TF-SkiMNet: Speech Enhancement Based on Inplace Modeling and Skipping Memory in Time-Frequency Domain

Zixuan Li, Shulin He, Jinglin Bai, Xueliang Zhang
cslzx@mail.imu.edu.cn
College of Computer Science, Inner Mongolia University, China

1. Introduction & Motivation

Neural networks that process both full-band and sub-band information have achieved state-of-the-art (SOTA) performance in speech processing, such as TF-GridNet and TF-CrossNet.

- **Point to Success:** We believe that the superior performance of full-sub band architectures can be attributed to their **inplace modeling** capability.
- **Limitation:** They rely on a computationally massive and resource-intensive Global Multi-head Attention (GMA) mechanism to capture global temporal dependencies. Limiting its suitability for deployment on edge devices

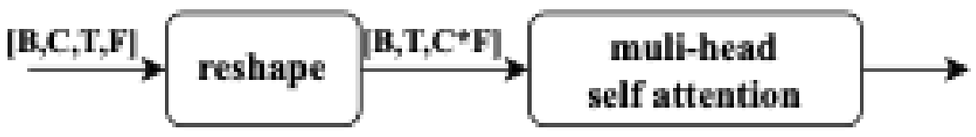

Fig.1 GMA Workflow

2. Our Contibutions

- **Efficient Global Modeling:** We extend the Skipping Memor (SkiM) to the time-frequency domain, enablingresource-efficient global temporal modeling within inplacenetworks at a low computational cost.
- **Novel Architecture:** Follow the principles of inplace modeling, we propose TF-SkiMNet, a fully causal network for speech enhancement, suitable for real-time systems.
- **SOTA Performance:** TF-SkiMNet achieves results comparable to SOTA models on single-channel tasks with 87% fewer MACs and demonstrates leading performance on multi-channel tasks under similar computational budgets.

3. TF-SkiMNet

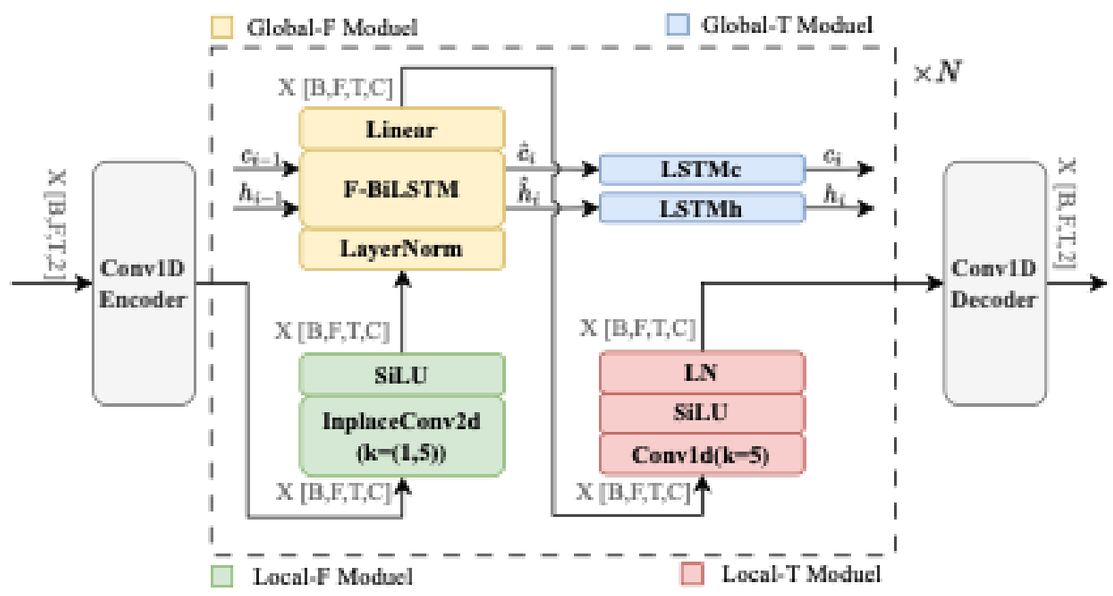

Fig.2 Architecture of TF-SkiMNet

Algorithm 2 TF-SkiMNet Forward Processing

Require: X, N $\{X$ is the input features in T-F domain; N denotes the number of layers in the network.$\}$

$X \leftarrow \text{Conv1dEncoder}(X)$
Initialize $h_0 \leftarrow \mathbf{0}$, $c_0 \leftarrow \mathbf{0}$
for $i = 1$ **to** N **do**
  $X \leftarrow \text{LayerNorm}(\text{SiLU}(\text{InplaceConv2d}^i(X)))$
  $X, \hat{h}_i, \hat{c}_i \leftarrow \text{F-BiLSTM}^i(X, h_{i-1}, c_{i-1})$
  $c_i \leftarrow \text{LayerNorm}(\text{LSTMc}_c^i(\hat{c}_i))$
  $h_i \leftarrow \text{LayerNorm}(\text{LSTMh}_h^i(\hat{h}_i))$
  $X \leftarrow \text{LayerNorm}(\text{SiLU}(\text{Conv1d}^i(\text{Linear}^i(X))))$
end for
$X \leftarrow \text{Conv1dDecoder}(X)$
return X

4. Experimental Results

Tab.1 STOI and PESQ for various methods on the single-channel speech enhancement task under different SNRs. Real-Time Factor(RTF) was measured on an Intel i7-11700 @ 2.50GHz.

Noise	Babble						Cafeteria						MACs	Params	RTF
Metrix	STOI(%)			PESQ			STOI(%)			PESQ			-	-	-
SNR	-5dB	0dB	5dB	-5dB	0dB	5dB	-5dB	0dB	5dB	-5dB	0dB	5dB	-	-	-
Mixture	58.50	70.35	81.18	1.54	1.82	2.12	57.52	69.86	81.06	1.46	1.77	2.12	-	-	-
CRN	77.62	86.96	92.03	2.01	2.45	2.81	77.62	86.93	92.11	2.01	2.42	2.79	2.58G	17.58M	0.036
GCRN	80.98	90.74	94.64	2.01	2.59	3.06	77.95	89.17	94.08	1.94	2.56	3.02	2.42G	9.77M	0.015
DCCRN	81.72	91.03	94.93	2.22	2.83	3.27	80.30	90.39	94.57	2.24	**2.86**	**3.24**	5.59G	3.67M	0.027
DPCRN	83.21	91.76	95.39	2.21	2.81	3.24	80.03	90.36	94.68	2.23	2.76	3.16	3.18G	0.81M	0.028
ICCRN	84.48	92.36	95.81	2.23	2.82	3.24	80.73	90.48	94.95	**2.26**	2.74	3.17	2.09G	0.46M	0.121
SkiM	82.52	89.97	93.93	2.14	2.57	2.92	78.20	87.89	92.90	2.00	2.44	2.82	4.74G	4.69M	0.022
TF-CrossNet	87.09	93.68	96.40	2.41	3.00	3.38	83.14	91.91	95.53	2.24	2.83	3.26	20.67G	1.47M	1.445
TF-CrossNet*	77.97	89.43	94.34	1.94	2.56	3.06	76.46	88.52	93.81	1.96	2.56	3.03	3.30G	0.13M	0.162
TF-SkiMNet	**85.62**	**92.70**	**95.81**	**2.33**	**2.89**	**3.29**	**82.16**	**91.20**	**95.13**	2.19	2.77	3.21	2.70G	0.49M	0.069
w/o Local-F	83.23	91.69	95.30	2.25	2.71	3.08	80.80	90.36	94.62	2.21	2.66	3.02	2.40G	0.48M	0.057
w/o Local-T	83.05	91.34	95.10	2.20	2.67	3.06	80.81	90.23	94.51	2.16	2.63	3.00	2.44G	0.48M	0.051
w/o Global-T	76.76	89.01	94.20	1.91	2.49	2.91	77.34	88.86	93.96	2.03	2.51	2.91	2.67G	0.16M	0.066
w/o LSTMch	77.48	89.25	94.28	1.99	2.52	2.92	76.98	88.76	93.89	2.03	2.53	2.90	2.67G	0.16M	0.066
Sub-Band	83.82	91.86	95.40	2.24	2.80	3.23	79.94	90.33	94.71	2.11	2.70	3.15	3.38G	0.20M	0.088

- TF-SkiMNet achieves comparable performance to the SOTA model TF-CrossNet while reducing MACs by 87%.
- For baselines with similar computational complexity, TF-SkiMNet significantly improves STOI and PESQ under various noise.

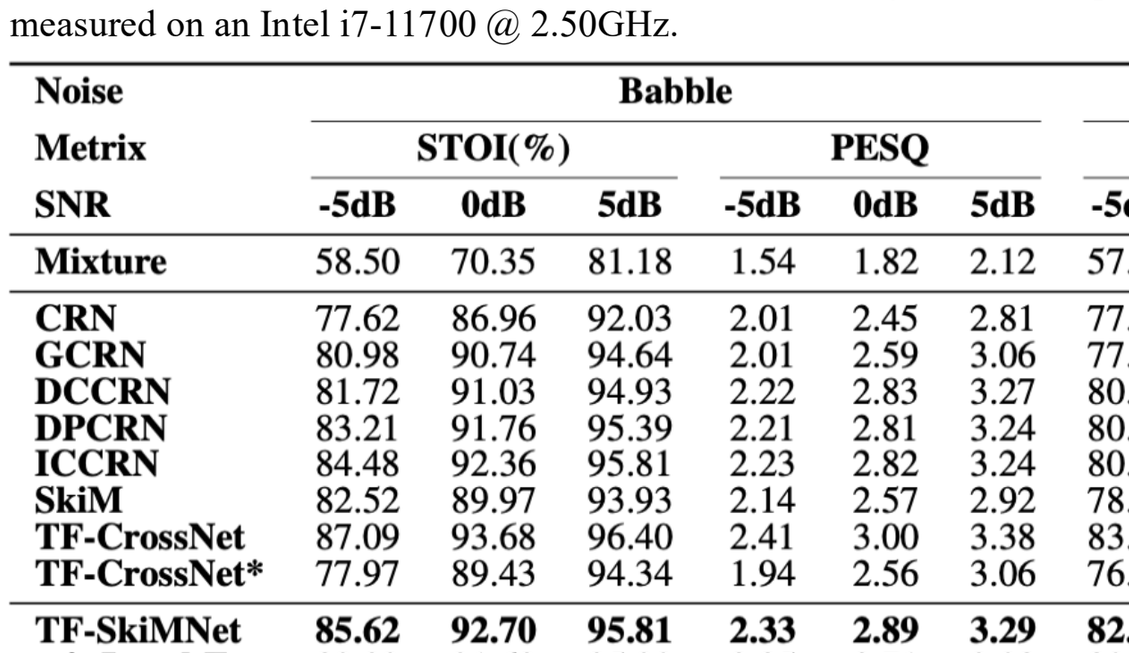

Fig3. Visualization of inference results for a -5 dB Babble noise mixture and the clustering results of the hidden states in the final layer.

- superior differentiation betweenspeech and noise
- Not introducing non-existent harmonics into the spectrum

5.Conclution

- This paper identifies the ‘inplace modeling’ capability as the key factor behind SOTA performance of full-sub band architectures, which serves as the foundation for the design of TF-SkiMNet.
- We proposed TF-SkiMNet, with a resource-efficient approach for utilizing full-band information in temporal modeling, which is is effective for both single-channel and multi-channel speech enhancemen.

Ablation studies:

- Removing local modeling along either the frequency or the time leads to noticeable performance degradation.
- w/o Global-T or w/o LSTMch Highlights the contribution of global temporal modeling.
- Sub-Band incurs higher computational costs, It still performs worse than TF-SkiMNet.

Milti-channel Speech Enhancement:

Model	Inplace	STOI	PESQ	SNR	WER	MACs	Params
Mixture	-	74.6	1.41	2.17	44.9	–	–
DPCRN	×	89.86	2.10	9.87	23.99	3.21G	0.81M
IGCRN	✓	90.35	1.92	10.05	21.55	4.27G	0.37M
ATT-MVDR	×	83.34	1.76	7.28	35.6	6.77G	37.99M
ABIC-MVDR	✓	91.47	2.32	10.57	18.89	4.06G	0.35M
OnlineSpatialNet*	✓	92.99	2.69	10.78	15.51	3.13G	0.30M
TF-SkiMNet	✓	**93.54**	**2.78**	**10.99**	**14.9**	2.72G	0.50M

- TF-SkiMNet achieves the lowest computational complexity while delivering superior performance.
- Better WER tan beamforming-based methods

6. Acknowledgements

This research was partly supported by the China National Nature Science Foundation (No. 61876214) and CCF-Lenovo Research Fund (Grant No.20240203).