实时多人对话的语音识别

陈卓
Microsoft

Content

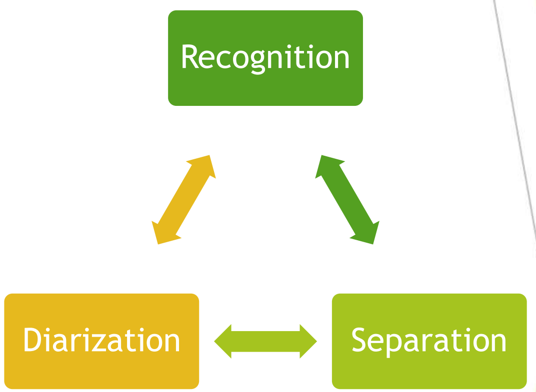
- Streaming conversation transcription: What and Why?
- Modularized solution: Continuous Speech Separation
- End to end solution: tokenized Serialize Output Training

Streaming conversation recognition

- “Who speak what at when”, on
  - Unsegmented continuous recordings
  - Different recording conditions & setup
  - Streaming recognition
- Legacy to borrow from previous speech systems
  - Long form audio recognition
  - Far-field speech processing
  - Speaker identification
- New challenges
  - Multi-speaker conversation
  - Speech overlap
  - Quick speaker turn

Multi-speaker processing: why

- Quick math: Word Error Rate(WER) impact of overlapped speech
  - Assuming:
    - Meeting words: 100
    - WER on single speaker： 10%
    - WER on fully overlapped speech: 80%
    - Overlap ratio: 10% (commonly 5%~25%)
  - What is the final WER and WER increase?
    - Error count: (100*0.9)*0.1 + (100*0.1) *0.8 =17
    - 10%-> 17%, 70% WER increase!
- Obstacle:
  - The multi-speaker audio breaks the fundamental assumption of previous speech systems

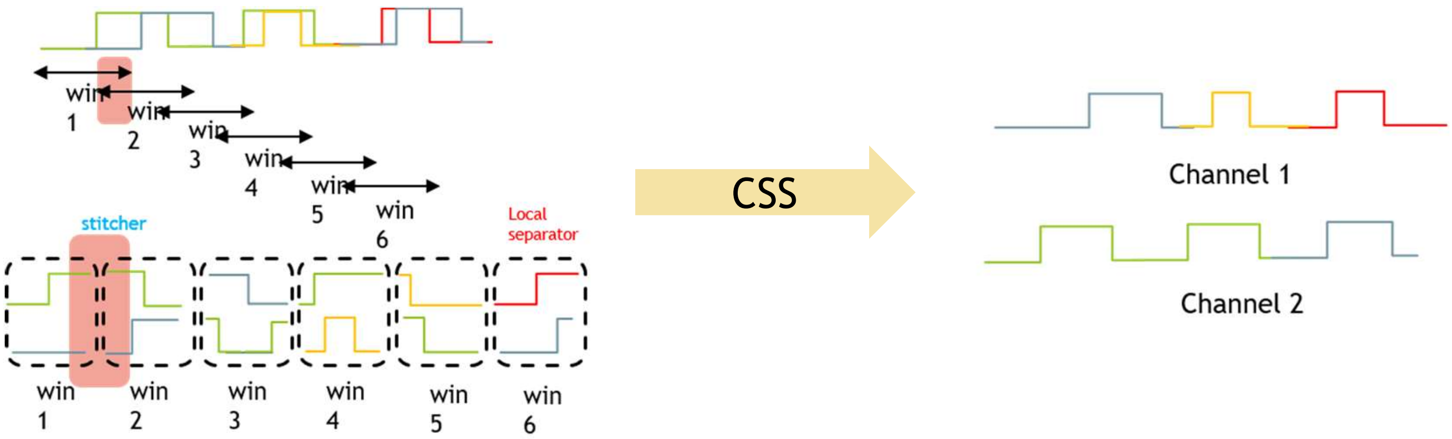


Solution for streaming multi-speaker processing

- **Modularized solution: Continuous Speech Separation**
  - Additional speech separation module for multi-speaker processing
  - Other modules remains unchanged

- End to end solution: tokenized Serialize Output Training
  - Modeling the multi-talker speech directly

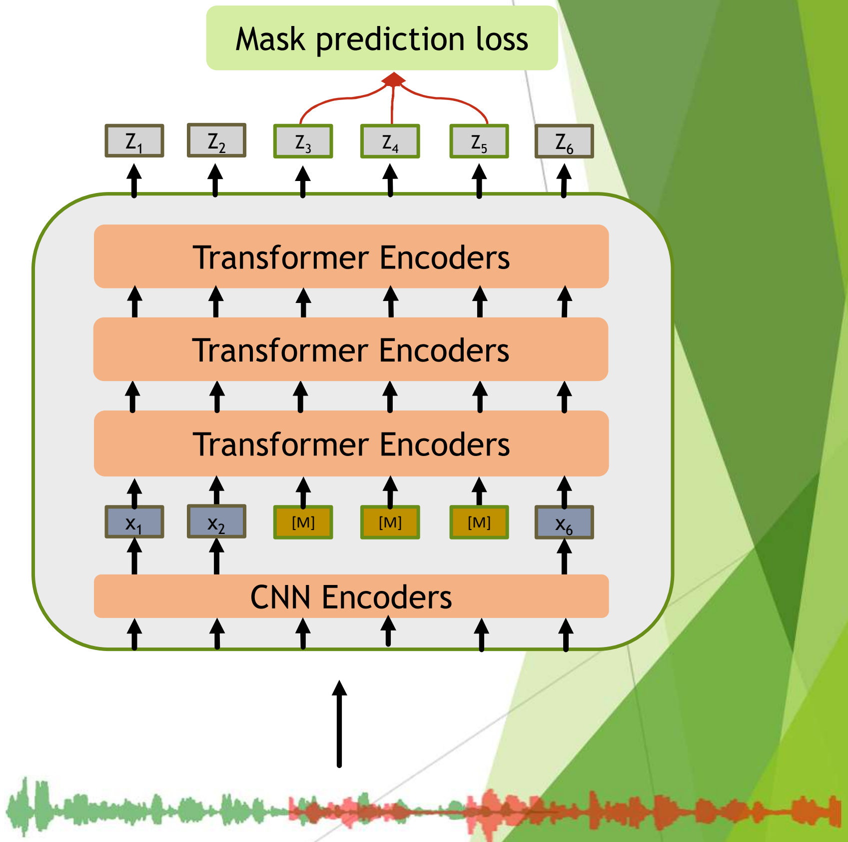
Continuous speech separation



- Basic components
  - Segmentor
  - Separator
  - Stitcher
- Properties
  - Processing the input mixture audio continuously
  - Short window separation ensures the 2 active speaker per window
  - Separated channels contains sparsely aligned, overlap free utterances for other speech components

WavLM

- A simple self supervised learning system specifically designed for non-ASR tasks
  - Pseudo labeling through clustering
  - Mask prediction loss
- Utterance mixing training
  - Artificially mixed training sample
  - Target at token from unmixed speech
  - Enforcing the speaker distinction in embedding
- State of the art performance in multiple tasks



WavLM

- A simple SSL system specifically designed for non-ASR tasks
  - Pseudo labeling through clustering
  - Mask prediction loss
- Utterance mixing training
  - Artificially mixed training sample
  - Target at token from unmixed speech
  - Enforcing the speaker distinction in embedding
- State of the art performance in multiple tasks
  - Multi-speaker ASR
  - Speech diarization
  - Speaker verification

Model	0S	0L	OV10	OV20	OV30	OV40
Conformer (SOTA)	4.5	4.4	6.2	8.5	11	12.6
HuBERT base	4.7	4.6	6.1	7.9	10.6	12.3
UniSpeech-SAT base	4.4	4.4	5.4	7.2	9.2	10.5
UniSpeech-SAT large	4.3	4.2	5.0	6.3	8.2	8.8
WavLM base+	4.5	4.4	5.6	7.5	9.4	10.9
WavLM large	4.2	4.1	4.8	5.8	7.4	8.5

Speech separation: LibriCSS

Model	spk_2	spk_3	spk_4	spk_5	spk_6	spk_all
EEND-vector clustering	7.96	11.93	16.38	21.21	23.1	12.49
EEND-EDA clustering (SOTA)	7.11	11.88	14.37	25.95	21.95	11.84
HuBERT base	7.93	12.07	15.21	19.59	23.32	12.63
HuBERT large	7.39	11.97	15.76	19.82	22.10	12.40
UniSpeech-SAT large	5.93	10.66	12.9	16.48	23.25	10.92
WavLM Base	6.99	11.12	15.20	16.48	21.61	11.75
WavLm large	6.46	10.69	11.84	12.89	20.70	10.35

Speaker diarisation: Callhome

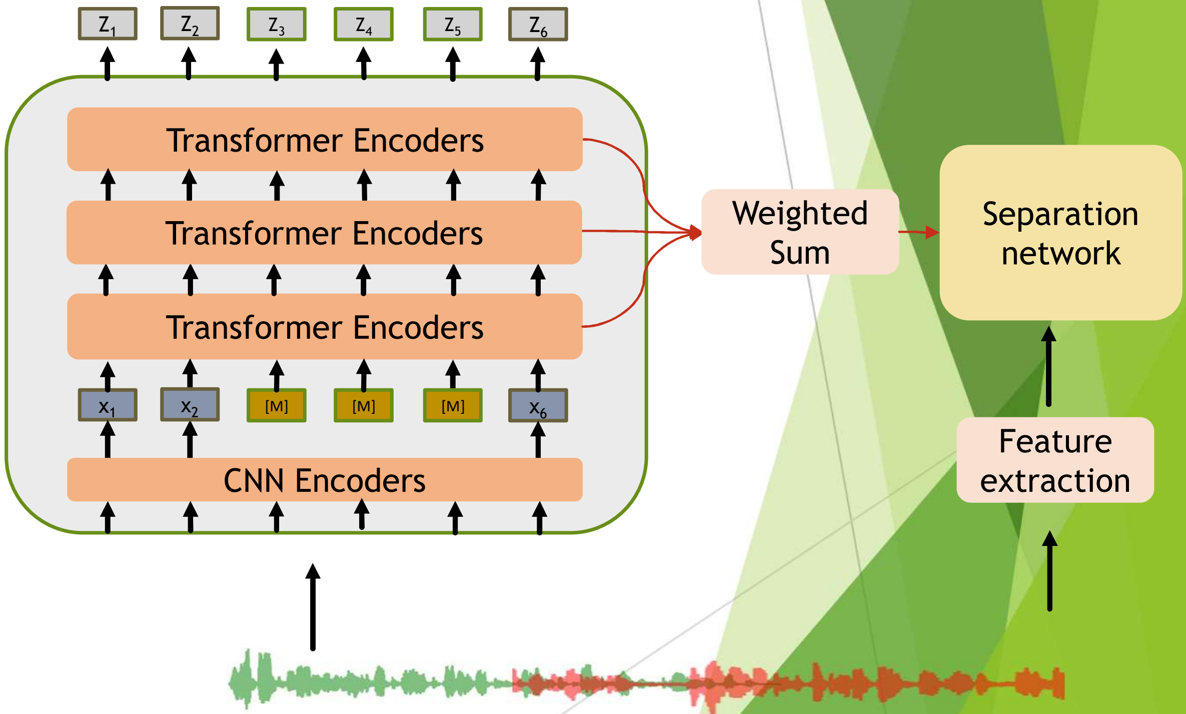
Results

#	User	Entries	Date of Last Entry	DCF ▲	EER ▲
1	Strasbourg-Spk	15	10/25/22	0.058 (1)	1.153 (2)
2	ravana	5	09/14/22	0.062 (2)	1.212 (3)
3	KristonAI	9	09/23/22	0.071 (3)	1.120 (1)
4	wigi	7	09/23/22	0.096 (4)	1.530 (4)

Speaker verification: Vox-celeb

WavLM based CSS: towards deployment

- WavLM based speech separation
  - Concatenation of acoustic feature and embedding
  - Weighted averaged embeddings from WavLM
  - Allowing separation network to access significantly larger data scale
- Towards real application
  - Performance improvement
    - Larger data scale
  - Computation reduction
    - Model configuration
    - Partial layer finetuning



WavLM based CSS: towards real application

- Better performance
  - Consistent performance gain as pretraining data increases
  - Small WavLM model still outperforms the baselines
- Computation reduction
  - Lower layer finetuning shows comparable performance with full finetuning
- Real meeting evaluating
  - 7% relative WER improvement
  - 38% computation reduction

ID	SSL		SS	RTF	WER (%)	
	Model	Data			Far-mix	Clean-mix
B1	-	-	SS-59	$\times$ 0.21	22.7	22.7
B2	-	-	SS-79	$\times$ 0.27	23.2	23.8
B3	-	-	SS-92	$\times$ 0.32	23.1	23.6
P1	WavLM Large	S	SS-59	$\times$ 0.55	21.5	22.8
P2	WavLM Large	M	SS-59	$\times$ 0.55	20.6	18.2
P3	WavLM Large	L	SS-59	$\times$ 0.55	19.1	17.5
P4	WavLM Large	L	SS-26	$\times$ 0.47	19.2	20.1
P5	WavLM Base	L	SS-26	$\times$ 0.25	20.4	19.2
P6	WavLM Small	L	SS-26	$\times$ 0.20	20.2	20.2

WER for Data scale and model configuration search

ID	SSL			SS	RTF	WER (%)	
	Model	$f^{\mathrm{wl}}(ms)$	FT-layers			Far-mix	Clean-mix
P3		20			$\times$ 0.55	19.1	17.5
L1	WavLM-Large	30	24	SS-59	$\times$ 0.46	21.9	24.8
L2		40			$\times$ 0.38	22.8	25.7
P4			24		$\times$ 0.47	19.2	20.0
S1			16		$\times$ 0.38	19.1	18.7
S2	WavLM-Large	20	12	SS-26	$\times$ 0.35	19.9	18.4
S3			8		$\times$ 0.31	19.7	18.6
S4			4		$\times$ 0.27	21.0	21.3

WER for computation reduction

ID	SSL		SS	RTF	AMI WER (%)		ICSI WER (%)	
	Model	FT-layers			dev	eval	dev	eval
B1	-	-	SS-59	$\times$ 0.21	21.6	25.0	23.2	20.7
S3'	WavLM Large	8	SS-26	$\times$ 0.31	19.1	22.6	17.8	16.5
S4'	WavLM Base	12	SS-26	$\times$ 0.25	19.4	22.9	18.6	17.2
S8'	WavLM Base	12	SS-9.5	$\times$ 0.19	19.5	22.9	18.0	17.0
S9'	WavLM Small	8	SS-9.5	$\times$ 0.13	19.6	23.3	18.3	18.5

WER on real meeting corpus

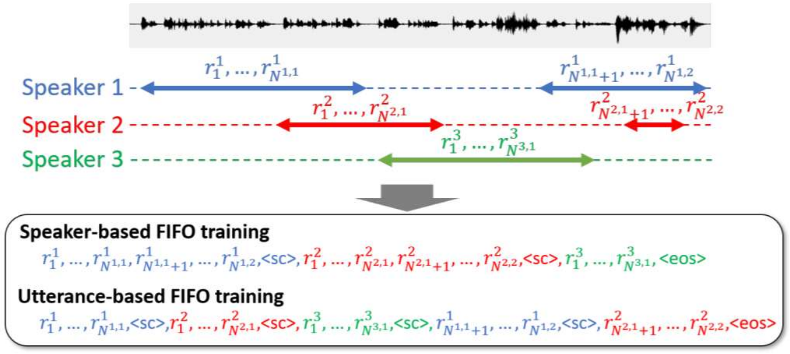
Solution for streaming multi-speaker processing

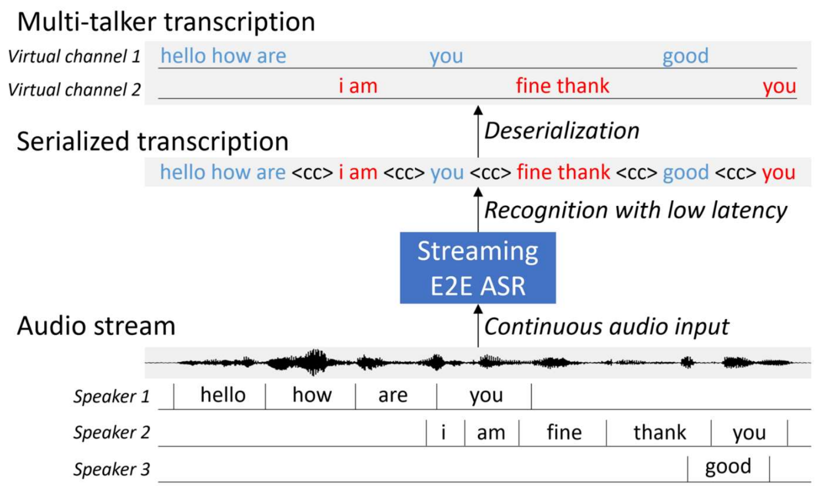
- Modularized solution: Continuous Speech Separation
  - Additional speech separation module for multi-speaker processing
  - Other modules remains unchanged

- **End to end solution: tokenized Serialize Output Training**
  - Modeling the multi-talker speech directly

SOT and tSOT

- Serialized Output Training
  - Utterance-wise Serialized output
  - Speaker based FIFO training
  - Sequence to sequence ASR backbone
  - Arbitrary number of overlapped speakers
  - Offline model
- tSOT
  - Token-wise serialized output
  - Transducer ASR backbone
  - Fixed number of outputting channel
  - Streaming model





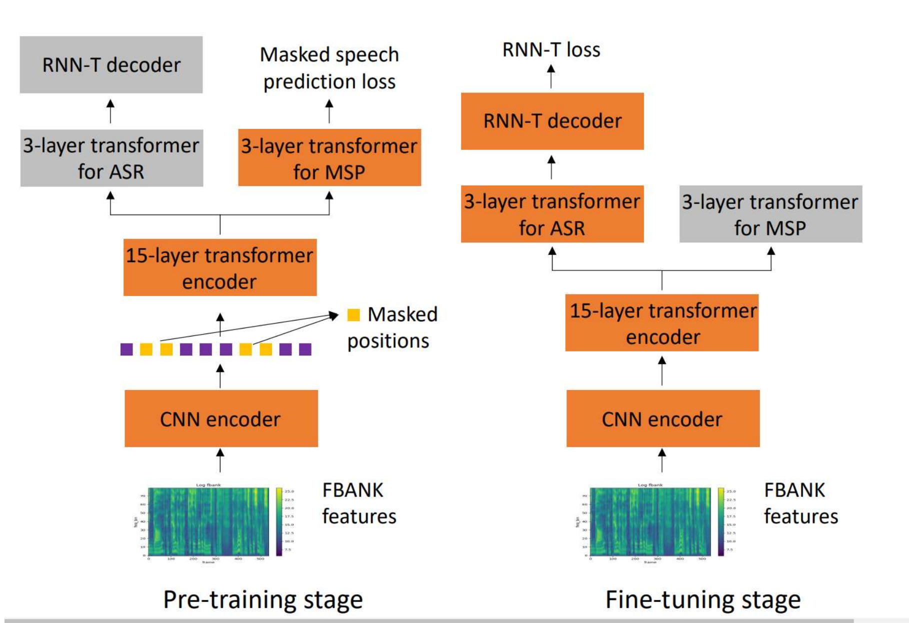
State of the art performance on LibriCSS testset

System	Algorithmic latency	WER (%) for different overlap ratio						
		0L	0S	10	20	30	40	Avg.
(Non-streaming ASR models with speech separation)								
BLSTM-CSS + Hybrid ASR [6]	1.2 sec‡ + (utterance length)*	16.3	17.6	20.9	26.1	32.6	36.1	24.9
Conformer-CSS + Transformer-AED-ASR w/ LM [9]	1.2 sec‡ + (utterance length)*	6.1	6.9	9.1	12.5	16.7	19.3	11.8
Conformer-CSS + Transformer-AED-ASR w/ LM [43]	1.2 sec‡ + (utterance length)*	6.4	7.5	8.4	9.4	12.4	13.2	9.6
(Streaming ASR models)								
SURT w/ DP-LSTM [44]	350 msec	9.8	19.1	20.6	20.4	23.9	26.8	20.1
SURT w/ DP-Transformer [44]	350 msec	9.3	21.1	21.2	25.9	28.2	31.7	22.9
Single-talker TT-18	160 msec	7.0	7.3	14.0	20.9	27.9	34.3	18.6
Single-talker TT-36	160 msec	**6.5**	6.7	13.1	20.4	27.0	34.0	18.0
t-SOT TT-18 (proposed)	160 msec	7.5	7.5	8.5	10.5	12.6	14.0	10.1
t-SOT TT-36 (proposed)	160 msec	6.7	**6.1**	**7.5**	**9.3**	**11.6**	**12.9**	**9.0**

- Better performance
- Low processing latency
- Simplistic implementation

Stronger together: SSL + tSOT

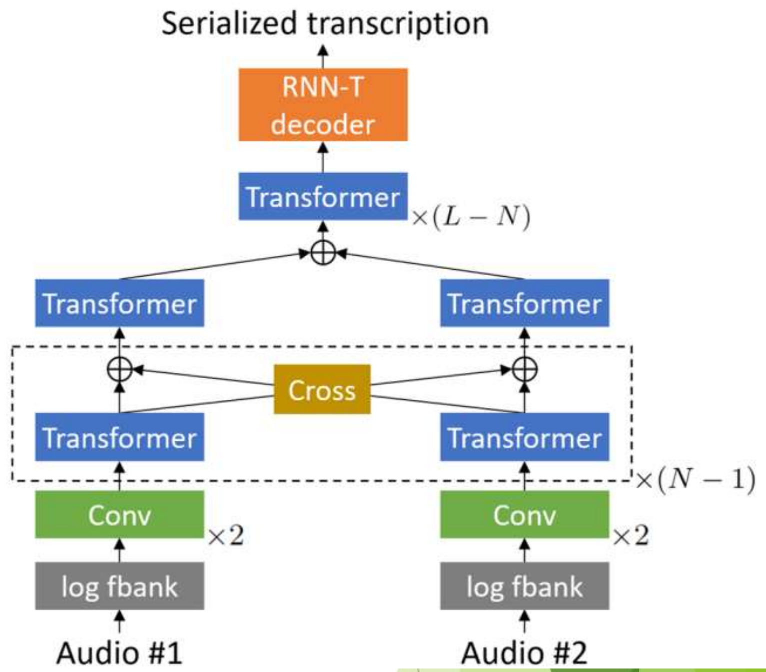
- Leveraging the advantage of the self supervised learning
  - WavLM style SSL for speaker information extraction
  - End to end ASR training
- Multi-fold exploration for SSL + tSOT
  - A bi-label WavLM style objective function
  - Tokenizer variation
  - Utterance mixing configuration
- Significant improvement over the pure supervised system



Pre-training		Dev WER (%)		Test WER (%)	
Objective	Quantizer	1spk	2spk	1spk	2spk
-	-	15.42	39.12	15.69	39.52
MSP	FBANK	13.17	36.13	13.20	35.29
Bi-label MSP	FBANK	13.29	25.68	13.90	25.78
MSP	HuBERT	10.77	17.24	11.30	17.25
Bi-label MSP	HuBERT	10.82	15.84	11.19	15.30
MSP	Phoneme	9.80	15.45	9.96	15.13
Bi-label MSP	Phoneme	**9.47**	**13.89**	**9.84**	**13.74**

Stronger together: CSS + tSOT

- Combine the advantage
  - Single or multichannel CSS
  - End to end ASR optimization
- Significantly advanced the state of the art on AMI dataset
  - From 17.7% to 15.5% (12.4% WERR)
  - Less data: 1M vs. 75K
  - Smaller network: 8B vs. 200M
  - Offline vs. Streaming modeling
  - 1ch vs. 8 ch



ID	Front-end configuration				Back-end configuration					Back-end training			Test segment	WER (%)	
	In	Out	Param.	Latency	Model	In	Cross	Param.	Latency	1ch-PT	2ch-PT	FT		dev	eval
B1	-	-	-	-	Single-talker TT18	1	-	82M	0.16 sec	75K	-	-	utt	38.0	40.8
B2	-	-	-	-	Single-talker TT18	1	-	82M	0.16 sec	75K	-	AMI	utt	27.3	30.3
B3	8	1	2M	0.8 sec	Single-talker TT18	1	-	82M	0.16 sec	75K	-	AMI	utt	25.8	27.9
B4	-	-	-	-	t-SOT TT18	1	-	82M	0.16 sec	75K-sim	-	-	utt-gr	35.5	40.3
B5	-	-	-	-	t-SOT TT18	1	-	82M	0.16 sec	75K-sim	-	AMI	utt-gr	21.6	25.3
B6	8	1	2M	0.8 sec	t-SOT TT18	1	-	82M	0.16 sec	75K-sim	-	AMI	utt-gr	20.7	23.0
P1	8	2	2M	0.8 sec	t-SOT 2ch-TT18	2	-	82M	0.16 sec	75K-sim	-	AMI	utt-gr	19.3	21.7
P2	8	2	2M	0.8 sec	t-SOT 2ch-TT18	2	-	82M	0.16 sec	75K-sim	75K-sim	AMI	utt-gr	18.6	21.1
P3	8	2	2M	0.8 sec	t-SOT 2ch-TT18	2	Eq. (1)	82M	0.16 sec	75K-sim	75K-sim	AMI	utt-gr	18.5	21.0
P4	8	2	2M	0.8 sec	t-SOT 2ch-TT18	2	Eq. (2)	84M	0.16 sec	75K-sim	75K-sim	AMI	utt-gr	18.3	20.6
P5	8	2	2M	0.8 sec	t-SOT 2ch-TT36	2	Eq. (2)	142M	0.64 sec	75K-sim	75K-sim	AMI	utt-gr	15.3	17.4
P6	8	2	2M	0.8 sec	t-SOT 2ch-TT36	2	Eq. (2)	142M	2.56 sec	75K-sim	75K-sim	AMI	utt-gr	14.4	16.5
P7	8	2	56M	0.8 sec	t-SOT 2ch-TT36	2	Eq. (2)	142M	2.56 sec	75K-sim	75K-sim	AMI	utt-gr	13.7	15.5

Conclusion

- Multi-talker problem is important for modern conversation transcription
- CSS provides a simple yet effective way for processing streaming conversation audio stream
- Self supervised learning significantly boost the performance for CSS models on real meeting data
- tSOT method shows strong performance for low latency multi-speaker ASR task
- Variations of tSOT show improved performance and achieves state of the art performance in AMI dataset

Thanks for attending ~

Questions?