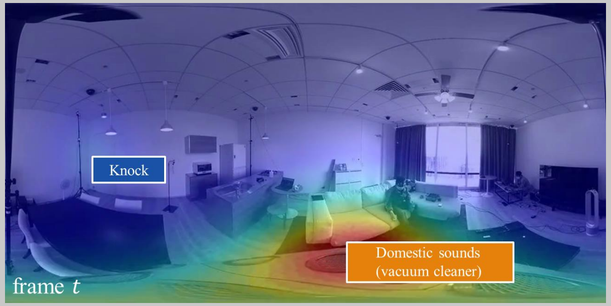
STARSS23: An Audio-Visual Dataset of Spatial Recordings of Real Scenes with Spatiotemporal Annotations of Sound Events

Kazuki Shimada, Archontis Politis, Parthasaarathy Sudarsanam,
Daniel Krause, Kengo Uchida, Sharath Adavanne, Aapo Hakala, Yuichiro Koyama,
Naoya Takahashi, Shusuke Takahashi, Tuomas Virtanen, Yuki Mitsufuji

Sony and Tampere University



Demo Video

- STARSS23 (Sony-TAU Real Spatial Soundscapes 2023)
- Real sound scene with multichannel audio, 360 video, and spatiotemporal annotation

STARSS23 Dataset

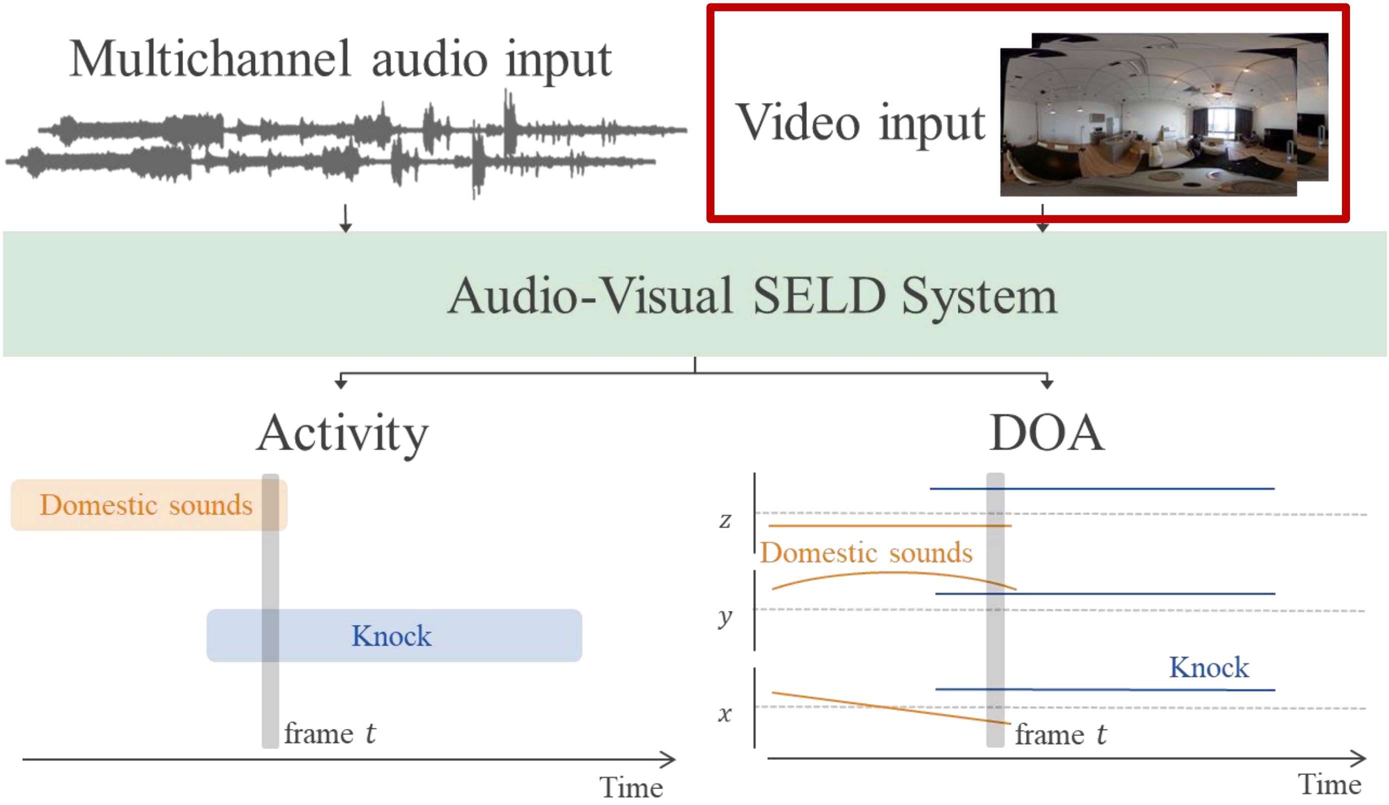
- What is STARSS23? - A unique dataset with multichannel audio, video, and spatiotemporal labels of sound events.

Dataset	Multichannel audio	Video	Label
STARSS23 (Ours)	OK	OK	Activity, Direction of Arrival (DOA), and Distance of Sound Events
VGG-SS [Chen+ '21]	-	OK	Bounding Box
YouTube-360 [Morgado+ '20]	OK	OK	-
AVRI [Qian+ '22]	OK	OK	Activity and DOA of Speech

- Why STARSS23? – The dataset enables spatial audio-visual tasks such as audio-visual sound event localization and detection (SELD). We can use spatial audio and aligned visual data with fine-grained annotation.
- Statistics? – 7.3 hours, 16 rooms, 57 participants, and 168 clips with full annotation.

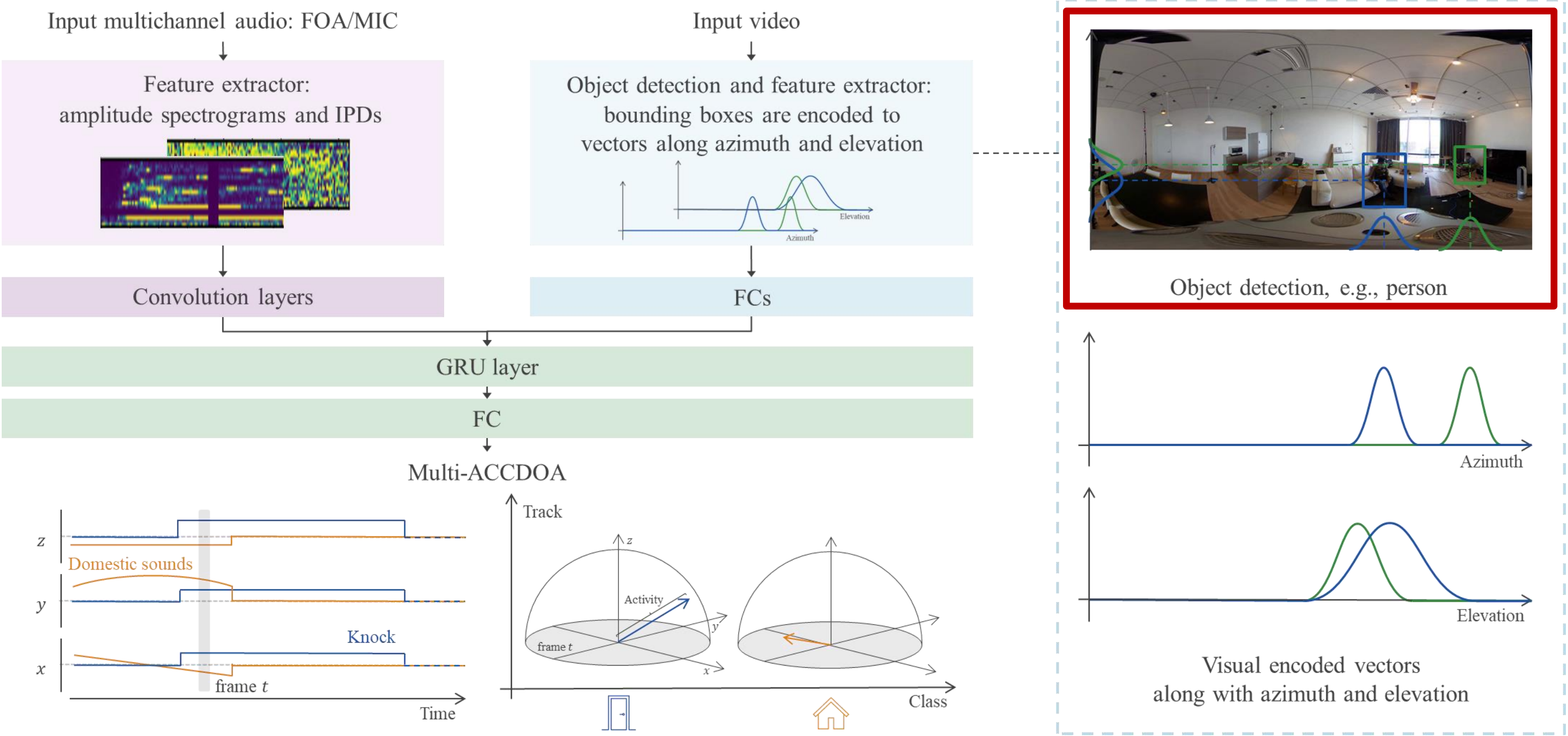
Audio-Visual Sound Event Localization and Detection (SELD)

- Sound events usually derive from visually perceptible source objects, e.g., footsteps from a walker.
- Using multichannel audio and video, audio-visual SELD system estimates class, activity, and direction of sound events.



Benchmark on Audio-Visual SELD

- Audio-visual SELD system uses convolutional recurrent neural network (CRNN).
- While audio inputs are spectrograms, visual inputs are bounding boxes of a person.
- Audio and visual embeddings are concatenated in the middle of the network.



Input multichannel audio: FOA/MIC

Feature extractor: amplitude spectrograms and IPDs

Convolution layers

Input video

Object detection and feature extractor: bounding boxes are encoded to vectors along azimuth and elevation

FCs

GRU layer

FC

Multi-ACCDOA

Object detection, e.g., person

Visual encoded vectors along with azimuth and elevation

Benchmark Result on Audio-Visual SELD

- Since the visual inputs are bounding boxes of a person, we especially focus on five classes related to a human body, i.e., female and male speeches, clapping, laughing, and footsteps.
- Audio-visual SELD system performs better in the body-related classes than audio-only one.
- If visual information matches the target classes of audio-visual SELD system, the audio-visual system can perform better.

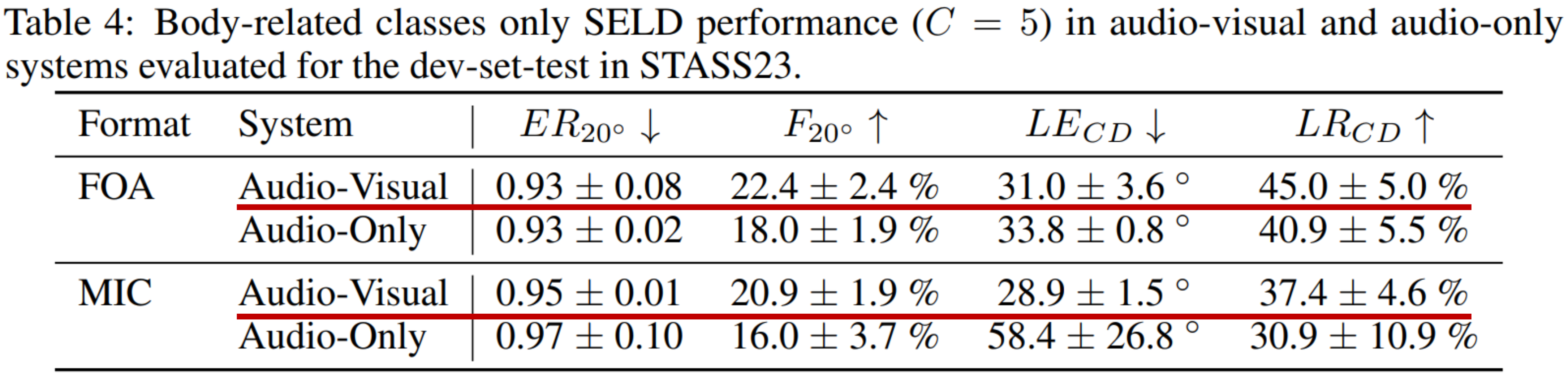


Table 4: Body-related classes only SELD performance ($C = 5$) in audio-visual and audio-only systems evaluated for the dev-set-test in STASS23.

Format	System	$ER_{20°} \downarrow$	$F_{20°} \uparrow$	$LE_{CD} \downarrow$	$LR_{CD} \uparrow$
FOA	Audio-Visual	0.93 ± 0.08	22.4 ± 2.4 %	31.0 ± 3.6 °	45.0 ± 5.0 %
	Audio-Only	0.93 ± 0.02	18.0 ± 1.9 %	33.8 ± 0.8 °	40.9 ± 5.5 %
MIC	Audio-Visual	0.95 ± 0.01	20.9 ± 1.9 %	28.9 ± 1.5 °	37.4 ± 4.6 %
	Audio-Only	0.97 ± 0.10	16.0 ± 3.7 %	58.4 ± 26.8 °	30.9 ± 10.9 %

Conclusion and Future Work

- STARSS23 consists of multichannel audio, 360 video, and spatiotemporal annotation.
- The dataset enables spatial audio-visual tasks, as shown in benchmark on audio-visual SELD.
- Both dataset and benchmark code are open under MIT license.

- Improve methods using audio-visual correspondence
- Tackle other spatial audio-visual tasks, e.g., spatial audio-visual generation tasks