Toward an Audio-Visual Dataset of Spatial Recordings of Real Scenes with Spatiotemporal Annotations of Sound Events

Kazuki Shimada[1], Archontis Politis[2], Parthasaarathy Sudarsanam[2], Daniel Krause[2]
Kengo Uchida[1], Sharath Adavanne[2], Aapo Hakala[2], Yuichiro Koyama[3]
Naoya Takahashi[1], Shusuke Takahashi[3], Tuomas Virtanen[2], Yuki Mitsufuji[1,3]
[1] Sony Research, Tokyo, Japan
[2] Audio Research Group, Tampere University, Tampere, Finland
[3] Sony Group Corporation, Tokyo, Japan

1. Introduction

Given multichannel audio input from a microphone array, a sound event localization and detection (SELD) system outputs a temporal activation track for each of the target sound classes along with one or more corresponding spatial trajectories, e.g., direction of arrivals (DOAs) around the microphone array, when the track indicates activity. Such a spatiotemporal characterization of sound scenes can be used in a wide range of machine cognition tasks, such as inference on the type of environment, tracking of specific types of sound sources, smart-home applications, scene visualization systems, and acoustic monitoring. The Sony-TAu Realistic Spatial Soundscapes 2022 dataset (STARSS22) [9] enables us to evaluate SELD systems in real sound-scene recordings with hand-labeled annotations. Currently STARSS22 is the only SELD dataset with real recordings, spatiotemporal annotations of multiple sound event classes, including moving source events. While this dataset is suitable for evaluating SELD systems with audio-only input, the dataset does not include video data.

As audio-visual data improves sound source separation quality [13, 2], audio-visual data in SELD tasks also have the potential to mitigate difficulties and ambiguities of the spatiotemporal characterization of the acoustic scene. For example, using video data, sounds of footsteps can be easily distinguished from other tapping sounds. Visible speakers in the video can also provide candidate positions of speaker-related sounds.

Several studies used audio-visual data for DOA estimation [10]. Qian *et al.* evaluated their audio-visual speaker DOA estimation system with Audio-Visual Robotic Interface (AVRI) dataset [10]. While the AVRI dataset helps evaluate audio-visual DOA estimation tasks, the evaluation focuses only on speech, not various sound events such as music, footsteps, and knocks.

There is broad interest in audio-visual spatial self-supervised learning [6] and audio-visual sound source localization [11, 7]. Morgado *et al.* proposed self-supervised learning with a spatial alignment method and evaluated it with the YouTube-360 dataset, which consists of first-order Ambisonic (FOA) signal and 360-degree video data from YouTube [6]. Senocak *et al.* evaluated their proposed audio-visual sound source localization method on the Flickr-SoundNet dataset, which consists of sound and image pairs [11]. These datasets are adequate to train neural networks (NNs) with audio-visual correspondence [6, 11] or evaluate localization performance in a visual scene [11]. However, there are no audio-visual datasets with hand-labeled DOA annotations of sound events around a microphone array.

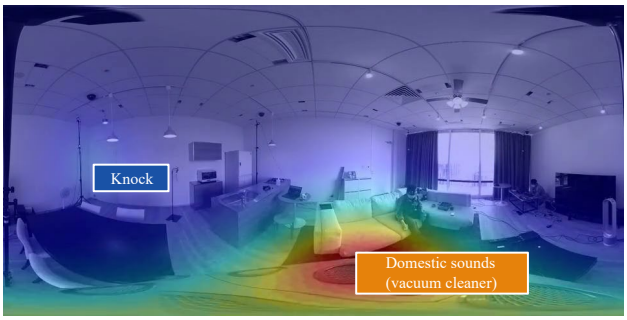

Figure 1. Still from overlay of 360° video, spatial acoustic power map generated from microphone array, and annotated event label.

We introduce a new audio-visual dataset, the Sony-TAu Realistic Spatial Soundscapes 2023 (STARSS23), consisting of multichannel audio, video, and spatiotemporal data annotation of sound events, as shown in Figure 1. The dataset enables us to evaluate SELD systems with audio-visual input. We developed and tested a simple audio-visual SELD system with STARSS23, which achieved lower localization error than an audio-only SELD system.

2. Related work

STARSS22 [9] has tackled real spatial recording of sound scenes with temporal and spatial annotation of sound events. STARSS22 was created at two different sites, Tampere, Finland and Tokyo, Japan, both using a similar recording and annotation procedure. The multichannel audio

data are delivered as two spatial recording formats: FOA and four-channel directional microphone recordings from a tetrahedral array configuration (MIC). There are 13 classes of target sound events, such as speech, footsteps, and knock. The dataset includes temporal activation and DOA labels of each target sound event around the microphone array. The dataset has real spatial recordings to evaluate SELD systems with audio-only input, but it does not release video data. Therefore, it cannot be used to evaluate SELD systems with audio-visual input.

There have been several studies on DOA estimation with audio-visual data [10]. Qian *et al.* proposed an audio-visual DOA estimation system, which processes log-mel spectrogram and the generalized cross correlation with phase transform (GCC-PHAT) features from audio input and face-bounding boxes from video input [10]. The face-bounding boxes are transformed into a concatenation of two Gaussian-like vectors. They represent the likelihoods of objects present along the image's horizontal and vertical axes to help DOA estimation. These features are integrated using attention mechanisms. The system then outputs the estimated DOA of speakers. The system was evaluated with the AVRI dataset, recorded using Kinect and a four-channel Respeaker array, along with annotation of location and voice activity detection. The audio and visual features are helpful for DOA estimation, and the dataset supports the evaluation of audio-visual speaker DOA estimation. However, the dataset is only for speech, not various sound events such as musical instruments, clapping, and doors.

There is broad interest in audio-visual spatial correspondence or localization, e.g., audio-visual spatial self-supervised learning [6] and sound source localization tasks [11, 7]. Morgado *et al.* proposed a self-supervised learning method with audio-visual spatial alignment, which uses FOA signal data and 360-degree video data in the YouTube-360 dataset [6]. The learned representation is evaluated for several downstream tasks, such as semantic segmentation and action recognition. Senocak *et al.* proposed a learning method to localize sound sources in a visual scene in an unsupervised and supervised manner [11]. The method was evaluated on the Flicker-SoundNet dataset, which consists of pairs of a sound and an image, along with bounding-box annotation of sound sources. While these methods and datasets are effective in learning the correspondence between audio and visual data semantically, they cannot be straightforwardly applied to SELD tasks, and the datasets do not have DOA labels for each target sound event around a microphone array.

3. Dataset

STARSS23 contains multichannel audio and video recordings of sound scenes in various rooms and environments, together with temporal and spatial annotations of prominent events belonging to a set of target classes. The dataset is available in Zenodo[1], and there is a demo video[2]. STARSS23 is an improvement on a multichannel audio dataset, i.e., STARSS22 [9]. One of the critical differences is releasing both audio and video data. We also increase the total number of recordings and add source distance information as additional annotations. We summarize the essential points as an audio-visual sound scene dataset while the details on the recording and annotation procedure can be found in the paper on STARSS22 [9].

As mentioned above, STARSS23 was created at two different sites, in Tampere and Tokyo. Recordings at both sites shared a similar process, organized in sessions corresponding to distinct rooms, participants, and sound-making props. In each session, various clips were recorded with combinations of that session's participants acting out simple scenes and interacting among themselves and with the sound props. The scenes were based on generic instructions on the desired sound events. The instructions were a rough guide to ensure adequate event activity and occurrences of the target sound classes in a clip. Participants improvise according to the instructions.

Each scene was captured with temporally aligned audio-visual sensors, i.e., a high-resolution 32-channel spherical microphone array (Eigenmike em32[3]), with a height set at 1.5 m and a 360-degree camera (Ricoh Theta V[4]) mounted about 10 cm above the spherical microphone array. For each recording session, a suitable position of the Eigenmike and Ricoh Theta V was determined to cover the scene from a central place while considering the intended scenarios and specific room constraints. The recordings were converted to two 4-channel spatial formats: FOA and MIC, converted from the original 32-channel recordings. After blurring visible faces, the simultaneous 360-degree videos were made available with the participant's consent. The video format is equirectangular, and the resolution is 1920×960. The video frames per second are 29.97.

The spatiotemporal annotations of STARSS23 consist of temporal activation, DOA, and source distance labels of the target sound event classes. The DOA and distance labels are based on tracking results of a motion capture system. The target sound events are the same 13 classes as in STARSS22. The classes are chosen to conform to the Audioset ontology [4], and are: ***female speech, male speech, clapping, telephone, laughter, domestic sounds, footsteps, door, music, musical instrument, water tap, bell, knock.***. The recordings contain natural background noise and directional interference sounds such as computer keyboard or shuffling cards.

[1] https://zenodo.org/record/7709052
[2] https://www.youtube.com/watch?v=ZtL-8wBYPow
[3] https://mhacoustics.com/products#eigenmike1
[4] https://theta360.com/en/about/theta/v.html

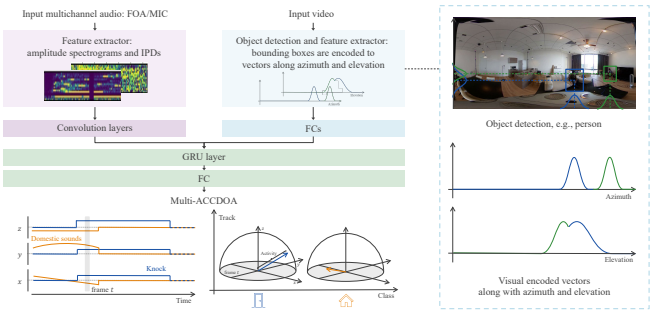

Figure 2. Audio-visual SELD baseline system uses a CRNN model with a Multi-ACCDOA output. From the input video frame, the system extract bounding boxes of objects. The bounding boxes are transformed to two vectors along azimuth and elevation.

Compared with STARSS22, STARSS23 maintains all the sessions of STARSS22 while adding four hours of material captured in Tampere. The dataset is split into a development set and evaluation set. The development set totals about 7 hrs 22 mins, of which 70 recording clips amounting to about 2 hrs were recorded in 4 rooms in Tokyo, and 98 recordings amounting to about 5 hrs were recorded in 12 rooms in Tampere. The development set is further split into a training part (dev-set-train, 40+50 clips in 2+7 rooms in Tokyo+Tampere) and testing part (dev-set-test, 30+48 clips in 2+5 rooms in Tokyo+Tampere) to support the development process. The evaluation set is about 3 hrs 33 mins, recorded in 2 rooms in Tokyo (35 clips) and 5 rooms in Tampere (44 clips). As the evaluation set is prepared for a competition, there are no publicly available annotations.

4. Baseline

Our baseline system for the audio-visual SELD task has a similar architecture to an audio-only SELD system based on SELDnet [1] and multiple activity-coupled Cartesian direction of arrival (Multi-ACCDOA) [12].

Before discussing our audio-visual SELD baseline system, we give a summary of the audio-only SELD system. The input is multichannel audio, from which the different acoustic features are extracted. This system takes a sequence of consecutive feature frames. It then predicts all the active sound event classes for each input frame and their respective spatial location, producing the temporal activity and DOA trajectory for each sound event class. In particular, a convolutional recurrent neural network (CRNN) is used to map the frame sequence to a Multi-ACCDOA sequence output, which encodes both sound event detection and DOA estimates in the continuous 3D space as a multi-output regression task. The CNN part encodes the audio features to audio embedding, then the RNN and fully connected layers (FCs) decode the embedding sequence to the Multi-ACCDOA sequence. Each sound event class in the Multi-ACCDOA output is represented by three regressors that estimate the Cartesian coordinates x, y, and z of the DOA around the microphone array. If the vector length represented by x, y, and z coordinates is greater than a threshold, the sound event is considered active. The corresponding x, y, and z values are their predicted DOAs.

While the audio-only system takes only audio input, our audio-visual SELD baseline system takes both audio and visual input. Visual input is a corresponding image at the start frame of the audio feature sequence. An object detection module (e.g.,YOLOX [3]) outputs bounding boxes of potential objects with the corresponding image. These bounding boxes are transformed to a concatenation of two Gaussian-like vectors, representing the likelihoods of objects present along the image's horizontal axis and vertical axis [10]. The Gaussian-like vectors are encoded to a visual embedding by FCs. The visual embedding and the audio embedding from the audio encoder are then concatenated. The concatenated feature sequence is fed into the decoder to output a Multi-ACCDOA sequence. Figure. 2 shows the architecture of our audio-visual SELD baseline system.

Here is the experimental settings. As audio features, multichannel amplitude spectrograms and inter-channel phase differences (IPDs) are used [12]. The short-term Fourier transform (STFT) is applied with a 20-ms frame length and 10-ms frame hop. Input features are segmented to have a fixed length of 1.27 sec. The shift length is set to 1.2 sec during inference. For a CRNN model, we stack three convolutional layers with kernel size 3×3 and a bidirectional GRU layer with hidden state size 256. The number of tracks in the Multi-ACCDOA format was fixed at $N = 3$ maximum simultaneous sources. We used 16 batch size and the Adam optimizer with a weight decay of 10^{-6}. The learning rate is set to 0.001. We validated and saved model weights in every 1,000 iterations to 20,000 iterations. The threshold for activity was 0.3 to binarize predictions during inference. The code is available in a GitHub repository[5].

5. Evaluation

We used four joint localization and detection metrics [5] with extensions from a previous study [8] as the SELD evaluation metrics to support multi-instance scoring of the same class. Two metrics are referred to as location-aware detection, and are error rate (ER_{20°) and F-score (F_{20°) in one-sec non-overlapping segments. We consider the prediction correct if the prediction and reference class are the same and the distance between them is below 20°. The other two metrics are referred to as class-aware localization and are localization error (LE_{CD}) in degrees and localization recall (LR_{CD}) in 1-sec non-overlapping segments, where the subscript refers to classification-dependent. Unlike the location-aware detection, we do not use any distance thresh-

[5] https://github.com/sony/audio-visual-seld-dcase2023

Table 1. SELD performance of audio-visual and audio-only baseline systems evaluated for dev-set-test in STASS23.

Input	$ER_{20^\circ} \downarrow$	$F_{20^\circ} \uparrow$	$LE_{CD} \downarrow$	$LR_{CD} \uparrow$
FOA + Video	1.07	14.3 %	48.4°	35.5 %
FOA	1.00	14.4 %	60.4°	32.7 %
MIC + Video	1.08	9.8 %	62.4°	29.2 %
MIC	1.03	11.4 %	77.3°	30.4 %

old but estimate the distance between the correct prediction and reference. We used the macro mode of computation. We first computed the above metrics for each sound class then averaged them to obtain the final system performance.

For a comparison between our audio-visual baseline and audio-only systems, we also experiment the audio-only system based on the same training data, i.e., the dev-set-train of STARSS23, and the same implementation. The only difference is the presence or absence of video input. We experiment the systems with both FOA and MIC formats.

The evaluation metric scores for the testing part of the development set, i.e., dev-set-test, are listed in Table 1. Our audio-visual baseline system exhibited lower localization error than the audio-only system while keeping comparable localization recall in FOA and MIC formats.

6. Conclusion

We introduces an audio-visual dataset, Sony-TAu Realistic Spatial Soundscapes 2023 (STARSS23), which consists of multichannel audio data, video data, and spatiotemporal annotation of sound events. The dataset enables us to evaluate sound event localization and detection (SELD) systems with audio-visual input. We also develop a simple audio-visual SELD baseline system, which handle multichannel audio data and video data. After feature extraction, a convolutional recurrent neural network (CRNN) outputs each target sound event's activation and direction-of-arrival (DOA). We compare the audio-visual SELD baseline system with an audio-only system on STARSS23. The audio-visual system achieves lower localization error than the audio-only system. We plan to integrate audio-visual sound source localization methods with the audio-visual SELD system.

7. Acknowledgments

We thank Akira Takahashi for his helpful code review and thank Atsuo Hiroe, Kazuya Tateishi, Masato Hirano, Takashi Shibuya, Yuji Maeda, and Zhi Zhong for valuable discussions about the annotation process.

The data collection and annotation at Tampere University have been funded by Google. This work was carried out with the support of the Centre for Immersive Visual Technologies (CIVIT) research infrastructure at Tampere University, Finland.

References

[1] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen. Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. *IEEE J. Selected Topics in Sig. Proc.*, 13(1):34–48, 2018. 3

[2] R. Gao, R. Feris, and K. Grauman. Learning to separate object sounds by watching unlabeled video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 35–53, 2018. 1

[3] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun. YOLOX: Exceeding YOLO series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 3

[4] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *IEEE Int. Conf. on Acoustics, Speech and Sig. Proc. (ICASSP)*, pages 776–780, 2017. 2

[5] A. Mesaros, S. Adavanne, A. Politis, T. Heittola, and T. Virtanen. Joint measurement of localization and detection of sound events. In *IEEE Work. on Applications of Sig. Proc. to Audio and Acoustics (WASPAA)*, pages 333–337, 2019. 3

[6] P. Morgado, Y. Li, and N. Nvasconcelos. Learning representations from audio-visual spatial alignment. *Advances in Neural Information Processing Systems*, 33:4733–4744, 2020. 1, 2

[7] A. Owens and A. A. Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 631–648, 2018. 1, 2

[8] A. Politis, A. Mesaros, S. Adavanne, T. Heittola, and T. Virtanen. Overview and evaluation of sound event localization and detection in DCASE 2019. *IEEE/ACM Trans. Audio, Speech, and Language Proc.*, 2020. 3

[9] A. Politis, K. Shimada, P. Sudarsanam, S. Adavanne, D. Krause, Y. Koyama, N. Takahashi, S. Takahashi, Y. Mitsufuji, and T. Virtanen. STARSS22: A dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events. *arXiv preprint arXiv:2206.01948*, 2022. 1, 2

[10] X. Qian, Z. Wang, J. Wang, G. Guan, and H. Li. Audio-visual cross-attention network for robotic speaker tracking. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:550–562, 2022. 1, 2, 3

[11] A. Senocak, T.-H. Oh, J. Kim, M.-H. Yang, and I. S. Kweon. Learning to localize sound source in visual scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4358–4366, 2018. 1, 2

[12] K. Shimada, Y. Koyama, S. Takahashi, N. Takahashi, E. Tsunoo, and Y. Mitsufuji. Multi-accdoa: Localizing and detecting overlapping sounds from the same class with auxiliary duplicating permutation invariant training. In *IEEE Int. Conf. on Acoustics, Speech and Sig. Proc. (ICASSP)*, pages 316–320, 2022. 3

[13] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba. The sound of pixels. In *Proceedings of the European conference on computer vision (ECCV)*, pages 570–586, 2018. 1