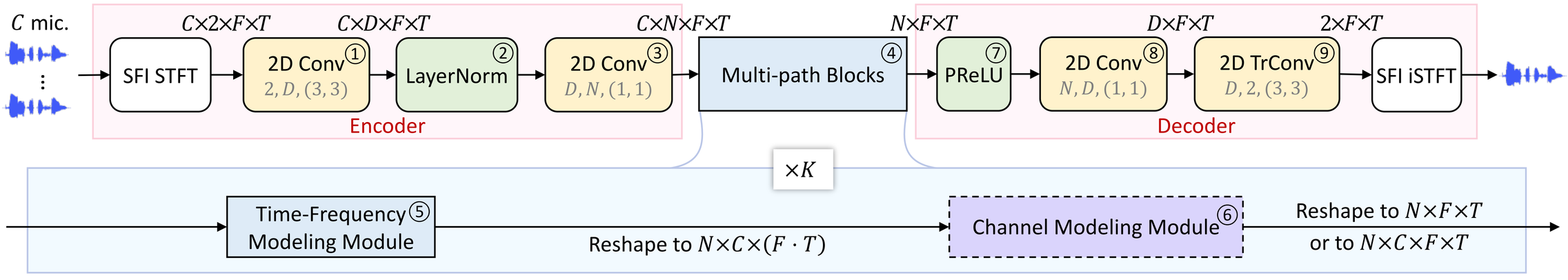
Improving Design of Input Condition Invariant Speech Enhancement

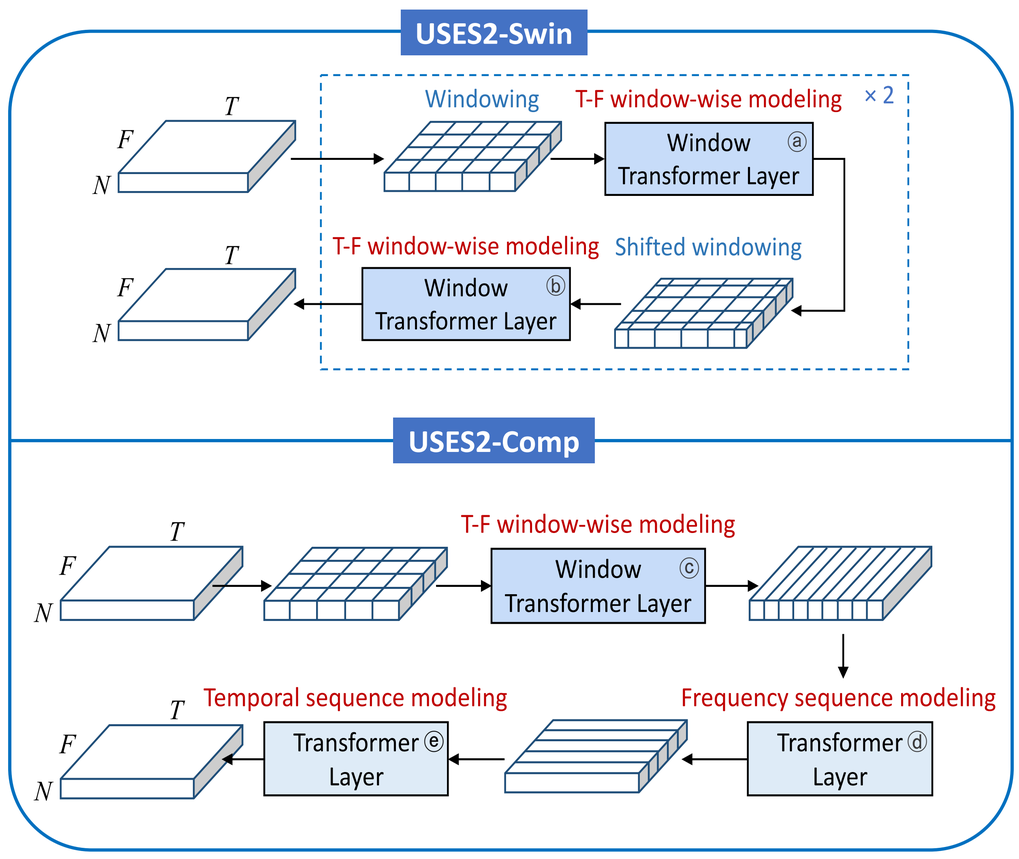
4011
SLP-P3.7

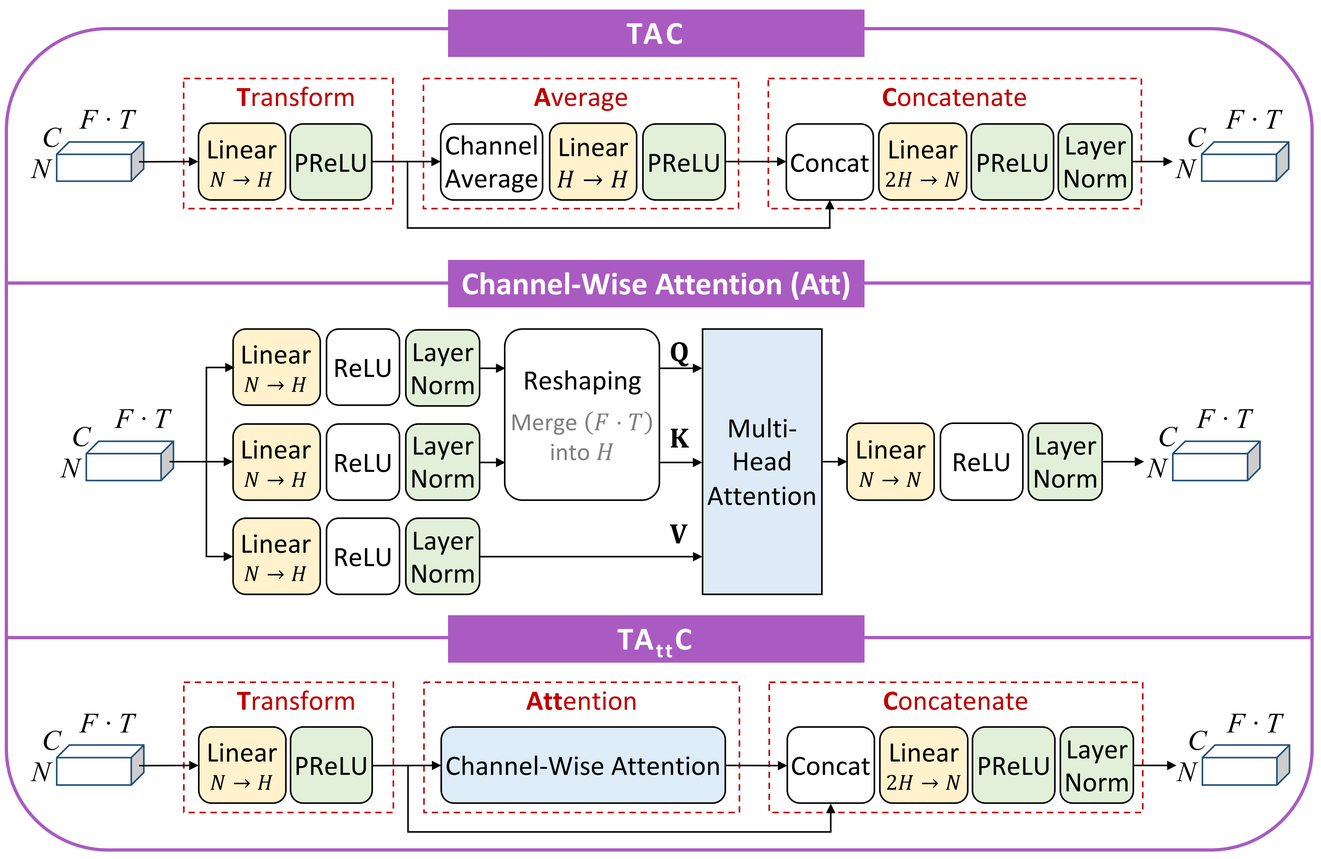
Wangyou Zhang[1,2,*], Jee-weon Jung[2,*], Yanmin Qian[1]

[1]Shanghai Jiao Tong University, China [2]Carnegie Mellon University, USA

C: #channels D: embedding dim N: bottleneck dim H: hidden dim F: frequency dim T: time dim K: #blocks







Highlights

1. Extension of our prior work [1]: input condition invariant speech enhancement (SE)
   - ❑ A single SE model unifying various input conditions including different sampling frequencies (SF), microphone variations, and input lengths
   - ❑ Decoupled 1ch and multi-channel modeling
2. Two-stage training strategy: 1ch → C-ch
3. Exploration of T-F window-wise modeling [2][3] in frequency-domain SE
4. A novel channel modeling design combining TAC [4] and channel-wise attention: TA$_{tt}$C
5. Open-source implementation in ESPnet

No.	Model	#Param	#MACs (G/s) 1ch	#MACs (G/s) 2ch	Test (Simu) SDR (dB)	Test (Simu) WER (%)	Test (Real) DNSMOS	Test (Real) WER (%)
1	No processing	-	-	-	7.5	5.8	1.46	**6.7**
2	USES (*baseline*) [1]	3.05 M	65.3	98.0	20.6	4.2	1.99 (2.94)	78.1 (11.0)
3	w/ decoupled proc.	3.05 M	60.8	98.0	18.2	5.2	2.32	27.8
4	+ 2-stage training	3.05 M	60.8	98.0	15.6	6.8	2.41	23.8
5	No.3 + Att×1 + TAC×2	3.02 M	60.8	97.5	**22.2**	**4.0**	1.48	99.0
6	+ 2-stage training	3.02 M	60.8	97.5	19.0	4.4	1.58	93.8
7	No.3 + TA$_{tt}$C × 3	3.47 M	60.8	116.9	18.7	5.2	2.40	28.8
8	+ 2-stage training	3.47 M	60.8	116.9	19.8	4.2	2.40	54.3
9	USES2-Swin	2.92 M	37.7	75.5	20.6	4.2	2.84	22.1
10	+ 2-stage training	2.92 M	37.7	75.5	21.1	4.1	2.80	24.9
11	USES2-Comp	2.53 M	52.4	83.0	20.4	4.2	2.89	15.6
12	+ 2-stage training	2.53 M	52.4	83.0	18.8	4.6	**2.96**	12.1

Experiments

1. **Investigations on CHiME-4 simulated and real data**
   - ❑ All models are only trained on 8 kHz data, and tested on 16 kHz data. (See the table for detailed results)
2. **Experiments on combined SE datasets covering diverse input conditions**
   - ❑ Combination of VCTK+DEMAND, DNS-2020, CHiME-4, REVERB, and WHAMR!
   - ❑ Siginificantly improved performance in realistic conditions
   - ❑ Check our paper for more information.

References

[1] Wangyou Zhang, *et al.*, "Toward Universal Speech Enhancement for Diverse Input Conditions," Proc. ASRU, 2023.

[2] Ze Liu, *et al.*, "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows," Proc ICCV, 2021.

[3] Jingyun Liang, *et al.*, "SwinIR: Image Restoration Using Swin Transformer," Proc. ICCV, 2021.

[4] Yi Luo, *et al.*, "End-to-End Microphone Permutation and Number Invariant Multi-Channel Speech Separation," Proc. ICASSP, 2020.