DCASE2022 Workshop

Sound Event Localization and Detection for Real Spatial Scenes: Event-Independent Network and Data Augmentation Chains

Presenter: *Wenwu Wang*[5]

Authors: *Jinbo Hu*[1,2], *Yin Cao*[3], *Ming Wu*[1], *Qiuqiang Kong*[4], *Feiran Yang*[1], *Mark D. Plumbley*[5], *Jun Yang*[1,2]

[1]Key Laboratory of Noise and Vibration Research, Institute of Acoustics, Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
[3]Xi'an Jiaotong Liverpool University, Suzhou, China
[4]ByteDance Shanghai, China
[5]Centre for Vision, Speech and Signal Processing, University of Surrey, UK

Institute of Acoustics Chinese Academy of Sciences

中国科学院大学 University of Chinese Academy of Sciences

Xi'an Jiaotong-Liverpool University 西交利物浦大学

ByteDance 字节跳动

UNIVERSITY OF SURREY

1. Introduction

Background

- Sound event localization and detection (SELD) contains sound event detection (SED) and direction-of-arrival (DoA) estimation.
- The dataset transforms from computationally generated spatial recordings to real-sound scene recordings in 2022.

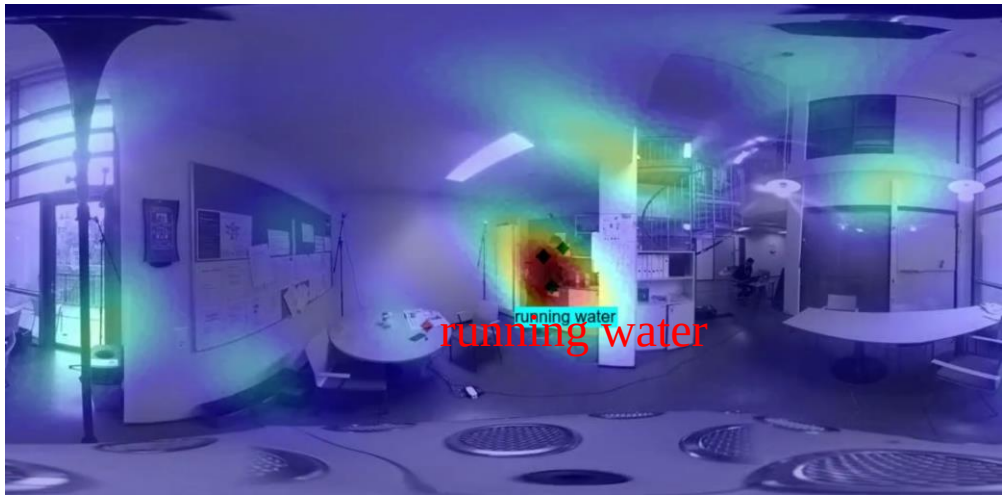

Fig. An instance of annotated real spatial sound scenes, provided by DCASE

Summary

- Our system is based on Event-Independent Network V2 (EINV2) with data augmentation chains.
- We generate simulated data by randomly convolving chosen samples of sound events with measured SRIRs.

2. The method

Event-Independent Network V2

- EINV2 uses three tracks to address up to three overlapped sound events.
- Multi-head self-attention blocks are replaced by Conformer blocks.

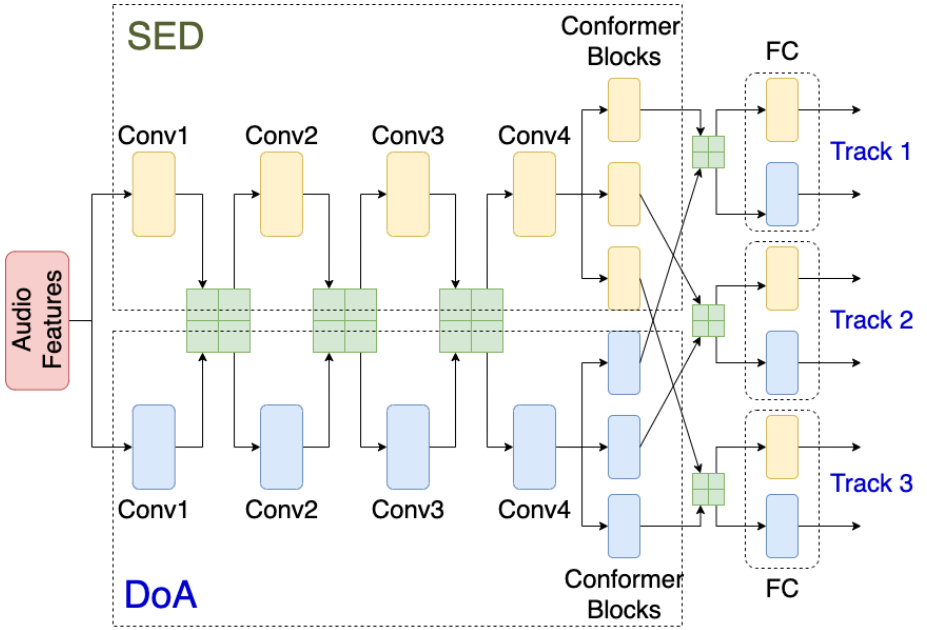

Fig. The architecture of **EINV2**

Data augmentation chains

- Combined by augmentation operations, which are randomly selected and linked in a chain.
- Augmentation operations include Mixup, Cutout, SpecAugment, and frequency shifting.
- Rotation of First-order Ambisonics (FOA) signals is an additional augmentation method.

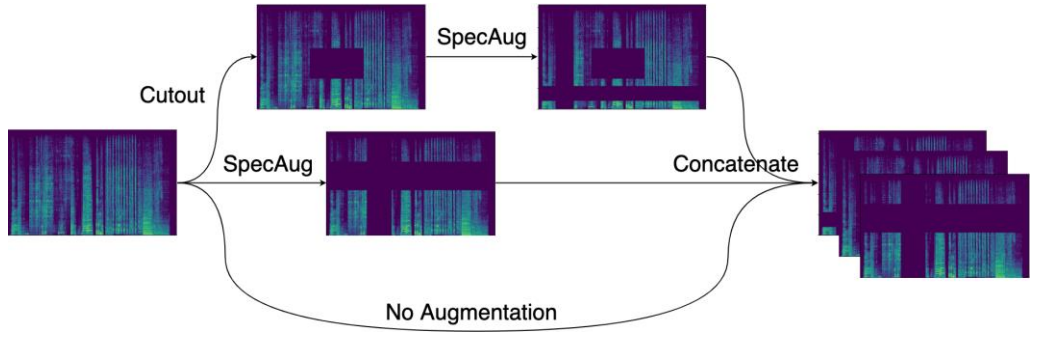

Fig. Diagram of **data augmentation chains**

2. The method

Simulated Data

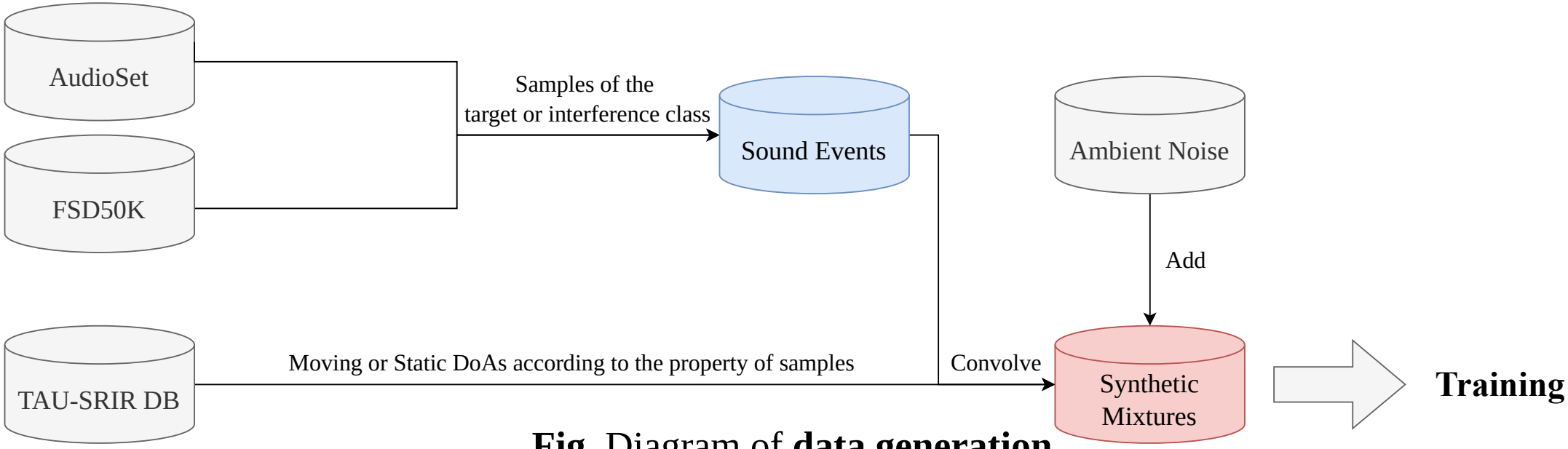

Fig. Diagram of **data generation**

- Samples of sound events are based on affinity of the labels in that dataset to the target/inference classes.
- The maximum polyphony of target classes is 3, excluding additional polyphony of interference classes.

3. Experiments

SELD performance of our submitted systems

- EINV2 with data augmentation chains performs better.
- The results demonstrate the effectiveness of our simulated data over the official dataset.

Tab. The metric scores on validation set and evaluation set

System	Datasets	Validation set				Evaluation (Blind test) set			
		ER_{20°	F_{20°	LE_{CD}	LR_{CD}	ER_{20°	F_{20°	LE_{CD}	LR_{CD}
Baseline FOA [5]	Official	0.71	21.0%	29.3°	46.0%	0.61	23.7%	22.9°	51.4%
EINV2 w/o dataAug chains	Official	0.75	32.3%	24.0°	56.1%	-	-	-	-
EINV2 w/ dataAug chains	Official	**0.56**	**42.4%**	**19.3°**	**61.4%**	-	-	-	-
System #1	A+B+C	**0.50**	48.4%	19.5°	65.7%	0.44	49.2%	16.6°	70.4%
System #2	A+B	**0.50**	**51.0%**	**16.4°**	**65.9%**	0.40	**57.4%**	**15.1°**	70.6%
System #3	A	0.53	48.1%	17.8°	62.6%	**0.39**	55.8%	16.2°	**72.4%**
System #4	B	0.53	45.4%	17.4°	62.5%	0.40	50.9%	15.9°	69.4%

3. Experiments

Class- and room-wise metric scores

- The highly skewed class-wise performance.
- Poor generalization ability to different rooms.

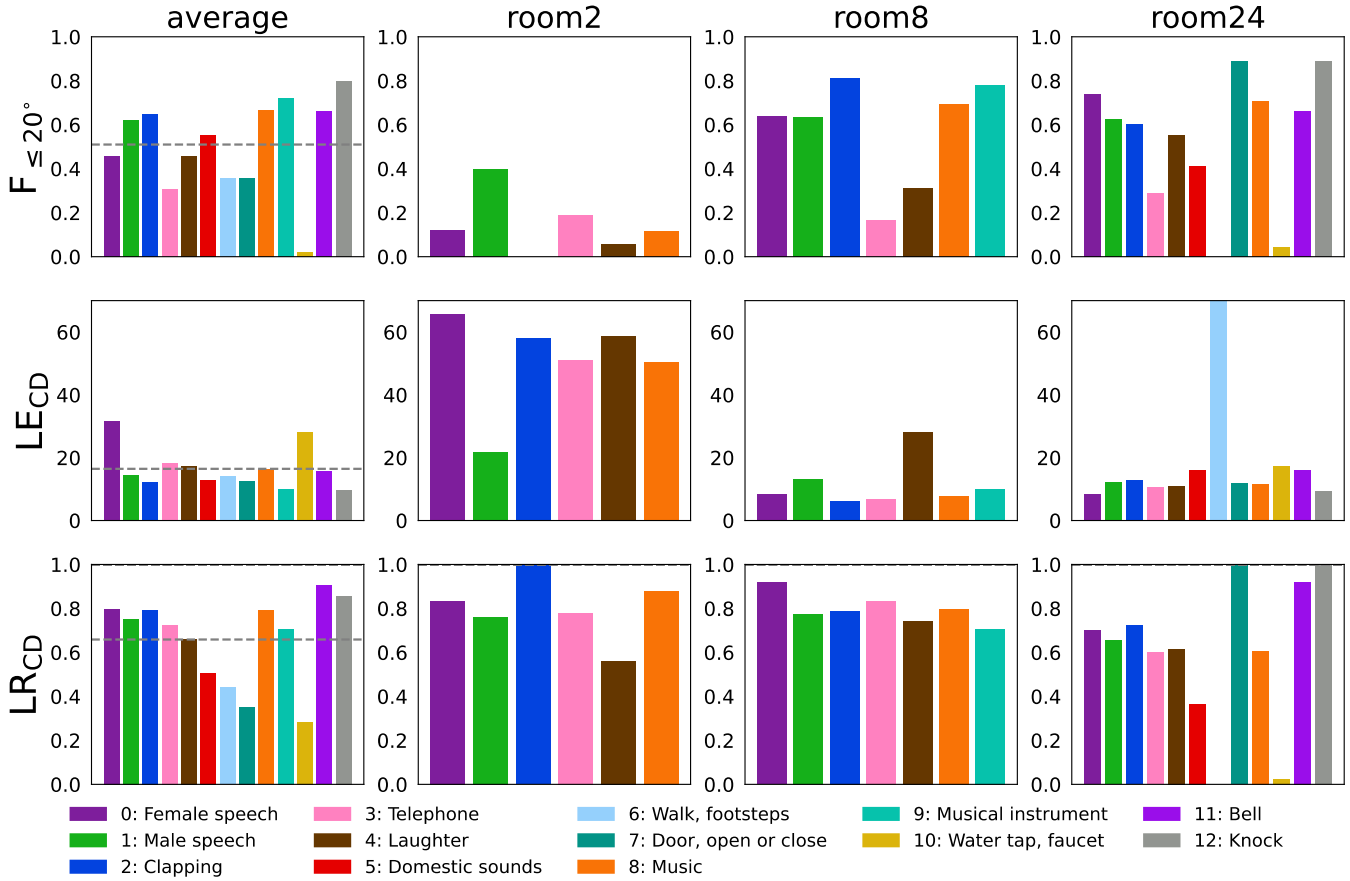

Fig. Metric scores of System #2 on validation set of STARSS22 in detail.

4. Conclusions

- This paper presents an approach using an Event-Independent Network V2 with a novel data augmentation method for real life sound event localization and detection.
- For this challenge, we synthesized more training samples which are convolved using sound events from FSD50k and AudioSet with collected room impulse responses from TAU-SRIR DB.
- Results indicate that the ability to generalize to different environments and unbalanced performance among different classes are two main challenges.
- Proposed method outperforms DCASE 2022 Task3 baseline model and ranked second.