.3	9.1	18.5	18.7	15.2	15.4	17.4	17.9
Sandglasset [38]	2.3	28.8	20.8	21.0	-	-	-	-
S4M [7]	3.6	38.4	20.5	20.7	-	-	16.9	17.4
SepReformer-T	3.7	10.4	22.4	22.6	17.2	17.5	19.7	20.2
SepReformer-S	4.5	21.3	23.0	23.1	17.3	17.7	20.6	21.0
DPRNN [45]	2.6	88.5	18.8	19.0	13.7	14.1	16.1	16.6
DPTNet [9]	2.7	102.5	20.2	20.3	14.9	15.3	16.7	17.1
Sepformer [66]	26.0	86.9	20.4	20.5	14.7	16.8	16.5	17.0
WaveSplit† [89]	29.0	-	21.0	21.2	16.0	16.5	16.6	17.2
A-FRCNN [32]	6.1	125.0	18.3	18.6	14.5	14.8	16.7	17.2
SFSRNet [60]	59.0	124.2	22.0	22.1	-	-	-	-
ISCIT† [51]	58.4	252.2	22.4	22.5	16.4	16.8	-	-
QDPN [59]	200.0	-	22.1	-	-	-	-	-
TF-GridNet [79]	14.5	460.8	23.5	23.6	-	-	-	-
SepReformer-B	14.2	39.8	23.8	23.9	17.6	18.0	21.7	22.1
SepReformer-M	17.3	81.3	24.2	24.4	17.8	18.1	22.1	22.5

(a) Comparison of SepReformer to existing models.

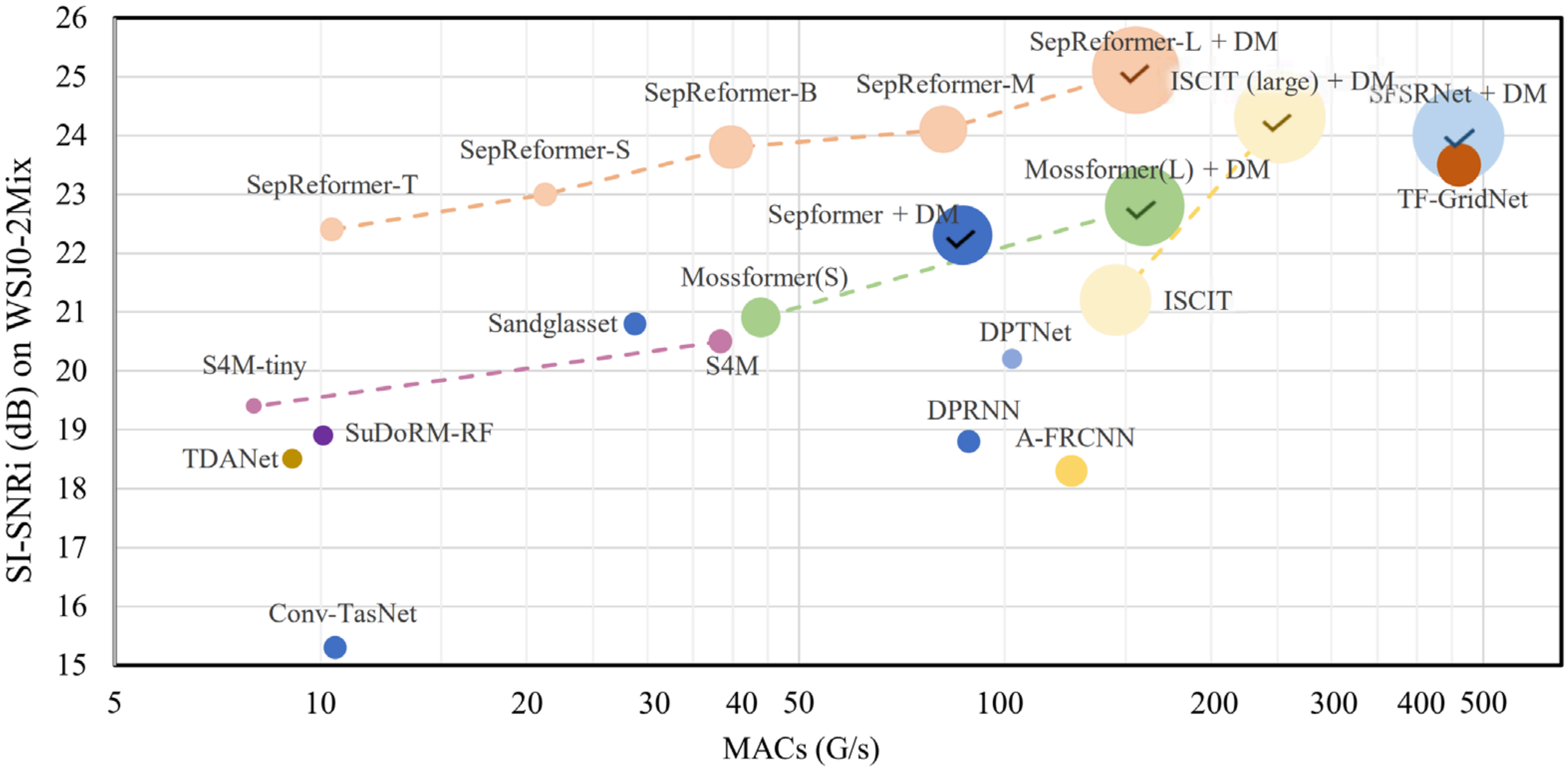


System	Params. (M)	MACs (G/s)	WSJ0-2Mix		WHAM!		WHAMR!	
			SI-SNRi (dB)	SDRi (dB)	SI-SNRi (dB)	SDRi (dB)	SI-SNRi (dB)	SDRi (dB)
Sepformer [67]	26.0	86.9	22.3	22.5	16.4	16.7	14.0	13.0
WaveSplit† [89]	29.0	-	21.0	21.2	-	-	13.2	12.2
SFSRNet [60]	59.0	466.2	24.0	24.1	-	-	-	-
ISCIT† [51]	58.4	252.2	24.3	24.4	16.9	17.2	-	-
QDPN [59]	200.0	-	23.6	-	-	-	14.4	-
Mossformer(L) [92]	42.1	86.1	22.8	-	17.3	-	16.3	-
Mossformer2(L) [93]	55.7	-	24.1	-	18.1	-	17.0	-
Separate And Diffuse [48]	-	-	23.9	-	-	-	-	-
SepReformer-L	59.4	155.5	25.1	25.2	18.5	18.7	17.1	16.0

(b) Comparison of SepReformer-L to existing large models with DM.

Overall Performance

- Performance vs. Computations



Conclusion

- SepRe method for efficient speech separation
- Global and Local Transformer Units for long sequence
- SepReformer achieved state-of-the-art performance

Thank you!