2024.04.17 (Wed)
SAM-P2.9
Poster Zone 6C

Paper Link

CST-FORMER

TRANSFORMER WITH CHANNEL-SPECTRO-TEMPORAL ATTENTION FOR SOUND EVENT LOCALIZATION AND DETECTION

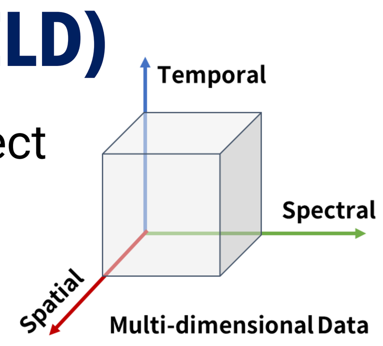
Yusun Shul, Jung-Woo Choi

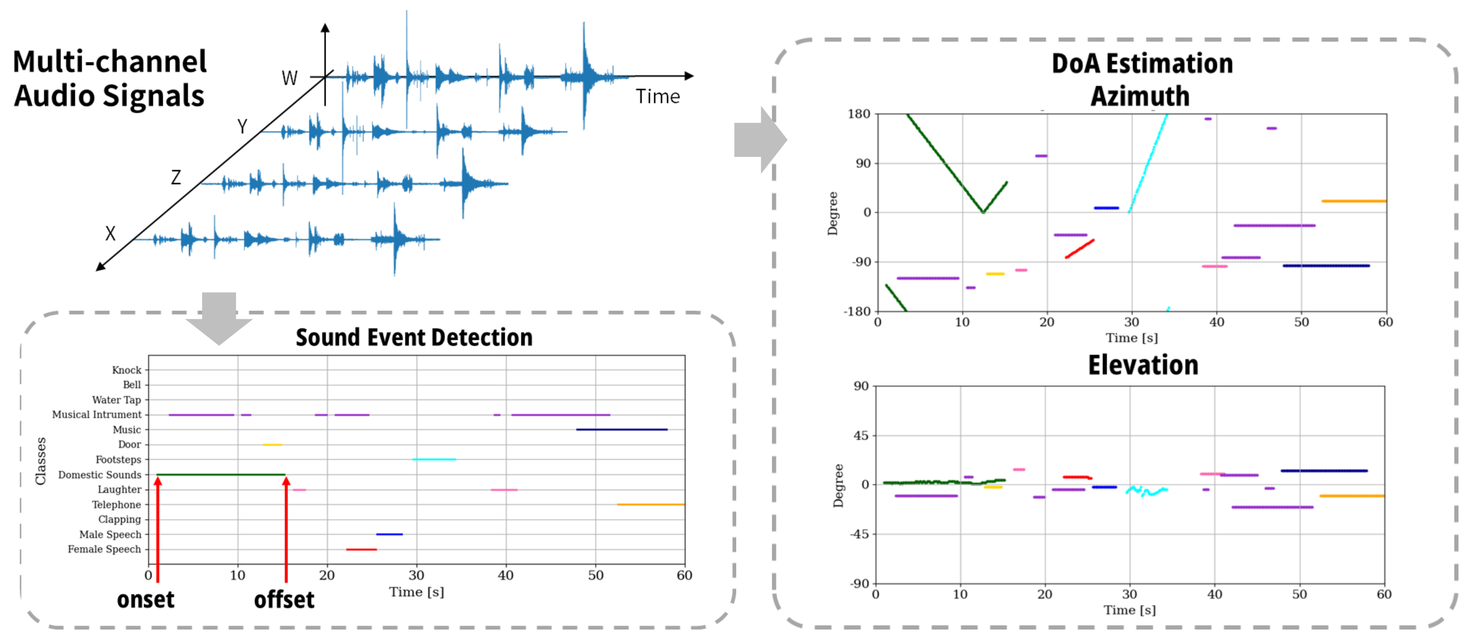
School of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST)

Github Link

Sound Event Localization and Detection (SELD)

- Sound Event Detection (SED): classify sound events and detect the onset and offset in the temporal domain
- Localization: estimate the direction-of-arrival (DoA) of SE
- Using multi-channel audio signals
- Importance of the **spatial**, **spectral**, and **temporal** information





Limitations of Conventional Models

- Focused on learning the temporal context of multichannel signals
- Limited usage of multidimensional data
- Channel & spectral information used as the embedding of temporal sequence
- Easily overfitted with a small number of real recorded data

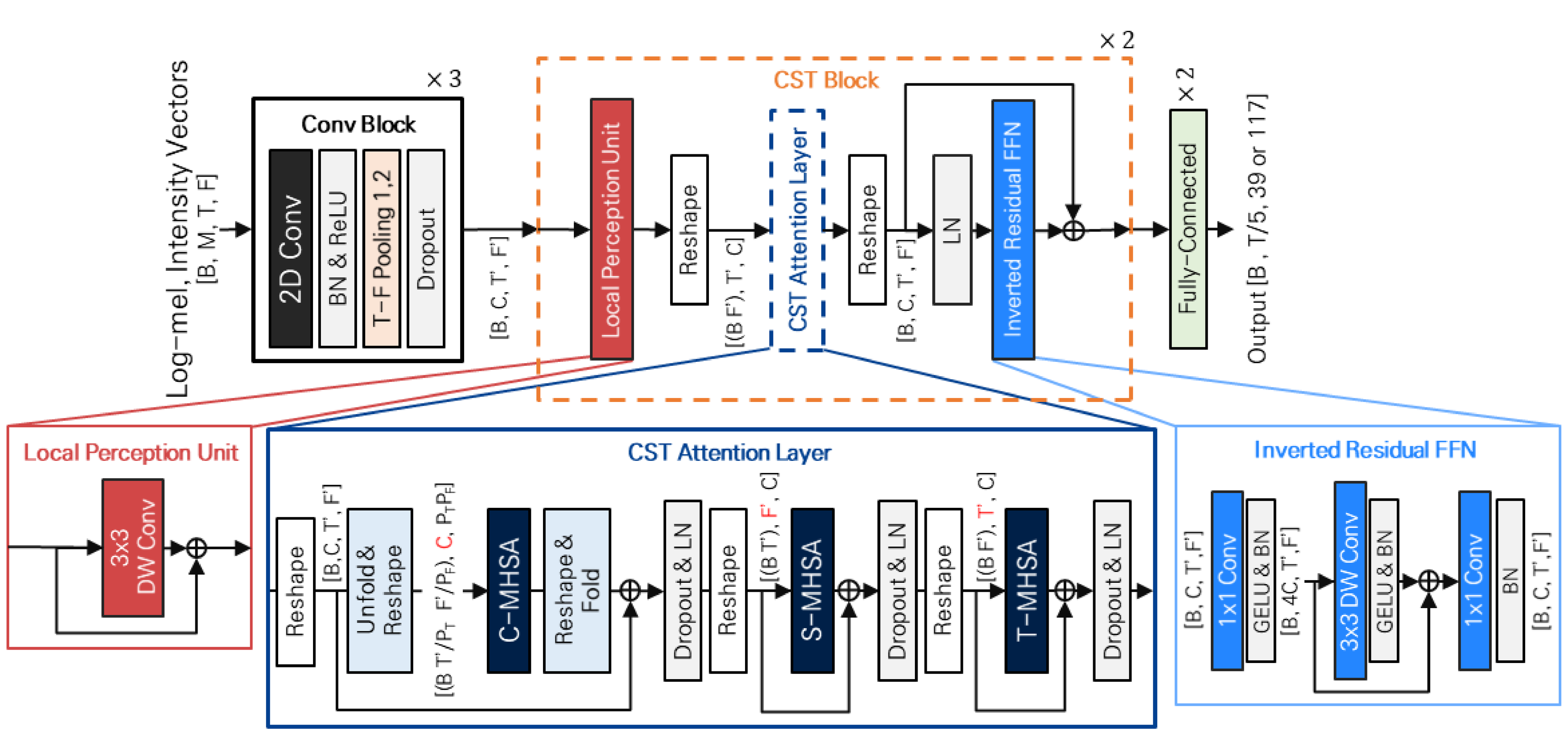
Model	Encoder	Decoder	Attention Domain	Pooling Location	Output Type	Parameter Size
2022 Baseline[10]	CNN	GRU	-	Front	Multi-ACCDOA	0.60M
2023 Baseline[1]	CNN	GRU, MHSA	**Time**	Front	Multi-ACCDOA	0.74M
ResNet-Conformer[16]	ResNet	Conformer	**Time**	Front Middle End	Multi-ACCDOA	58M
EINV2[6]	CNN	Conformer	**Time**	Front	Multi-task	85M
DST Attention[21]	CNN	DST-MHSA	Frequency Time	Front	Multi-ACCDOA	0.30M
CST-former (Proposed)	CNN	CST-transformer	**Channel Frequency Time**	Front Middle	Multi-ACCDOA	0.39M

Contributions

- **Transformer** with **multidimensional attention** layers for SELD
  - Attention modules learning context of each channel (spatial), spectral, and temporal domain
  - **Two embedding generation methods** for channel attention (CA) : Divided Channel Attention, Unfolded Local Embedding

Proposed Architecture

Channel-Spectro-Temporal Transformer (CST-former)



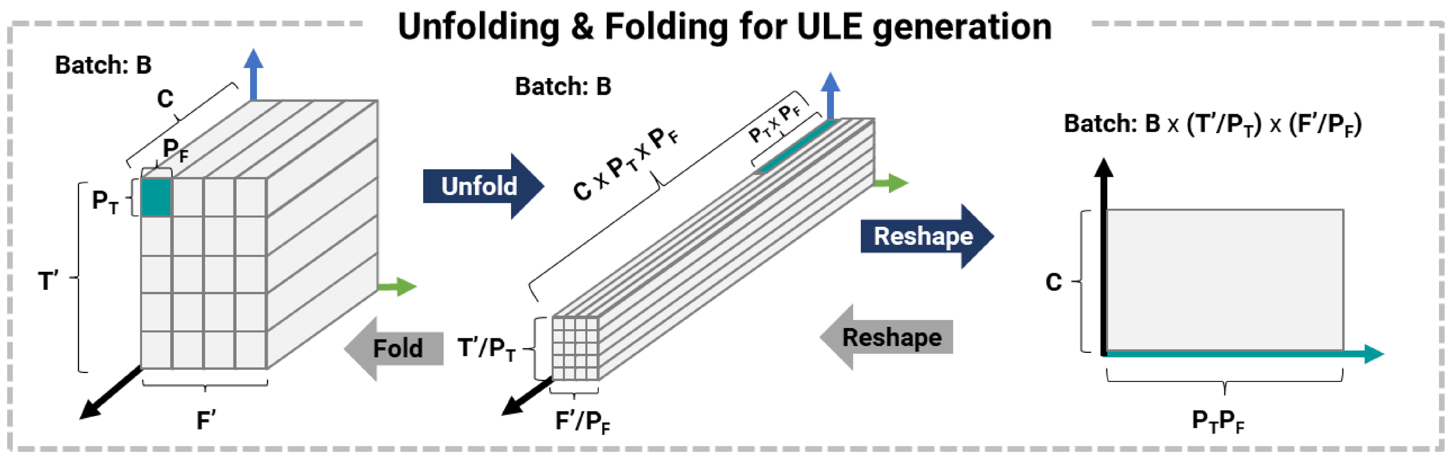
- **CST Block**
  - Structures from convolution meets transformer (**CMT**)[22]
    - **Local Perception Unit (LPU)**
      : Local temporal and spectral information extracted by 3x3 depth-wise convolution
    - **Inverted Residual Feed Forward Network (IRFFN)**
      : Substitutes the FFN of conformer
- **CST Attention Layer**
  - Independent attention layers for different domains
  - Spectral and temporal attention uses the encoded channel as embedding
  - Two different ways of **embedding generation** for channel attention

Embedding Generation Methods for CA

	Conv Block	LPU, IRFFN	C-MHSA	S-MHSA	T-MHSA
DCA	[(B M), 1, T, F]	[(B M), C, T', F']	[(B T' F'), M, C]	[(B T' M), F', C]	[(B F' M), T', C]
ULE	[B, M, T, F]	[B, C, T', F']	[(B T'/P_T F'/P_F), C, P_TP_F]	[(B T'), F', C]	[(B F'), T', C]

- **Divided Channel Attention (DCA)**
  - The microphone input channel (M) is not encoded in the conv block and is utilized as the sequence of CA
  - Encoded channel information (C) from the conv block is used as embedding
- **Unfolded Local Embedding (ULE)**
  - The microphone input channel (M) is encoded in the conv block
  - Encoded channel information (C) is used as the sequence of CA
  - ULE generated with the unfold layer is utilized as the embedding for the CA
  - **Local T-F bins are affected by CA** while **maintaining** the **global T-F context**

Unfolding & Folding for ULE generation



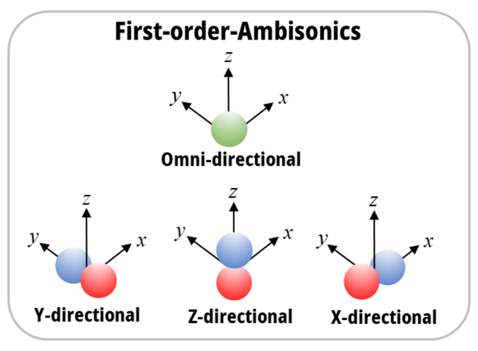
- **Time-Frequency Pooling**
  - Different kernels in the conv block
  - Matching the temporal resolution of the target label
  - Minimizing the computational cost without sacrificing performance

	Layer 1	Layer 2	Layer3
T-F pooling 1 (Front)	(5,2)	(2,2)	(1,1)
T-F pooling 2 (Middle)	(1,1)	(2,2)	(2,5)

Experimental Results

Data	Type	Room Variations	Duration
Train	Real	12	~5 h
Test	Real	4	~2.5 h
Train	Synthetic	10	20 h

- **Data**
  - DCASE 2022 / 2023 challenge task3
    : First-order-Ambisonics (FoA) format, Maximum polyphony of five, 13 different classes
  - 4 Log-mel spectrograms & 3 intensity vectors



- **Ablation Studies for CST-former**
  - Results on test data of **DCASE 2022** task3 dataset

Model	CMT	CA	Pooling Location	SELD Score↓	Error Rate↓	F-score↑	Localization Error↓	Localization Recall↑
2022 Baseline[10]	-	-	Front	0.5345	0.72	24.0	26.6	49.0
2023 Baseline[1]	-	-	Front	0.5006 (6.3%↓)	0.70	29.1	23.8	53.9
DST Attention[21]	-	-	Front	0.4861 (9.0%↓)	0.68	31.4	22.6	54.7
	O	-	Front	0.4563 (14.6%↓)	0.66	36.6	22.5	59.1
CST-former (Proposed)	-	DCA	Front	0.4749 (11.1%↓)	0.68	33.7	22.6	56.9
		ULE	Front	0.4698 (12.1%↓)	0.67	36.4	21.3	54.5
	O	DCA	Front	0.4480 (16.2%↓)	0.65	36.8	23.2	61.9
		ULE	Front	0.4286 (19.8%↓)	0.66	40.0	21.3	**66.4**
			Middle	**0.4162** (22.1%↓)	**0.59**	**42.6**	**20.5**	61.3

- **Verification with DCASE challenge task3 datasets**
  - Performance compared with various SELD models (**DCASE 2022**)

Model	Pooling Location	SELD Score↓	Error Rate↓	F-score↑	Localization Error↓	Localization Recall↑
2022 Baseline[10]	Front	0.5345	0.72	24.0	26.6	49.0
ResNet-Conformer[16]	Front	0.4928 (7.8%↓)	0.72	27.0	25.4	62.0
	Middle	0.4794(10.3%↓)	0.70	32.0	21.2	58.0
	End	0.4710 (11.9%↓)	0.71	31.0	22.3	64.0
EINV2[6]	Front	0.5000 (6.4%↓)	0.75	32.0	24.0	56.0
CST-former (Proposed)	Front	0.4286 (19.8%↓)	0.66	40.0	21.3	**66.4**
	Middle	**0.4162** (22.1%↓)	**0.59**	**42.6**	**20.5**	61.3

  - Performance on **DCASE 2023** challenge task3 test dataset

Model	Pooling Location	SELD Score↓	Error Rate↓	F-score↑	Localization Error↓	Localization Recall↑
2023 Baseline[1]	Front	0.4791	0.57	29.9	22.0	47.7
DST Attention[21]	Front	0.4345 (9.3%↓)	0.58	39.5	20.0	55.8
CST-former (Proposed)	Front	0.4111 (14.2%↓)	0.58	42.5	18.4	61.1
	Middle	**0.4019** (16.1%↓)	**0.56**	**42.7**	**17.9**	**62.0**

Conclusion

- **[CST-former]** Distinct **multidimensional attention mechanisms** for SELD task
- **[ULE]** Embedding generation for CA, utilizing the **unfolded local temporal and spectral information** as embedding
- Significant performance improvements even without data augmentation

shulys@kaist.ac.kr, jwoo@kaist.ac.kr