NEURAL INFORMATION PROCESSING SYSTEMS

KAIST EE

SMART SOUND LAB

DeepASA

An Object-Oriented Multi-Purpose Network for Auditory Scene Analysis

Dongheon Lee, Younghoo Kwon, Jung-Woo Choi†
{dongheon0115, k0hoo, jwoo}@kaist.ac.kr

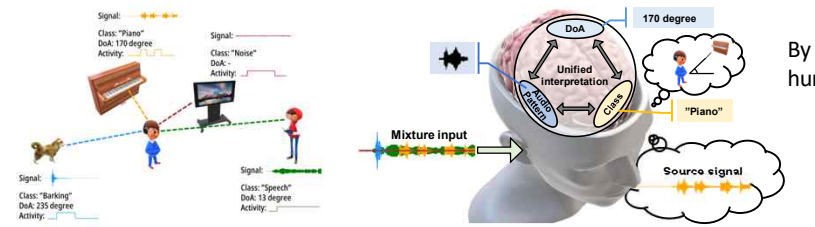
Hugging Face | NeurIPS | Paper | ASA2 dataset

Highlights

Universal model performing three auditory tasks
(Separation, Event Detection, Direction of Arrival Estimation)

- Result: Achieving the SOTA across all tasks through complementary cue integration
- **OOP**: **Object-Oriented Processing** for analyzing Object Features
- **CoI**: **Iterative reciprocal reasoning** for the fusion and refinement of estimated parameters (SED ↔ DoA)

Auditory Scene Analysis (ASA)



By analyzing **spatial**, **temporal**, and **spectral relations**, human can conduct various auditory scene analysis tasks

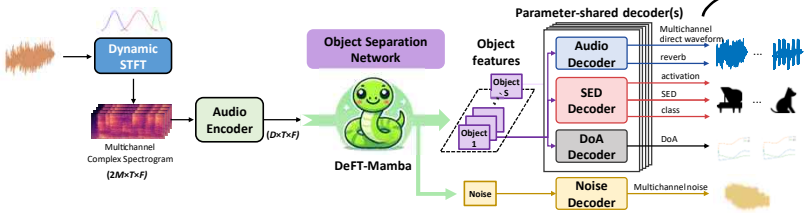
1. Direction-of-Arrival Estimation (DoAE)
2. Sound Event Detection (SED)
3. Universal Source Separation (USS)

Architecture of DeepASA

Object-Oriented Processing (OOP) with parameter-shared Multihead Decoders

- **DeFT-Mamba**: Object feature separation through Transformer + Mamba architecture
- **Multihead decoder**: Simultaneous estimation of audio signals, SED and DoA parameters

✓**Unified object feature** → Multi-task synergy & less permutation ambiguity

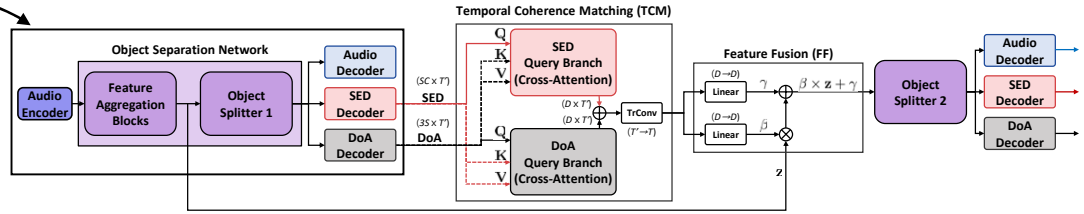


Chain-of-Inference (CoI)

Chain-of-Inference (CoI): Iterative reasoning for robust feature refinement

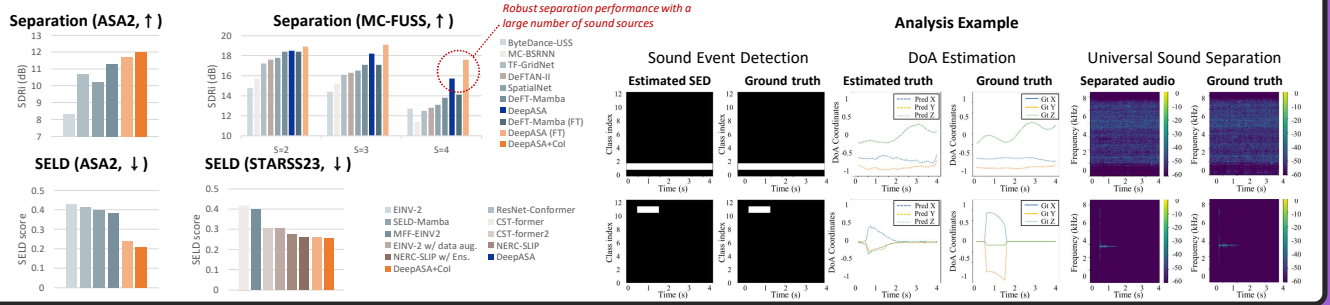
- **Temporal Coherence Matching (TCM)**: Early estimation refinement via temporal consistency between SED and DoA
- **Feature Fusion (FF)**: Re-injecting fused auditory clues into object representations

✓**Iterative feature refinement** → Positive feedback loop, simultaneously enhancing all downstream tasks



Comparison with SOTA models on Various Tasks

ASA2: new dataset for USS & SELD (**Event Detection, Localization & Classification**) on the moving sound source



More fun things in the paper!

1. **External Benchmarks**
   SOTA performance on the STARSS23 (SELD) and MC-FUSS (separation)
2. **Real-world Robustness**
   Robust performance in real office environments with unseen RIRs
3. **Dynamic STFT**
   Time-varying learnable window captures fast transient events (e.g. knock)
4. **Noise Decoder**
   Explicit noise separation boosts 'domestic sound' classification
5. **Anechoic / Reverb Separation**
   Separating Direct/Reverb Sharpens DoAE