SkiM: Skipping Memory LSTM for Low-latency Real-time Continuous Speech Separation

[1]Chenda Li, [2]Lei Yang, [2]Weiqin Wang, [1]Yanmin Qian

[1] X-LANCE Lab, Shanghai Jiao Tong University
[2] Samsung Research China - Beijing (SRC-B)

SJTU Cross Media Language Intelligence Lab
上海交通大学跨媒体智能实验室

SAMSUNG

Highlights

- Real-time continuous speech separation
- 17.1 dB SDR improvement
- Computational cost is reduced by 75%
- Latency less than 1 ms on low-power device

Low-latency and Real-time Processing

- Time-domain model
  - Smaller encoder stride
  - More frames of feature
- Graph-PIT criterion
  - Meeting-level training
  - Long duration

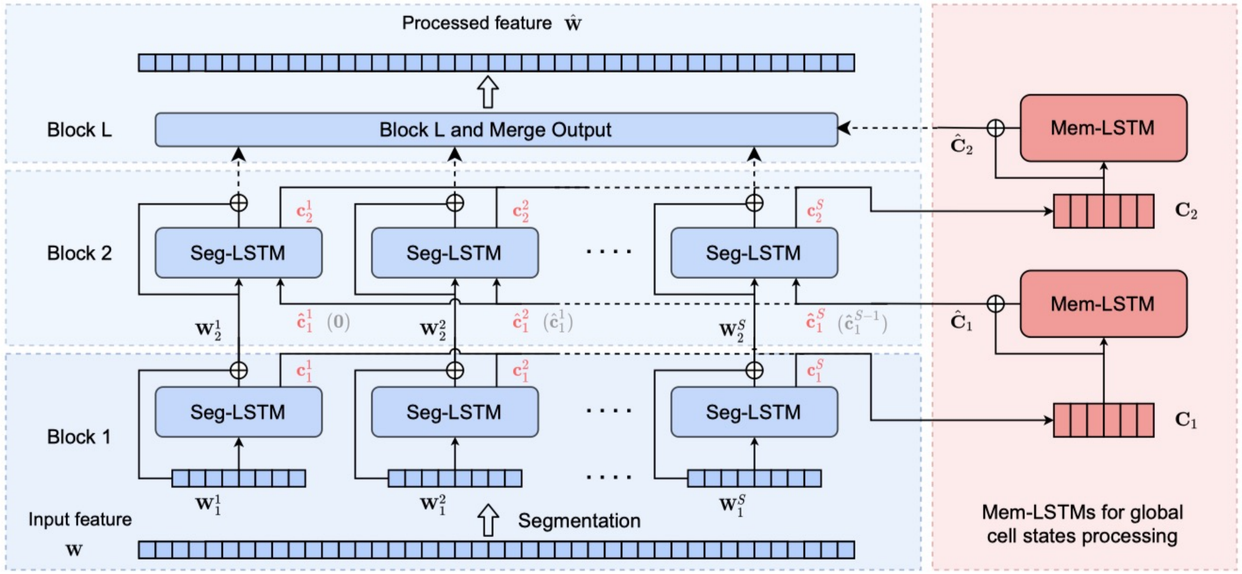
→ Super-long sequence (> 100k frame per minute)

Experimental Details

- Dataset:
  - A simulated meeting style dataset derived from LibriSpeech
  - Each session lasts for ~90 seconds
  - 3-5 active speakers in each session
  - Overlap ratio is between 50% to 80%

The Proposed Skipping Memory (SkiM) LSTM Model

- Segmentation-LSTM
  - Local processing
  - Sequence -> State vector
- Memory-LSTM
  - Process state vectors
  - Global information sync
- Compared to DPRNN
  - More efficient long sequence modeling
  - Less computational cost



Experimental Results

Table 1: The overall STOI, SDR improvement (SDRi) and high-overlap SDR improvement (SDRi50) comparison on different models.

Model	Causal	Stride Size	Model size (M)	MACs (G/s)	SDRi (dB)	SDRi50 (dB)
TCN	no	20	3.4	2.7	13.5	5.9
DPRNN	no	20	9.6	14.6	**19.2**	9.0
	yes	20	4.9	7.5	16.6	7.6
	yes	10	4.9	14.7	16.8	7.7
SkiM	no	20	15.9	**3.8**	18.7	**9.2**
	yes	20	6.0	**2.0**	**17.3**	**8.0**
	yes	10	6.0	**3.9**	**17.1**	**7.8**

Table 2: Real-time factor (RTF) and latency evaluation for causal models. (Tested out with a single-core Intel Ivy Bridge CPU @ 1.9GHz)

Model	Stride size	Ideal latency	MACs (G/s)	RTF	Latency
DPRNN	20	1.25 ms	7.5	0.98	2.47 ms
	10	0.625 ms	14.7	1.98	null
SkiM	20	1.25 ms	**2.0**	**0.23**	1.54 ms
	10	0.625 ms	3.9	0.46	**0.92** ms

Table 3: Ablation studies for Mem-LSTMs in SkiM. Replace the global-synchronized hidden (**h**) and cell (**c**) states with zeros (**0**) or unprocessed local states (**id**)

Model	Mem-LSTM	Model size (M)	MACs (G/s)	SDRi (dB)	SDRi50 (dB)
SkiM	h, c	15.9	3.8	18.7	9.2
	h, 0	10.4	3.8	17.8	8.6
	0, c	10.4	3.8	15.6	7.8
	0, 0	4.9	3.8	12.5	7.1
	id, id	4.9	3.8	12.5	7.1

Table 4: Comparison with other models on WSJ0-2mix Benchmark. (*):MACs per second estimated by us.

Model	Model size (M)	MACs (G/s)	SI-SNRi	SDRi
DPCL++ [28]	13.6	-	10.8	-
ADANet [29]	9.1	-	10.4	10.8
WA-MISI-5 [5]	32.9	-	12.6	13.1
Conv-TasNet-gLN [8]	5.1	3.2*	15.3	15.6
Deep CASA [30]	12.8	-	17.7	18.0
FurcaNeXt [31]	51.4	-	-	18.4
DPRNN-KS2 [10]	2.6	38.9*	18.8	19.0
DPRNN-KS8 [10]	2.6	9.8*	17.0	17.3
SepFormer [11]	26.0	32.1*	**20.4**	**20.5**
SkiM-KS2	5.9	19.7	18.3	18.7
SkiM-KS8	5.9	4.9	17.4	17.8