Multi Event Localization by Audio-Visual Fusion with Omnidirectional Camera and Microphone Array

Wenru Zheng[1], Ryota Yoshihashi[1], Rei Kawakami[1], Ikuro Sato[1,2], and Asako Kanezaki[1]
[1]Tokyo Institute of Technology
[2]Denso IT Laboratory Inc
zheng.titech@gmail.com, yoshi@d-itlab.c.titech.ac.jp,
reikawa@sc.e.titech.ac.jp, isato@c.titech.ac.jp, kanezaki@c.titech.ac.jp

Abstract

Audio-visual fusion is a promising approach for identifying multiple events occurring simultaneously at different locations in the real world. Previous studies on audio-visual event localization (AVE) have been built on datasets that only have monaural or stereo channels in the audio; thus, it was hard to distinguish the direction of audio when different sounds are heard from multiple locations. In this paper, we develop a multi-event localization method using multi-channel audio and omnidirectional images. To take full advantage of the spatial correlation between the features in the two modalities, our method employs early fusion that can retain audio direction and background information in images. We also created a new dataset of multi-label events containing around 660 omnidirectional videos with multi-channel audio, which was used to showcase the effectiveness of the proposed method.

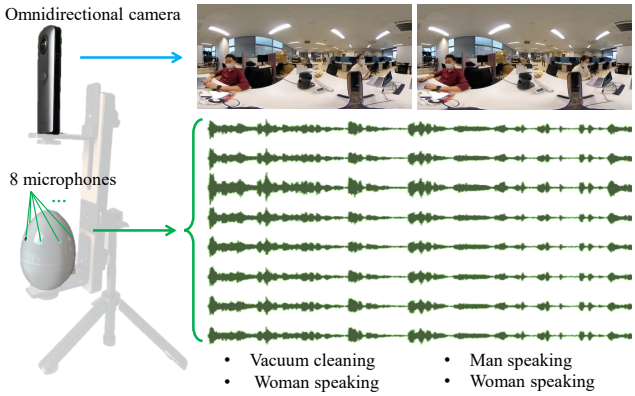


Figure 1. Illustration of our dataset that has visual input from omnidirectional camera Theta V (above) and 8-channel audio input captured by 360°-covered microphone array TAMAGO-03 (below). Multi-labels are given to each video frame. The camera and microphones are spatially calibrated.

1. Introduction

Sight and hearing play important roles in human perception of real-world events, as integration of vision and audio is carried out by quite an extensive part of the brain. In machine learning, ways of obtaining multi-modal representations for vision and audio have been extensively studied [2–4].

Among the audio-visual learning, audio-visual event (AVE) localization is a task to identify temporal segments with event labels from an audio-visual input. A major approach of AVE localization computes audio and visual features, which are later concatenated to classify event labels for each segment [19, 22, 29], or minimizes distances between segment features of the two modalities [13] so that they share good semantic representations. Most studies use an audio-guided visual attention mechanism (see [22]) that can capture rough sound-source location that serves as spatial attention for the visual input. However, since they use *late fusion* of the two modal features, the spatial information is likely partially lost at the time of concatenation of the two features. Also, strong spatial cues such as those in visual input are hard to obtain from datasets that contain only stereo or single-channel audio althoguh they are widely used [6]. The partial loss of spatial resolution can be problematic, especially when multiple sound sources are present in the input.

To overcome the aforementioned limitations in the AVE localization task, we develop an AVE localization method for datasets that contain multi-channel audio. We also propose a model based on *early multi-modal fusion* that can fully utilize spatial information not only from video but from multi-channel audio. To evaluate our model, we created a new multi-label multi-channel AVE (MLMC-AVE) dataset[1] that contains omnidirectional visual input and 8-

[1]The dataset is published on https://github.com/zwr17/

channel audio input captured by a microphone array as shown in Fig. 1. Unlike existing AVE datasets, where each video includes exactly a single event, our dataset contains multi-label videos, where a single frame can have multiple event labels. With this newly created dataset, we demonstrate that our model based on *early fusion* outperforms the existing method [22] based on *late fusion*, which is in fact contrary to the findings from [22], where single-label single-channel AVE dataset is used. The contributions of this study are listed as follows:

- We propose an AVE localization method that early fuses features from video and multi-channel audio so that rich spatial information from each modality can be effectively integrated.
- We present a multi-label multi-channel AVE (MLMC-AVE) dataset, where videos are captured by an omnidirectional camera and audio data are recorded by a 360°-covered 8-channel microphone array. The dataset contains frames, in which multiple event labels are given.
- We confirmed that our method consistently outperforms an off-the-shelf AVE localization method on the MLMC-AVE dataset.

2. Related Work

Audio-Visual Learning. Audio-visual learning (AVL) has been studied to overcome the limitations of recognition tasks based on a single modality by modeling the relationship between visual and audio information [30]. Audio-visual representation learning aims to find the pattern of representation of each modality automatically. Since the audio and visual information are synchronized in videos, the audio can be used for self-supervision. Aytar *et al.* [4] introduce a deep convolutional architecture to recognize natural sound by distilling the discriminative knowledge from the visual teacher network to audio student network. Considering the relationship between audio and visual information, some studies [2,3,11] propose networks that can learn better representation of each modality by recognizing whether or not the visual input is a counterpart corresponding to the audio input. Owens *et al.* [17] develop an early-fusion multimodal network by leaning a task to recognize whether the two modalities are synchronized or not. AVL studies are rather focused on unlabeled videos that are available on the Internet, and the use of spatial correlation between audio and visual information has not been extensively explored.

Audio-visual separation and localization. Audio-visual separation and localization are important tasks of AVL. The tasks isolate the sound source and specify its location by observing the sound and visual pairs in a video. Early work utilizes audio-visual synchrony [7], and early techniques in audio-visual separation can be found in a review by Rivet *et al.* [20]. Recent approaches mostly focus on unsupervised learning owing to the large amount of unlabeled videos on the Internet. Given a video as an input, Zhao *et al*. [28] showed that image regions can be identified and connected to each sound source by introducing a wisely considered pre-text task. The task is to separate sound sources that are extracted from two videos and artificially mixed. Senocak *et al.* presented a two-stream network with an attention mechanism to process each modality for sound source localization [21]. As the sources on the Internet typically have monaural or stereo audio, there have not been many studies on multi-channel audio. Although a few studies tackled audio-visual localization with multi-channel audio [1, 12], they did not consider event category estimation. Our focus is tied to the strength of multi-channel audio that has strong spatial correlation to the visual information.

Audio-visual event localization. Audio-visual event (AVE) localization aims to identify audio-visual event labels from an input video where labels are provided along a time axis. AVE localization methods typically have two modal streams, and the focus of research has been on how to attend from one modality to the other to obtain good representations and to have good synchronization between modalities [5, 23–27]; therefore, many architectures and attention mechanisms have been proposed. For instance, Tian *et al.* [22] proposed an audio-guided visual attention to incorporate spatial attention. Lin *et al.* [13] and Ramaswamy [18] both considered the attention between global and local features to have inter and intra-interactions between the modalities. Liu *et al.* [14] suggested a bi-directional model where both the forward and backward attention in the two modalities were considered. Xia and Zhao [24] introduced a multi-modal background-suppression mechanism to maintain semantically focused attention. Since the datasets in those methods only include monaural or stereo audio information, spatial attention in audio features and its strength have not been extensively studied.

Apart from event recognition, augmented reality is an application area where multi-channel-audio and omnidirectional-video processing is highlighted to support human perception. In active speaker localization, Jiang *et al.* [9] proposed an end-to-end method with images of 360-degree version and N-channel audio collected by a multi-channel microphone array. Inspired by it, this work pay attention to the way of extracting richer spatial information from audio inputs so that multiple events can be finely localized with visual cues.

MLMC−AVE.

3. Dataset

3.1. Overview of our dataset

Our capturing device is a combination of 8-channel microphone array, TAMAGO-03, and an omnidirectional camera, Theta V, as shown in Fig. 1. The TAMAGO is an inexpensive hand-held microphone array with 8 microphones. It can be connected to a computer via USB and be operated as a USB audio. Its sampling rate is 16 kHz, and its bit depth is 24 bit. Open-sourced robot audition software, HARK [16], can be used for sound localization, sound source separation, and speech recognition. Theta V is a camera that can record 360-degree still images or videos with high image quality. Videos were recorded with 1920 $\times$ 960 resolution and the frame rate was 16 frames per second. The microphones and the camera were calibrated so that they share the center of rotation and have aligned coordinate system.

We defined 12 event categories assuming an indoor office environment with multiple sound sources. The defined categories are *man speaking, woman speaking, walking, typing, kettle boiling, writing on board, alarming, opening the door, opening the drawer, coughing, printer working,* and *cleaner working*. There is also a category *nothing* for scenes without any event; thus, the total number of categories in our dataset is 13. 660 videos were collected, and the distribution of each category is shown in Fig. 2. Each video contains at least one AVE. Note the following bias in the data set due to the fact that the subjects were predominantly male: *man speaking* samples comprise 29% of the total while *woman speaking* samples comprise only 9%.

The comparison of our dataset and several existing datasets is shown in Table 1. Tian *et al.* [22] created the Audio-Visual Event (AVE) dataset, which is widely used for the AVE localization task. It consists of 4,143 videos collected from YouTube, encompassing 28 event types, temporally tagged with audio-visual event boundaries. The collection contains a wide range of audio-visual events from several domains, including human activities, animal behaviors, musical performances, and vehicle sounds (*e.g.*, *man speaking, woman speaking, dog barking*, etc.). It is a monaural dataset, and only one audio-visual event is marked in each video, even when several events occur.

The FAIR-Play [6] and REC-street [15] datasets used their own recording devices and have multiple audio channels. The FAIR-Play dataset [6] contains 1,871 10-second video clips captured with GoPro and a professional binaural microphone in a music room. The authors captured 20 volunteers playing various instruments such as cello, drum, and piano in solo, duet, and multi-player performances. However, the dataset is created to generate spatial sound, and thus the category labels do not exist. The REC-street dataset [15] utilized a Theta-V 360 camera with a coupled TA-1 spatial audio microphone to record outdoor scenes. It

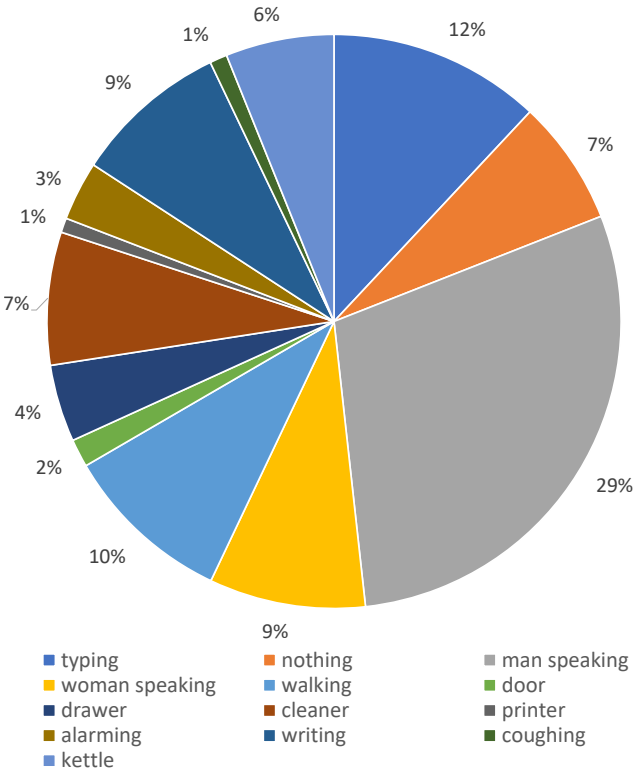


Figure 2. The distribution of each category in MLMC-AVE.

includes 43 videos that are 3.5 hours long in total. They also collected 360-degree videos with spatial audio using queries from YouTube which is referred to as YT-ALL. YT-MUSIC (397 videos) and YT-clean (496 videos) are the subsets of it that contain musical performances and super-imposed sources such as people talking in the meeting room, respectively. The dataset is for self-supervised generation of spatial audio; therefore, similar to the FAIR-Play dataset [6], the category labels are not provided in the dataset.

To summarize, our dataset is unique compared to the existing ones because the sound is recorded with multiple microphones and the frames contain multiple events that occur simultaneously, *i.e.*, each frame can have multiple labels whereas those in existing ones have no or single label.

3.2. Process for obtaining the data

Collection. To ensure variation in the dataset, we recorded videos in several different places with different people and items. The numbers of people, places and items that are used are presented in Table 2. Each video is 15 seconds in length. It took about 10 days to record the dataset.

Labeling. The labeling was done by one of the authors. The events that can be both seen and heard were thoroughly labeled. There were cases that were difficult to categorize as the dataset is a collection of real events. We list those

Table 1. Comparison of our dataset MLMC-AVE and existing audio-visual datasets.

Dataset	Video Type	#Channels	#Videos	#Categories	Category Example
AVE [22]	Normal	1	4,143	28	Man speaking, Dog barking
Fair-PLAY [6]	Normal	2	1,871	-	Not mentioned
REC-street [15]	Omnidirectional	4	43	-	Not mentioned
MLMC-AVE (Ours)	Omnidirectional	8	660	12	Man speaking, clapping

Table 2. Detailed information of MLMC-AVE.

Item	Value
#Label for each sample	$1 \sim 3$
Total video duration	660 (videos) $\times$ 15 (seconds)
The number of person	10
The number of places	6
Items	kettle, cleaner, printer, white board, laptop

difficult cases in the following.

- There were cases where the events were hard to hear (e.g., *writing, door,* etc.). As long as the event was audible to the labeler in the audio file, it was marked.
- When multiple events occurred simultaneously, there were cases where one of the sound was covered by the other and barely audible. In those cases, as long as the event was visible, it was labeled as an AVE.
- When the subject of the event was completely obscured by other things and was not visible in the video, it was not labeled as an AVE.
- Each segment is one-second long, but an event may occur for a short period of time that is less than a second. Those were still marked as AVEs in the dataset.

3.3. Challenge

Some examples of our dataset are shown in Fig. 3. Estimating multiple AVEs occurring simultaneously is difficult as shown in the figure, since some events can be only visible in a small part of the image, yet the difficulty can be mitigated by utilizing the audio-visual information. Complex background, small difference in human motion, noise from uncontrollable sources, and the randomness of the event occurrence make our dataset lively but challenging. For example, there may be multiple objects that can produce sounds, and their sizes may vary from as small as an alarm clock to as large as a printer. There can also be several people in the scene, each of whom can be the subject of an event. Many people wore masks in the scene, and yet the task requires classification of not speaking, coughing, and talking. The dataset includes some background sound conducted by air-conditioning or some machines outside the window that are not marked as AVEs.

4. Method

The goal of the AVE localization is to predict a label or possibly multiple labels of each segment in a video clip. We let T denote the number of segments in a video clip. The labels for the t-th segment ($t = 1, \cdots, T$) are given as $y_t = \left\{ y_t^k | y_t^k \in \{0, 1\} \right\}_{k=1}^{N+1}$ where N represents the number of the defined event categories. We added an extra element in the label set y_t to explicitly indicate that no event occurs. In our setting, each video segment may have zero, one, or multiple labels ($\sum_{k=1}^{N+1} y_t^k \geq 1$), whereas a segment in a common benchmark dataset (e.g., [22]) may have either zero or one label ($\sum_{k=1}^{N} y_t^k \leq 1$).

The overall architecture of the proposed model is shown in Fig. 4. Similar to the previous study [22], our architecture consists of three components: feature extraction, audio-spatial fusion, and classification. The details of each component are described below.

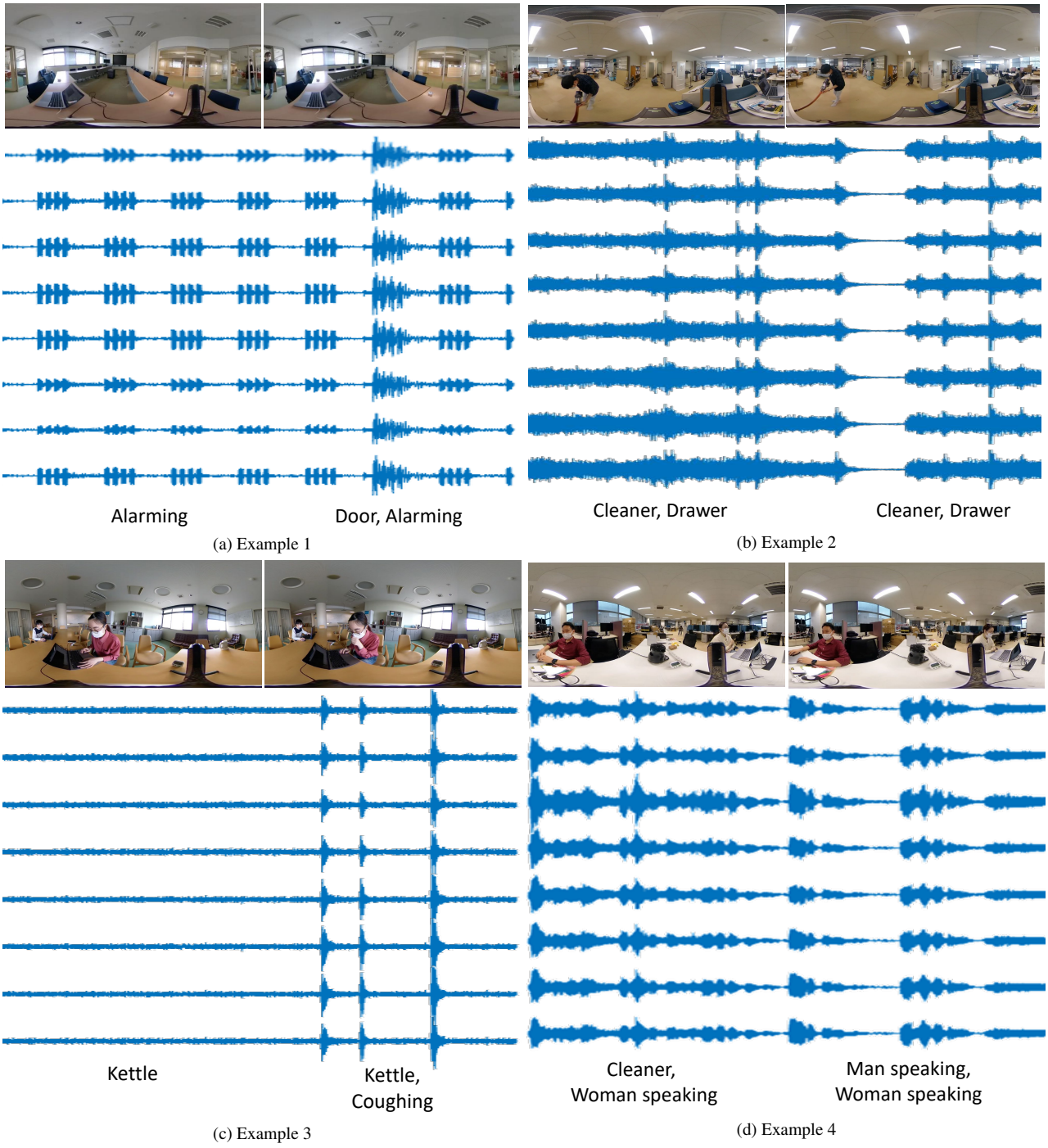
4.1. Feature extraction

The visual and audio inputs are individually pre-processed by pre-trained convolutional neural networks. To process the audio data, we first take the wav-format audio data from the video or the microphone(s) and compute the audio feature through the VGG-like CNN model pre-trained on the AudioSet dataset [8]. The extracted audio feature is denoted by $\boldsymbol{A} \in \mathbb{R}^{T \times d_{\mathrm{a}}}$, where d_{a} refers to the audio feature dimension.

To process the visual data, each frame image is extracted from the video and passed through ResNet pre-trained on ImageNet to obtain the frame-wise feature. The frame-wise feature vectors are then averaged within one second to yield the visual feature, which is denoted as $\boldsymbol{V} \in \mathbb{R}^{T \times H \times W \times d_{\mathrm{v}}}$, where H, W, and d_{v} denote the feature-map height, width, and the feature-channel dimension, respectively.

4.2. Audio-spatial fusion

After individually extracting visual and audio features, the next process is to extract the spatial correlation between visual and audio features by devising how to attend to the



(a) Example 1 (b) Example 2

(c) Example 3 (d) Example 4

Figure 3. Examples of omni-directional images, 8-channel sound waves, and category labels in our MLMC-AVE dataset.

features. This is especially important when multiple events occur simultaneously in different locations. As shown in Fig. 4, we adopt *early fusion*, where audio features from 8 channels are replicated and concatenated to visual features. In contrast, the baseline model [22] adopts *late fusion*, where explicit spatial information of visual features is lost at the fusion phase.

In addition to the aforementioned approach that uses a simple stack of audio features from multiple channels, we also investigate two alternative strategies to generate audio-visual attention.

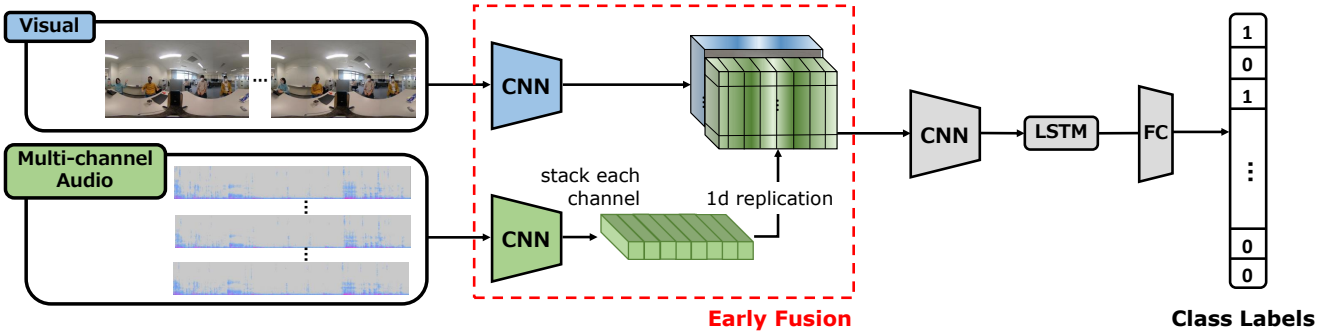


Figure 4. Overall architecture of the proposed model. To extract the spatial correlation between audio and visual features, we adopt *early fusion*, where audio features from 8 channels are replicated and concatenated to visual features.

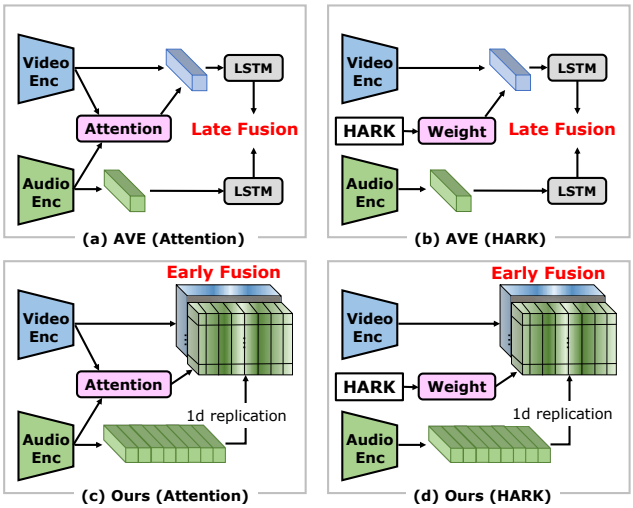


Figure 5. Summary of audio-spatial fusion methods: (a) late fusion with audio-guided visual attention [22], (b) late fusion with HARK, (c) early fusion with attention, and (d) early fusion with HARK.

Audio-Guided Visual Attention [22]. The first strategy is to follow a method of previous work [22], where the attention weight vector is generated by the audio-guided visual attention. The softmax attention weights of the audio-guided attention is computed from both visual and audio features. The spatial atetntion is only applied to visual features.

Multi-channel audio using an external sound arrival direction estimation method. The second strategy is to utilize a method of sound source separation and sound arrival direction estimation from multi-channel audio for attention generation with spatial dimension. In particular, we used an open-sourced robot audition software HARK [16] for the sound arrival direction estimation. Since the output from HARK is the predicted direction θ, we transferred it into the weight vector that gives a high probability for the range of $[\theta - 30°, \theta + 30°]$ to generate spatial attention map. The probability is modeled as a Gaussian, simulating a high center weight and continuous decrease of weight on the periphery. We set the mean as θ and the standard deviation as 16.0.

Summary of various fusion methods. The summary of audio-spatial fusion methods is provided in Fig. 5. In addition to our proposed model shown in Fig.4, which has a simplest architecture, we implemented those variants in Fig. 5 to test the two attention generation strategies combined with late and early fusion strategies, yielding four variations of architectures. They will be compared in detail in the experimental section.

4.3. Classification

In our new dataset, multiple events occasionally occur at the same time; thus, it is a multi-label classification problem. Here, we transform a conventional multi-label problem into a set of binary classification problems for different labels using the binary cross entropy loss:

$$l(\boldsymbol{x}, \boldsymbol{y}) = -\sum_{n=1}^{N+1} y_n \cdot \log x_n + (1 - y_n) \cdot \log (1 - x_n) \,. \quad (1)$$

where $\boldsymbol{x} = [x_1, \ldots, x_{N+1}]^\top$ and $\boldsymbol{y} = [y_1, \ldots, y_{N+1}]^\top$ are the output of the network and the target labels for a training sample, respectively. The inference time of multi-label AVE is about 0.35 seconds per frame.

5. Results

We used the AVE dataset [22] for the single-channel evaluation and our new MLMC-AVE dataset for the evaluation with multi-channel audio. The AVE dataset contains 4,143 10-second videos, each containing at least two seconds of at least one event of 28 different categories. The 4,143 videos include 3,339 training videos, 402 validation videos, and 402 testing videos. We used the classification

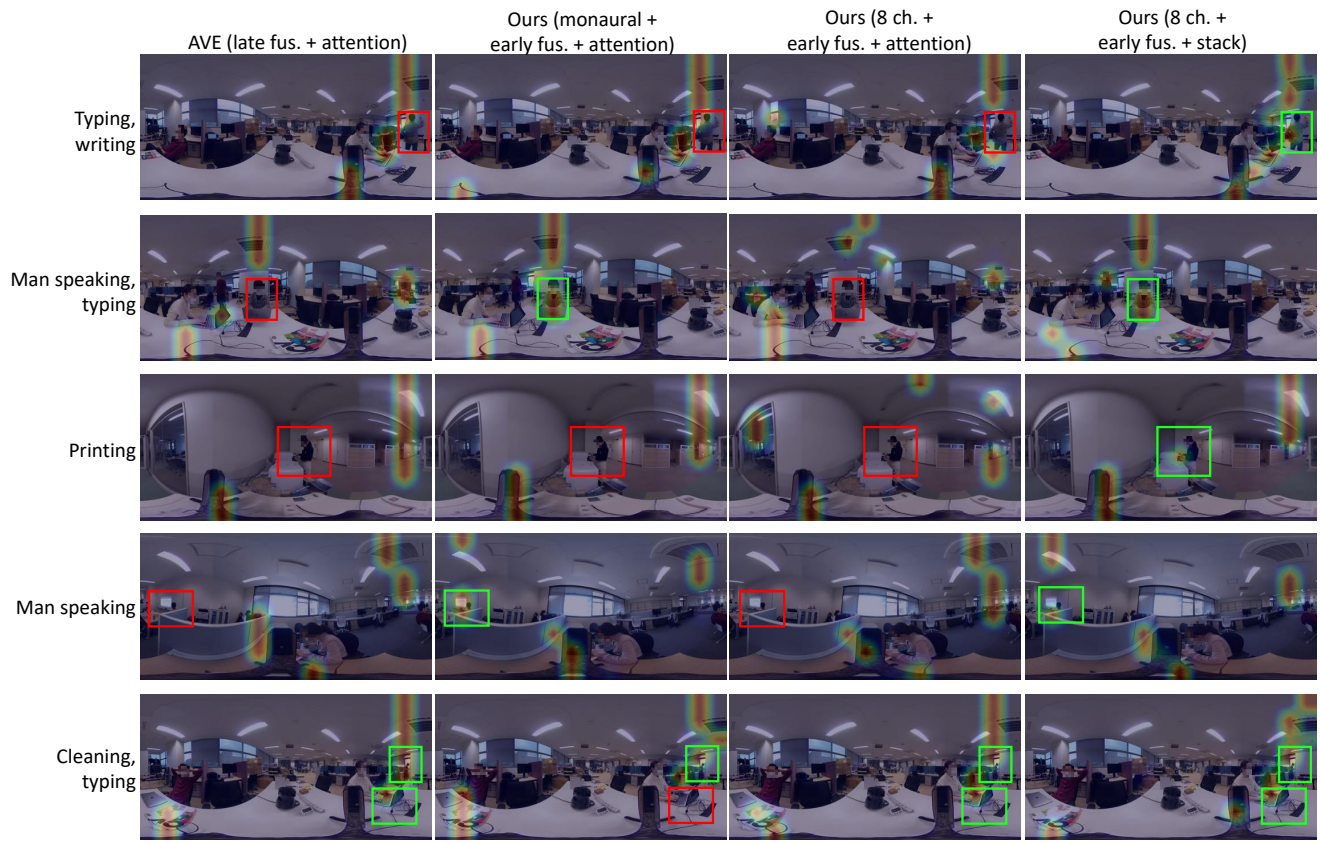


Figure 6. Attention maps for different methods. Green and red rectangles represent correct/wrong attention, respectively.

accuracy of each one-second segment as the evaluation metric. In order to evaluate the effectiveness of our method on *multi-label* classification problem, we also conducted experiments on a subset of our dataset that mostly consists of single-label samples. The subset contains about 300 videos, which can be divided into four categories: *man speaking, woman speaking, typing*, and *nothing*. We refere to the subset as "single-label multi-channel AVE (SLMC-AVE)" dataset.

The comparison of classification accuracy is shown in Table 3. When single-channel audio is input, the performance of AVE [22] slightly outperforms our model on all datasets. This result indicates that late fusion on visual and audio features performs better than early fusion when single-channel audio is used, which is consistent with the results of [22]. However, when multi-channel audio is used, early fusion performs better on all datasets. This result suggests that the early fusion approach, which takes into account the spatial correlation between audio and visual features, is more effective in the case of multi-channel audio because it can distinguish sound information by location. Regarding the audio-spatial fusion approach, the use of HARK was effective on the single-label dataset but performed worse than audio-visual attention on the multi-label dataset. This may be due to the fact that it is difficult to separate sounds using HARK in situations where different sounds are coming from different locations at the same time. In contrast, the proposed method, which has no special processing and simply stacks multi-channel audio as shown in Fig. 4, performed the best. Even with such a simple method, since our method uses early fusion to concatenate visual and

Table 3. Classification accuracy comparison. M denotes the number of audio channels. "Aud. Ops." shows methods for audio-spatial fusion described in Sec. 4.2, where "Stack" indicates a simple multi-channel stack shown in Fig. 4.

Method	M	Aud. Ops.	Dataset		
			AVE [22]	SLMC	MLMC
AVE [22] (late fus.)	1	Attention	**0.72**	0.56	0.57
	8	HARK	-	0.64	0.42
Ours (early fus.)	1	Attention	0.70	0.53	0.56
	8	Attention	-	0.57	0.59
	8	HARK	-	0.66	0.48
	8	Stack	-	**0.67**	**0.65**

Table 4. Comparison of Hamming loss for different categories in MLMC-AVE. M denotes the number of audio channels. "Aud. Ops." shows methods for audio-spatial fusion described in Sec. 4.2, where "Stack" indicates a simple multi-channel stack shown in Fig. 4.

Method	M	Aud. Ops.	man speaking	woman speaking	typing	cleaner	kettle	walking	door
AVE [22] (late fus.)	1	Attention	**0.13**	0.16	0.67	0.13	0.67	0.33	0.73
	8	HARK	0.18	0.19	0.50	0.17	0.57	0.41	0.79
Ours (early fus.)	1	Attention	0.16	0.19	0.69	0.13	0.67	0.33	0.75
	8	HARK	0.18	0.17	0.56	0.15	**0.41**	0.37	0.75
	8	Stack	**0.13**	**0.15**	**0.35**	**0.10**	0.48	**0.28**	**0.70**

audio features that differ from place to place, there is a high possibility that a large weight is placed on the locations that are strongly related to AVEs in the CNN at the latter stage.

Figure 6 shows attention maps of different methods in Table 3. We calculated attention weights for the features taken from just before LSTM and then obtained the sum of the absolute values to generate activation-based spatial attention maps [10]. As shown in the figures, ours (8 ch.+early fus.+stack) generally produces the best visualization result. However, we also found that the ceiling, air-conditioning and other locations outside of the event also drew attention simultaneously, which may have influenced classification results.

We also show the Hamming loss in several categories in Table 4. Hamming loss is calculated as the Hamming distance between y_{true} and y_{pred}, with smaller values indicating better performance:

$$\text{Hamming Loss } = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\left(y_j^{(i)} \neq \hat{y}_j^{(i)}\right),$$

where $y_j^{(i)}$ denotes the jth label of the ith sample. Hamming loss measures a percentage of the number of incorrectly predicted labels over the total number of labels in all samples. In general, we can see that our approach using multi-channel audio achieves lower losses. As a comparison of categories, scores in *man speaking, woman speaking* and *cleaner* are better than in the other categories, while *Door* is the most difficult category for all methods.

6. Conclusion

In this paper, we have presented a method for the AVE localization task when multiple audio-visual events occur simultaneously at different locations. The proposed method uses an omnidirectional camera and a microphone array to acquire different multi-channel audio, each corresponding to different locations in camera images. To maximize the spatial correlation between visual and audio information, we examined several different network architectures based on early and late fusion. We also created a new multi-label multi-channel AVE dataset for evaluation. Experimental results show that, in contrast to the single-label single-channel AVE dataset where late-fusion-based methods are promising, the proposed early-fusion-based method outperforms the existing late-fusion-based method on our dataset. Furthermore, we demonstrated that our method is able to properly focus attention on the sounding locations in images. Application of the proposed method to other more baseline methods, especially the latest vision transformer (ViT)-based methods, and comparative experiments are important future tasks.

Acknowledgement

This work is an outcome of a research project, Development of Quality Foundation for Machine-Learning Applications, supported by DENSO IT LAB Recognition and Learning Algorithm Collaborative Research Chair (Tokyo Tech.).

References

[1] Sharath Adavanne, Archontis Politis, Joonas Nikunen, and Tuomas Virtanen. Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 13(1):34–48, 2018. 2

[2] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of International Conference on Computer Vision (ICCV)*, pages 609–617, 2017. 1, 2

[3] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 435–451, 2018. 1, 2

[4] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, volume 29, 2016. 1, 2

[5] Bin Duan, Hao Tang, Wei Wang, Ziliang Zong, Guowei Yang, and Yan Yan. Audio-visual event localization via recursive fusion by joint co-attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4012–4021, 2021. 2

[6] Ruohan Gao and Kristen Grauman. 2.5 d visual sound. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 3, 4

[7] John Hershey and Javier Movellan. Audio vision: Using audio-visual synchrony to locate sounds. In *Proceedings of*

Advances in Neural Information Processing Systems (NIPS), volume 12, 1999. 2

[8] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 131–135. IEEE, 2017. 4

[9] Hao Jiang, Calvin Murdock, and Vamsi Krishna Ithapu. Egocentric deep multi-channel audio-visual active speaker localization. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10544–10552, 2022. 2

[10] Nikos Komodakis and Sergey Zagoruyko. Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2017. 8

[11] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, volume 31, 2018. 2

[12] Daniel Krause, Archontis Politis, and Konrad Kowalczyk. Comparison of convolution types in cnn-based feature extraction for sound source localization. In *2020 28th European Signal Processing Conference (EUSIPCO)*, pages 820–824. IEEE, 2021. 2

[13] Yan-Bo Lin, Yu-Jhe Li, and Yu-Chiang Frank Wang. Dual-modality seq2seq network for audio-visual event localization. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 2002–2006. IEEE, 2019. 1, 2

[14] Shuo Liu, Weize Quan, Yuan Liu, and Dong-Ming Yan. Bi-directional modality fusion network for audio-visual event localization. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4868–4872. IEEE, 2022. 2

[15] Pedro Morgado, Nuno Nvasconcelos, Timothy Langlois, and Oliver Wang. Self-supervised generation of spatial audio for 360 video. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, volume 31, 2018. 3, 4

[16] Kazuhiro Nakadai, Hiroshi G Okuno, and Takeshi Mizumoto. Development, deployment and applications of robot audition open source software hark. *Journal of Robotics and Mechatronics*, 29(1):16–25, 2017. 3, 6

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

[18] Janani Ramaswamy. What makes the sound?: A dual-modality interacting network for audio-visual event localization. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4372–4376. IEEE, 2020. 2

[19] Janani Ramaswamy and Sukhendu Das. See the sound, hear the pixels. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2970–2979, 2020. 1

[20] Bertrand Rivet, Wenwu Wang, Syed Mohsen Naqvi, and Jonathon A Chambers. Audiovisual speech source separation: An overview of key methodologies. *IEEE Signal Processing Magazine*, 31(3):125–134, 2014. 2

[21] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4358–4366, 2018. 2

[22] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 247–263, 2018. 1, 2, 3, 4, 5, 6, 7, 8

[23] Yu Wu, Linchao Zhu, Yan Yan, and Yi Yang. Dual attention matching for audio-visual event localization. In *Proceedings of International Conference on Computer Vision (ICCV)*, pages 6291–6299, 2019. 2

[24] Yan Xia and Zhou Zhao. Cross-modal background suppression for audio-visual event localization. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19957–19966, 2022. 2

[25] Haoming Xu, Runhao Zeng, Qingyao Wu, Mingkui Tan, and Chuang Gan. Cross-modal relation-aware networks for audio-visual event localization. In *Proceedings of ACM Multimedia*, page 3893–3901, 2020. 2

[26] Cheng Xue, Xionghu Zhong, Minjie Cai, Hao Chen, and Wenwu Wang. Audio-visual event localization by learning spatial and semantic co-attention. *IEEE Trans. on Multimedia*, pages 1–1, 2021. 2

[27] Jiashuo Yu, Ying Cheng, Rui-Wei Zhao, Rui Feng, and Yuejie Zhang. Mm-pyramid: Multimodal pyramid attentional network for audio-visual event localization and video parsing. In *Proceedings of the 30th ACM International Conference on Multimedia*, page 6241–6249, 2022. 2

[28] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 570–586, 2018. 2

[29] Jinxing Zhou, Liang Zheng, Yiran Zhong, Shijie Hao, and Meng Wang. Positive sample propagation along the audio-visual event line. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8436–8444, 2021. 1

[30] Hao Zhu, Man-Di Luo, Rui Wang, Ai-Hua Zheng, and Ran He. Deep audio-visual learning: A survey. *International Journal of Automation and Computing*, 18(3):351–376, 2021. 2