SoundCompass

Navigating Target Sound Extraction with Effective Directional Clue Integration in Complex Acoustic Scenes

Dayun Choi & Jung-Woo Choi*
cdy3773@kaist.ac.kr, jwoo@kaist.ac.kr

SMART SOUND LAB
KAIST EE
School of Electrical Engineering

AASP-P22.7
May 7th (Thu) 16:30-18:30
Poster Area 24

Paper Link | Demo Link

Highlights

Target sound extraction (TSE) on multichannel audio

- Multichannel Target Extraction using Directional Clue
- Centered on **Spectral Pairwise INteraction (SPIN)** input feature & **Spherical harmonics (SH)-based** directional clue

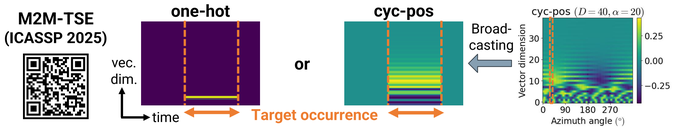
Research Motivation

Limitations of input feature selection

- Relying on **hand-crafted** spatial features (e.g., inter-channel level/phase difference (ILD/IPD))
- Suboptimal in capturing essential spatial information

Directional Clue Encoding & Fusion

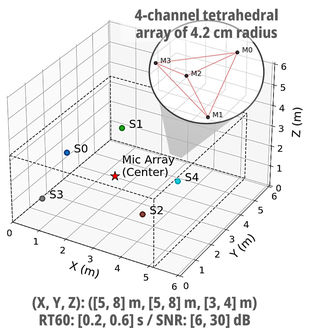
- **Microphone-position-dependent** representations (e.g., TPD)
- **Discrete** encodings ignoring angular continuity and periodicity
- **Target-activity-dependent** fusion via temporal broadcasting of one-hot or cyclic positional (cyc-pos) embeddings



Dataset & Evaluation Metrics

Auditory Scene Analysis V2 (ASA2) dataset

- Synthetic mixtures of 16 kHz audio, including **2–5 foreground sources from 13 audio classes** and one background noise



① Signal quality evaluation

- **Target signal isolation and noise suppression performance**
- **SNR and SI-SNR improvements** compared to the mixture

② Spatial consistency evaluation

- Evaluation of **inter-channel relationship preservation**
- Mean absolute error between GT and estimation of **ILD, IPD, and ITD via GCC-PHAT** across all microphone pairs

SoundCompass Framework

Spectral Pairwise INteraction (SPIN) input feature

- Pairwise multiplication of cos/sin components derived from the phase of multichannel complex short-time Fourier transform (STFT) across channels
- **Comprehensive and stable spatial modeling** via **learnable correlations**

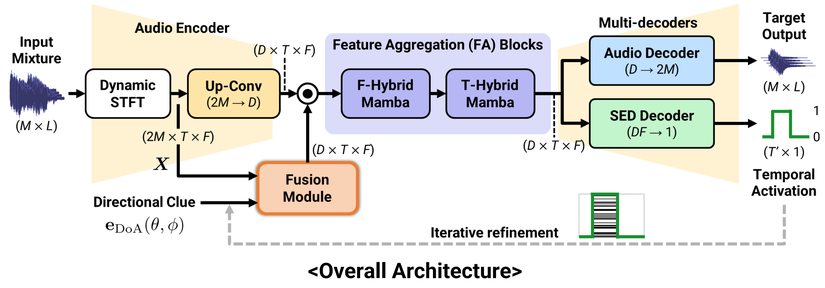
$$\mathbf{SPIN}(t,f) = \mathbf{P}_{t,f}\mathbf{P}_{t,f}^{\top} \text{ where } \mathbf{P}_{t,f} = \text{stack}_{\,0\le m<M}\left(\cos(\angle \boldsymbol{X}_m(t,f)), \sin(\angle \boldsymbol{X}_m(t,f))\right)$$

Spherical harmonics (SH)-based direction-of-arrival (DoA) clue

- **Continuous directional representation without discretization and coordinate separation** unlike one-hot and cyc-pos embeddings

$$\mathbf{e}_{\text{DoA}} = \text{stack}_{\substack{0\le n\le N \\ -n\le m\le n}}\left(\text{Re}(Y_n^m(\theta,\phi)), \text{Im}(Y_n^m(\theta,\phi))\right)$$

$$\text{where } Y_n^m(\theta,\phi) = \sqrt{\frac{(2n+1)}{4\pi}\frac{(n-m)!}{(n+m)!}}P_n^m(\cos\theta)e^{im\phi}$$



<Overall Architecture>

Subband-wise clue fusion

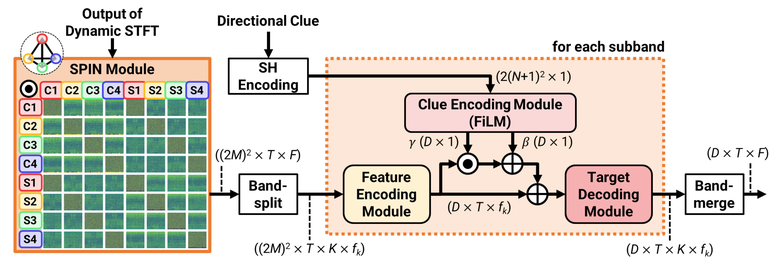
- **Frequency-overlapping subband processing** tailored to frequency-varying nature of spatial features to enable **fine-grained spatial conditioning**

Iterative refinement with temporal clue

- **Recursive injection of target temporal activation** from the sound event detection (SED) decoder to **overcome initial temporal uncertainty**

Loss functions

- A linear combination of SNR and SI-SNR loss (9:1) for audio decoder & binary cross entropy (BCE) loss for SED decoder



<Fusion Module>

Experimental Results

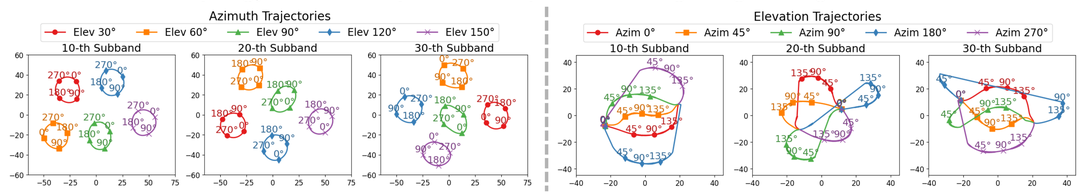
Performance comparisons across models and structural variations of SoundCompass

- Demonstrating **effective and efficient DoA-based TSE framework** and importance of each component

Model gradient clipping norm (gc-norm): 5	SNR Metrics ↑		Spatial Errors ↓			Complexities ↓	
	SNRi (dB)	SI-SNRi (dB)	ΔILD (dB)	ΔIPD (rad)	ΔITD (us)	Param.	Mult-Adds
DeepASA [25]	15.636	12.976	0.261	0.896	44.829	5.46 M	74.85 G
SSDQ (w. TPD) [12]	5.949	-1.171	-	-	-	3.91 M	21.22 G
DSENet (w. cyc-pos (θ, ϕ)) [18]	16.419	16.025	-	-	-	4.88 M	86.89 G
Proposed (Base)	17.865	16.717	0.099	0.805	10.302	2.70 M	20.49 G
remove an interaction in SPIN	17.171	15.770	0.124	0.812	15.777	2.59 M	20.49 G
** gc-norm: 20 (above), 5 (this row)	5.663	15.854	0.115	0.821	11.765	2.59 M	20.49 G
replace SH to cyc-pos (θ, ϕ)	17.696	16.538	0.100	**0.782**	12.747	2.70 M	20.49 G
remove a band-split structure	17.524	16.238	0.104	0.808	14.513	2.16 M	20.49 G
add an SED decoder	17.884	16.780	0.098	0.800	9.993	4.09 M	23.46 G
refine iteratively (×2)	**18.196**	**17.079**	**0.093**	0.789	**9.714**	+ 3.48 M	+ 24.01 G

t-SNE trajectories of the FiLM scale (γ) parameters

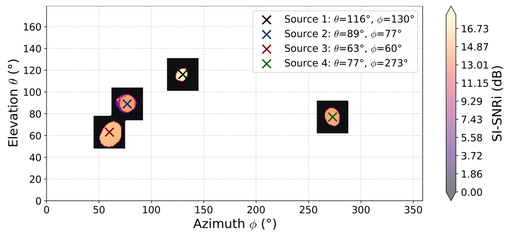
- **Continuous topology of subband-specific latent patterns for target angle change**



SI-SNRi change with respect to the clue direction

- **Robust discrimination** of spatially and temporally overlapping sources in complex acoustic scenes

Extraction in a cuboid room of size (X, Y, Z) = (5.57, 5.20, 3.79) m with an RT60 of 0.32 s



Spectrogram example