6DoF SELD: SOUND EVENT LOCALIZATION AND DETECTION USING MICROPHONES AND MOTION TRACKING SENSORS ON SELF-MOTIONING HUMAN

Masahiro Yasuda, Shoichiro Saito, Akira Nakayama, Noboru Harada

NTT Corporation, Japan

ABSTRACT

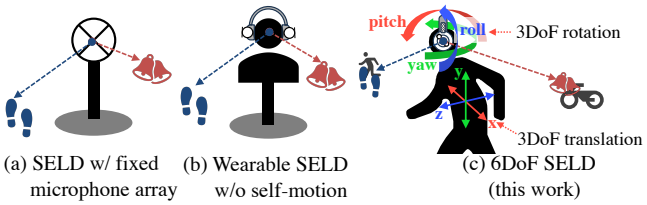
We aim to perform sound event localization and detection (SELD) using wearable equipment for a moving human, such as a pedestrian. Conventional SELD tasks have dealt only with microphone arrays located in static positions. However, self-motion with three rotational and three translational degrees of freedom (6DoF) shall be considered for wearable microphone arrays. A system trained only with a dataset using microphone arrays in a fixed position would be unable to adapt to the fast relative motion of sound events associated with self-motion, resulting in the degradation of SELD performance. To address this, we designed *6DoF SELD Dataset*[1] for wearable systems, the first SELD dataset considering the self-motion of microphones. Furthermore, we proposed a multi-modal SELD system that jointly utilizes audio and motion tracking sensor signals. These sensor signals are expected to help the system find useful acoustic cues for SELD on the basis of the current self-motion state. Experimental results on our dataset show that the proposed method effectively improves SELD performance with a mechanism to extract acoustic features conditioned by sensor signals.

Index Terms— sound event localization and detection, motion tracker, six degrees of freedom, microphone array, dataset

1. INTRODUCTION

Sound event localization and detection (SELD) is a combined task of sound event detection, which estimates the class of event and its onset/offset time, and sound source localization [1, 2]. In this study, we newly defined and addressed 6DoF SELD, a SELD using microphone arrays worn by a self-moving human in six degrees of freedom (6DoF). Here, 6DoF is the sum of three rotational and three translational degrees of freedom, corresponding to behaviors such as walking, looking around, and bending over. The output of 6DoF SELD is similar to that of SELD, but the direction of arrival (DOA) of the sound source is estimated in relative coordinates for the head's orientation. For example, when the sound source is fixed, the estimated DOA moves in the opposite direction of the head motion. One promising application is pedestrian safety assistance through notification of approaching vehicles and humans. Another application is in immersive communication, where the status of the surrounding environment is shared remotely [3, 4]. Moreover, SELD on moving vehicles, such as autonomous cars [5, 6] and surveillance drones [7], can also be considered an application of 6DoF SELD. As a practical constraint in these applications, the system shall be developed under the causal constraint of being able to operate online [8].

Conventional SELD systems mainly use first-order ambisonics (FOA) signals as input and the deep neural networks (DNN) as a regression or classification function for SELD, as shown in Fig. 1 (a), (b) [9–14]. In particular, Fig. 1 (a) shows the most investigated SELD systems that use FOA signals. Such a system would be useful in applications where a device placed in a room, such as a smart speaker, analyzes indoor events. Although the SELD using wearable microphone arrays has rarely been addressed, our previous work [14] addressed wearable SELD by using a microphone array attached to a head and torso simulator (HATS).



(a) SELD w/ fixed microphone array (b) Wearable SELD w/o self-motion (c) 6DoF SELD (this work)

Fig. 1: Conventional and our problem settings of SELD

On the other hand, applications that support moving humans, such as pedestrian safety support, require the ability to handle self-motion microphones. In this problem setting, rapid relative motion of the surrounding sound sources is caused by self-motion, especially rotation. For example, given a look-back motion in one second, all surrounding sound sources move at 180 deg./sec., faster than a 60 km/h car crossing 3 meters from a human. Considering that moving sound sources degrade the localization performance in the conventional SELD task [15], this rapid relative motion should also cause degradation. Adapting to such rapid system changes is considered more difficult for online systems that use only past observations.

In contrast to the difficulties in 6DoF SELD, it has been reported that humans can localize sound sources more accurately when their heads are moving rather than stationary [16–18]. The major reason for this is that dynamic cues, such as dynamic changes in inter-time difference and inter-level difference, play an important role in source localization during self-motion [16]. Considering these human auditory characteristics, utilizing dynamic cues in the 6DoF SELD system is a promising strategy. For this purpose, training data in the 6DoF SELD situation needs to be collected to train a system that can capture the dynamic cues of acoustic features. Furthermore, if the system can observe self-motion in the same way humans use semicircular canals, it is expected to capture dynamic cues more effectively. In fact, the effectiveness of using head rotation information in binaural source localization has been reported [19]. Observation of self-motion is the low-cost option by using inertial sensors, which are commonly used in wearable devices.

Therefore, we propose and publish a new dataset for 6DoF SELD, called *6DoF SELD Dataset*[1]. Unlike conventional SELD datasets, our dataset is designed to identify sound events that occur around a moving human. It uses headphone-type equipment with three motion tracking sensors and 18-channel microphones to measure the position and posture of the head and acoustic signals.

We also propose a new multi-modal SELD system that combines acoustic signals with velocity and angular velocity observations from motion tracking sensors. The system introduces sensor

[1]The dataset is available in `https://github.com/nttrd-mdlab/6dof-seld` (DOI: 10.5281/zenodo.10473531)

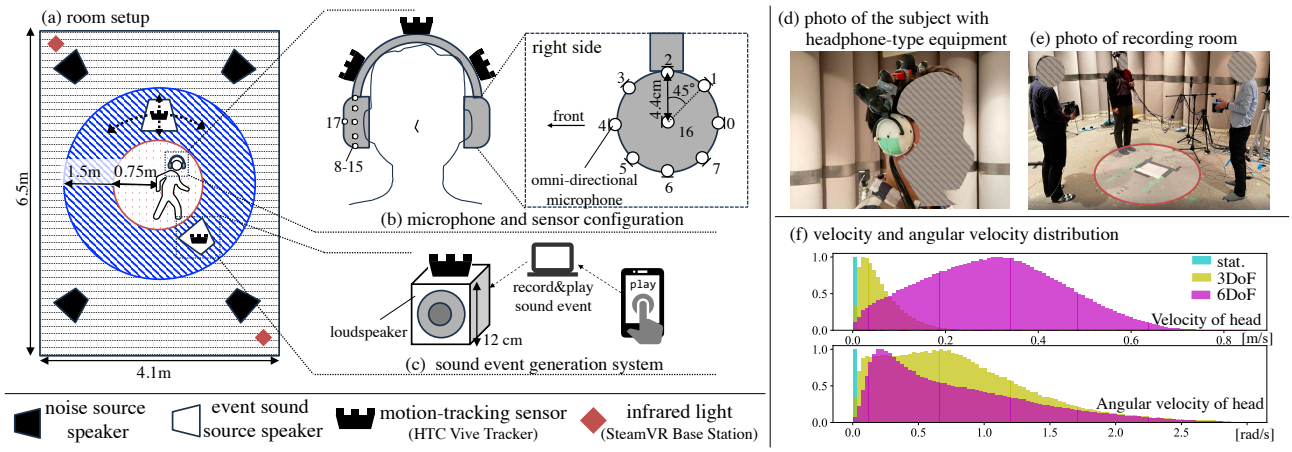


Fig. 2: Recording setup and equipment configuration for *6DoF SELD Dataset*. In (a) and (e), red circles indicate the range of movement of the subject and blue circles indicate the range of the sound source position.

Table 1: Specification of conventional and proposed SELD datasets.

Dataset	Amount of data	Event generation	Other modalities	Self-motion of mic.	Subjects	Source movement	# of classes	# of mic.	SNR [dB]	Rooms
TAU-NIGENS Spatial Sound Events 2021 [11]	13.3h	IR-based	-	stat.	-	Moving	12	4	6 - 30	18
STARSS23 [2]	7.4h	Real	Video	stat.	-	Moving	13	4	natural	16
Wearable SELD [14]	8.3h	IR-based	-	stat.	1 (HATS)	Fixed	12	12	10 - 20, clean	3
Ours: 6DoF SELD	20.1h	Speaker	Motion tracker	stat./3DoF/6DoF	3 (Human)	Fixed	12	18	6 - 20	3

signal-based excitation of the acoustic features to mimic humans' ability to utilize dynamic cues. It was implemented by introducing a multi-modal transfer module (MMTM) proposed for multi-modal speech enhancement and action recognition into SELDNet, the baseline model for the DCASE 2023 task3. Numerical experimental results showed that training the system using the data with the proposed dataset improved 6DoF SELD performance compared to using the data from a stationary microphone array. In addition, using sensor information of velocity and angular velocity was shown to effectively improve 6DoF SELD performance.

2. PROPOSED DATASET

In this section, we describe an overview and specifications of our proposed *6DoF SELD Dataset* and a comparison with conventional SELD datasets.

2.1. Dataset overview

We propose a *6DoF SELD Dataset*[1] for detecting and localizing sound events from the view of a self-motioning human. Unlike conventional SELD datasets with a fixed microphone array [2, 9–11] or a wearable SELD dataset with HATS [14], we record sound events with headphone-type equipment worn by a subject with 6DoF self-motion, i.e., walking and looking around. The 18-channel microphone array and three motion tracking sensors are installed in headphone-type equipment. Motion tracking sensors allow us to observe the position and posture of the head. By time differentiating the position and posture acquired by the motion tracking sensors, it is also possible to simulate the observation of head motion by a more practical sensor such as a 6-axis inertial measurement units (IMUs).

2.2. Dataset and equipment specifications

Figure 2 shows the recording setup and equipment configuration for the *6DoF SELD Dataset*. The recording was conducted in a variable reverberation room, as shown in Fig. 2 (a). A human wearing headphone-type equipment moves in the red circle area, and sound events are played randomly in the blue area. Fig. 2 (b) shows the details of the headphone-type equipment. Each left and right plastic earpad has an 8-channel microphone on the outer edge and a 1-channel microphone on the center. In the experimental section of this paper, only microphones 0, 4, 8, and 12 of these channels were used. Future studies could include array processing using two circular microphone arrays or a more realistic setup using only the two central microphones. Three motion trackers are attached to the headband. The head position and posture are observed as centroid and posture of a triangular composed by these three motion trackers. Fig. 2 (c) shows the sound event generation system. Sound events are generated by randomly playing audio clips from two loudspeakers. Variations of directions of arrival of sound events are reproduced by manually moving the speakers to various heights and angles. The sound source position is recorded by the motion tracking sensor in the absolute coordinate system of the room and then converted to a relative coordinate system to the central human's head on the basis of the observed head posture information.

Table 1 shows the specifications of the *6DoF SELD Dataset* and the conventional SELD dataset. Our dataset generates sound events by playing back pre-recorded sound samples from the speakers as described above. Impulse response (IR)-based generation has the advantage of increased data volume and a high degree of control over experimental conditions but is unsuitable for self-motion microphones. Using real-life sound events, as in STARSS23, is most compatible with real-life scenarios, but collecting a large amount of labeled data is difficult. Our dataset uses a speaker-based playback of sound events, allowing us to record sound events with the self-motion microphone and collect a sufficient amount of labeled data (20.1 hours). Our dataset and STARSS23 record multimodal signals as a dataset for SELD. The STARSS23 records 360° video of a human generating sound events. Although video modalities can be used to record body movements, they are not suitable for our pur-

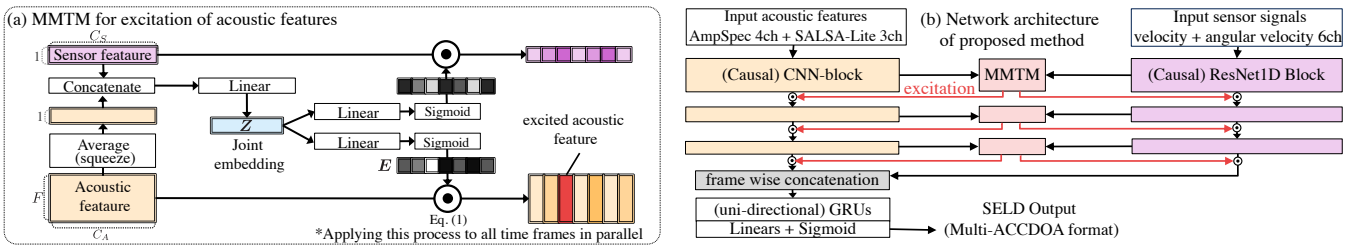


Fig. 3: (a) MMTM for excitation of acoustic features on the basis of sensor signals. C_A and C_S are the number of channels of the sensor and acoustic features, F is the number of dimensions of the acoustic features, and $\odot$ denotes the Adamar product. (b) Network architecture of proposed multi-modal SELD system. "AmpSpec" denotes the amplitude of the spectrogram.

pose of SELD with wearable equipment, so we used motion tracking sensors. The self-motion of the microphone is included only in our dataset. The dataset is divided into three subsets ("stat.", "3DoF", and "6DoF") in accordance with the self-motion condition. In "stat." the subject is seated in a chair; in "3DoF," the subject makes rotational movements of the head and body while standing; and in "6DoF," the subject walks, changes direction, and swivels within a circle of $0.75m$. These subsets' actual velocity and acceleration distributions are shown in Fig. 2 (f). Sound events consisting of 12 classes were recorded by an 18-channel microphone (Fig. 2-(b)) attached to headphone-type equipment. Signal-to-noise ratios (SNR) were controlled to 6, 10, and 20 dB, allowing quantitative analysis of the system's noise robustness. The recording room is equipped with a variable reverberation room, where the reverberation time is controlled in three steps ($T_{60}^{500\text{Hz}} = 0.12, 0.30, 0.41$ sec).

All microphones used for recording were Hosiden KUB4225, with a sampling frequency of 48k and a bit depth of 16 bits. A combination of HTC Vive Tracker (2018) and HTC SteamVR Base Station 2.0 was adopted for the motion tracking sensors. The sensor signals were recorded at a nonuniform sampling rate of about 40 fps and then downsampled to a uniform sampling rate of 20 fps.

3. PROPOSED METHOD

In this section, we describe the proposed wearable SELD system utilizing the joint feature of audio and sensor signals.

3.1. Basic concept

For a 6DoF SELD system to maintain robust performance during self-motion, it is desirable to exploit appropriate acoustic features as cues depending on the state of self-motion. Estimating self-posture from acoustic signals has been explored [20–22], but online estimation of velocity and angular velocity, which are considered relevant to the dynamic cue, is still difficult. Utilizing tracking sensor observations, especially velocity and angular velocity, is expected to enable these self-motion states to be acquired and the system to more appropriately utilize dynamic cues to improve SELD performance.

Therefore, we propose a multi-modal SELD system that combines acoustic signals with velocity and angular velocity signals obtained by tracking sensors. This system mimics the mechanism by which people utilize dynamic cues as excitation of specific acoustic features on the basis of velocity and angular velocity. It is represented by weighting the C_A channel and F dimensional acoustic features $\boldsymbol{\phi}^c \in \mathbb{R}^F$ ($c \in [1, C_A]$) by the excitation vector $\boldsymbol{E} \in \mathbb{R}^{C_A}$, which depends on velocity $\boldsymbol{\nu} \in \mathbb{R}^3$ and angular velocity $\boldsymbol{\omega} \in \mathbb{R}^3$:

$$[\tilde{\boldsymbol{\phi}}^1, \ldots, \tilde{\boldsymbol{\phi}}^C] = [\boldsymbol{\phi}^1, \ldots, \boldsymbol{\phi}^C] \odot \boldsymbol{E}(\boldsymbol{\nu}, \boldsymbol{\omega}) \quad (1)$$

As an implementation of this principle, we used MMTM [23], a method for fusing convolutional neural network (CNN) features of multi-modal signals. Fig. 3 (a) shows the block diagram of the MMTM applied to our problem setting. In MMTM, the acoustic features obtained at each layer of the CNN are first averaged in the frequency axis. This operation is called "squeeze" in the original MMTM. From these squeezed acoustic features and sensor features, a joint embedding $\boldsymbol{Z} \in \mathbb{R}^{D_z}$ is then extracted. Finally, The excitation for each modality feature is computed based on $\boldsymbol{Z}$. Note that in our implementation, the original MMTM is modified to avoid squeezing the time axis of the features to excite the appropriate acoustic features at each time frame. Since the excitation to acoustic features is dependent on the sensor signal through the joint embedding $\boldsymbol{Z}$, MMTM is an appropriate implementation of the principle of Eq. 1.

3.2. Implementation details

Fig. 3 (b) shows the network architecture of the proposed method. The inputs to the system are 4-channel acoustic signals and 6-channel velocity/angular velocity signals. The 4-channel acoustic signals are composed of channels 0, 4, 8, and 12 of the 18 microphones in the dataset. It corresponds to the microphones embedded in the front and rear of the left and right earpads. From these acoustic signals, 4-channel amplitude spectrograms and 3-channel SALSA-Lite features [24] are extracted as input features. The velocity/acceleration signals are extracted by first-order differentiation of the position/angle obtained from the tracking sensor observation signals while smoothing them with a Savitzky-Golay filter [25].

The DNN used Causal-SELDNet, a causal modification of SELDNet, which is the baseline model of DCASE 2023 task 3 [12]. For causal modification, there are three modifications to the network. The first is the use of causal convolution, which uses only past time frames for convolution in the CNN; the second is the change of the bi-directional gated recurrent unit (GRU) to a uni-directional GRU; and third is the removal of multi-head self-attention (MHSA). The DNN for sensor signals was a 1D CNN with ResNet-like skip connections, as in previous studies of action recognition [26]. The 1D CNN was also modified to use causal convolution. 2D and 1D CNNs for extracting features from acoustic and sensor signals consisted of the same number of blocks and time frames, and MMTM excitation was applied to all outputs for each time frame. The acoustic and sensor features obtained as outputs of the CNNs were concatenated in the feature axis and input to the subsequent uni-directional GRUs. Multi-ACCDOA was used as the loss function for training [27].

4. EXPERIMENTS

4.1. Experimental setup

Hyper-parameter: For the short-time Fourier transform, a Hanning window of 1024 points and a shift of 0.025 sec. were used. From the extracted spectrograms, the frequencies corresponding to 50 - 9050 Hz were cut out, resulting in a frequency dimension of 64. All model parameters of Causal-SELDNet are the same as in the

DCASE2023 baseline model. The 1D CNN used for the encoder of the sensor signal consisted of three ResNet blocks. The number of CNN filters was (64, 32, 16), kernel size, stride, and padding were 5,1,2, respectively. The acoustic and sensor features extracted by the CNN and 1D CNN were concatenated for each time frame. The Adam optimizer was used for learning, with an initial learning rate of 0.01 [28]. Learning was concluded in 100 epochs, and the parameters obtained in the epoch with the lowest validation loss were adopted.

Comparison methods: To evaluate the effectiveness of the proposed method, the following conditions were compared:

(A) Baseline (stat.) A method using Causal-SELDNet trained on the "stat." subset of our dataset. This is a similar situation to the conventional Wearable SELD Dataset [14], which deals with conditions where the head is fixed. We used this system to investigate the performance degradation of a system trained with fixed microphone data under self-motion conditions.

(B) Baseline A method using Causal-SELDNet trained on all the data in our dataset. A variant of this method (B/3), trains Causal-SELDNet on 1/3 of the data in the dataset. Since this is the same amount of data as (A), the change in performance can be validated when training data with and without the self-motion microphone under the same amount of data.

(C) Audio-SENet A method using modified Causal-SELDNet that replaces the CNN with the squeeze and excitation network (SENet) [29]. This method can be considered as an uni-modality version of the method (E) using MMTM. Comparing it with (E), the change in performance due to the use of sensor signals can be validated.

(D) Sensor-concat This method is a variation of (E) that directly concatenates sensor signals and acoustic features extracted using CNN. By comparing this method with (E), the performance improvement by using sensor signals for the excitation of acoustic features based on self-motion can be clarified.

(E) Sensor-MMTM Proposed method described in Sec. 2

Evaluation metrics: We adopt the same metrics with DCASE 2023 task3 [30]. The metrics used for event detection were location-dependent F1-score $F_{\leq\Theta}$ and error rate $ER_{\leq\Theta}$. These are calculated by counting as true positives (TP) if the event class matches the ground truth label and the event localization is correct within the threshold angle Θ. In this experiment, we adopt $\Theta = 20^\circ$ as in DCASE 2023 task3. Class-dependent localization error LE_{CD} and localization recall LR_{CD} were used as metrics for event localization. LE_{CD} is the angular error between the estimated source location and the ground truth, calculated using only TP time frames; LR_{CD} represents recall of the number of active source estimations. All experiments were performed three times for different initial parameters. All experimental results are shown with the standard error of the metrics obtained from the three experiments.

4.2. Result

Table 2 compares SELD performances with and without self-motion of a subject in the training data. First, a performance comparison for all test data shows that (B/3) outperforms (A) on all metrics. Next, comparing performance under different self-motion conditions in (A), a performance gap exists between the "stat." and "3DoF" conditions. It suggests that systems trained only on the stationary microphone data cannot adequately cope with the rotation motion. In addition, the "3DoF" subset performs poorly compared to the "6DoF" subset. It is considered because "3DoF" contains faster rotational motion than "6DoF" as shown in Fig. 2 (f). On the other hand, in (B/3), the performance gap between the "stat." and "3DoF" conditions is reduced for the "3DoF" and "6DoF" conditions. These results suggest that 6DoF SELD requires not only a dataset under conventional stationary conditions but also a dataset including the proposed self-motion.

Table 2: SELD performance for different DoF of self-motion included in the training data. The "stat." "3DoF" and "6DoF" indicate that the microphone includes data for stationary, rotating, and rotating/translating cases, respectively.

	motion		SELD performance			
	train	test	$ER_{\leq 20^\circ} \downarrow$	$F_{\leq 20^\circ} \uparrow$	$LE_{CD} \downarrow$	$LR_{CD} \uparrow$
(A) Baseline (stat.)	stat.	all	$0.63_{\pm 0.007}$	$39.0_{\pm 0.5}$	$25.6_{\pm 0.2}$	$83.3_{\pm 0.04}$
		stat.	$0.48_{\pm 0.003}$	$53.8_{\pm 0.5}$	$19.5_{\pm 0.1}$	$85.8_{\pm 0.04}$
		3DoF	$0.71_{\pm 0.01}$	$32.8_{\pm 0.4}$	$28.4_{\pm 0.2}$	$82.2_{\pm 0.2}$
		6DoF	$0.69_{\pm 0.02}$	$31.9_{\pm 1.3}$	$28.5_{\pm 0.5}$	$81.8_{\pm 0.1}$
(B/3) Baseline (1/3 data)	all	all	$0.55_{\pm 0.01}$	$45.7_{\pm 0.6}$	$23.0_{\pm 0.1}$	$84.6_{\pm 0.5}$
		stat.	$0.51_{\pm 0.01}$	$50.6_{\pm 1.1}$	$20.5_{\pm 0.3}$	$86.1_{\pm 0.4}$
		3DoF	$0.57_{\pm 0.008}$	$44.9_{\pm 0.5}$	$24.2_{\pm 0.1}$	$83.9_{\pm 0.3}$
		6DoF	$0.59_{\pm 0.005}$	$42.0_{\pm 0.5}$	$24.2_{\pm 0.1}$	$83.6_{\pm 0.8}$

Table 3: Comparison of SELD performance for different network architectures and input modalities.

	SELD performance			
	$ER_{\leq 20^\circ} \downarrow$	$F_{\leq 20^\circ} \uparrow$	$LE_{CD} \downarrow$	$LR_{CD} \uparrow$
(B) Baseline	$0.55_{\pm 0.003}$	$49.1_{\pm 0.2}$	$21.6_{\pm 0.1}$	$85.2_{\pm 0.1}$
(C) Audio-SENet	$0.54_{\pm 0.003}$	$51.1_{\pm 0.2}$	$21.2_{\pm 0.2}$	$\underline{85.9}_{\pm 0.1}$
(D) Sensor-concat	$\underline{0.53}_{\pm 0.005}$	$\underline{51.4}_{\pm 0.4}$	$\underline{20.9}_{\pm 0.1}$	$85.2_{\pm 0.2}$
(E) Sensor-MMTM	$\mathbf{0.51}_{\pm 0.003}$	$\mathbf{54.1}_{\pm 0.1}$	$\mathbf{20.0}_{\pm 0.1}$	$\mathbf{86.1}_{\pm 0.3}$

Table 3 compares SELD performances for different network architectures and input modalities. The (D) and (E), which use velocity and angular velocity extracted from head tracking sensors as input features, perform better on all metrics than (B) and (C), which only use audio modality. The performance improvement in (D) indicates the effectiveness of using sensor features for temporal modeling. The performance improvement in (E) indicates that MMTM-based excitation of acoustic features based on sensor signals is effective for SELD with self-motion. This fact is consistent with the property that changes in acoustic features obtained when a human moves his/her head can be used as dynamic cues for source localization. In (D), where the sensor signal is directly input to the GRU responsible for temporal modeling, a certain performance improvement is also observed compared to (B) and (C). These results indicate that the use of tracking sensors in 6DoF SELD improves feature extraction and temporal modeling, thus enhancing the performance of SELD.

5. CONCLUSION

We designed a *6DoF SELD Dataset* and proposed a multi-modal sound event localization and detection (SELD) system that combines motion tracking sensor signals with acoustic signals. Our dataset provides recordings of acoustic events around a subject moving at 6DoF. The data is captured using a headphone-type device with embedded microphones and motion tracking sensors. The proposed method utilizes dynamic cues by applying excitations to the acoustic features in accordance with the velocity and angular velocity extracted from the sensor signals. Validation experiments on our dataset showed that learning the system using a dataset that includes self-motion improves SELD performance during movement. Furthermore, it also demonstrated that using sensor signals can improve SELD performance. Therefore, the proposed dataset and system effectively perform SELD on a self-motioning human.

6. 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. sel. top. signal process.*, vol. 13, 2019.

[2] K. Shimada, A. Politis, P. Sudarsanam, D. Krause, K. Uchida, S. Adavanne, A. Hakala, Y. Koyama, N. Takahashi, S. Takahashi, T. Virtanen, and Y. Mitsufuji, "STARSS23: An audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events," 2023.

[3] J. Herre and S. Disch, "MPEG-I immersive audio – reference model for the virtual/augmented reality audio standard," *Audio Engineering Society*, vol. 71, no. 5, pp. 229–240, 2022.

[4] D. McGrath, S. Bruhn, H. Purnhagen, M. Eckert, J. Torres, S. Brown, and D. Darcy, "Immersive audio coding for virtual reality using a metadata-assisted extension of the 3gpp evs codec," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 730–734.

[5] Y. Xu, Q. Kong, W. Wang, and M. D. Plumbley, "Surrey-cvssp system for DCASE2017 challenge task4," in *Tech. Rep., DCASE2017*, 2017.

[6] D. Lee, S. Lee, Y. Han, and K. Lee, "Ensemble of convolutional neural networks for weakly-supervised sound event detection using multiple scale input," in *Tech. Rep., DCASE2017*, 2017.

[7] X. Chang, C. Yang, X. Shi, P. Li, Z. Shi, and J. Chen, "Feature extracted DOA estimation algorithm using acoustic array for drone surveillance," in *Proc. of IEEE 87th Veh. Technol. Conf.*, 2018.

[8] K. Nagatomo, M. Yasuda, K. Yatabe, S. Saito, and Y. Oikawa, "Online sound event localization and detection for real-time recognition of surrounding environment," *Applied Acoustics*, vol. 199, pp. 108961, 2022.

[9] S. Adavanne, A. Politis, and T. Virtanen, "A multi-room reverberant dataset for sound event localization and detection," in *Proc.of 4th Workshop on Detection and Classification of Acoust. Scenes and Events (DCASE)*, 2019.

[10] A. Politis, S. Adavanne, and T. Virtanen, "A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection," in *Proc. Workshop Detect. Classif. Acoust. Scenes Events (DCASE)*, 2020.

[11] A. Politis, S. Adavanne, D. Krause, A. Deleforge, P. Srivastava, and T. Virtanen, "A dataset of dynamic reverberant sound scenes with directional interferers for sound event localization and detection," *arXiv preprint arXiv:2106.06999*, 2021.

[12] "DCASE 2023: Sound event localization and detection evaluated in real spatial sound scenes," `https://github.com/sharathadavanne/seld-dcase2023`, Accessed: 2034-09-05.

[13] Q. Wang, Y. Jiang, S. Cheng, M. Hu, Z. Nian, P. Hu, Z. Liu, Y. Dong, M. Cai, J. Du, and C. Lee, "The nerc-slip system for sound event localization and detection of dcase2023 challenge," Tech. Rep., DCASE2023 Challenge, June 2023.

[14] K. Nagatomo, M. Yasuda, K. Yatabe, S. Saito, and Y. Oikawa, "Wearable SELD dataset: Dataset for sound event localization and detection using wearable devices around head," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 156–160.

[15] T. Nguyen, K. Watcharasupat, Z. Lee, N. K. Nguyen, D. Jones, and W. Gan, "What makes sound event localization and detection difficult? insights from error analysis," in *Proc. 6th Workshop Detect. Classif. Acoust. Scenes Events (DCASE)*, 2021.

[16] H Wallach, "The role of head movements and vestibular and visual cues in sound localization," *Experimental Psychology*, vol. 27, no. 4, pp. 339–368, 1940.

[17] M. Kato, H. Uematsu, M. Kashino, and T. Hirahara, "The effect of head motion on the accuracy of sound localization," *Acoustical Science and Technology*, vol. 24, pp. 315–317, 2003.

[18] I. Toshima and S. Aoki, "The effect of head movement on sound localization in an acoustical telepresence robot: Telehead," in *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2006, pp. 872–877.

[19] G. García-Barrios, D.A. Krause, A. Politis, J.M. Gutiérrez-Arriola, and R. Fraile, "Binaural source localization using deep learning and head rotation information," in *Proc. 30th Eur. Signal Process. Conf. (EUSIPCO)*, 2022.

[20] G. Lian, Y. Wakabayashi, T. Nakashima, and N. Ono, "Self-rotation angle estimation of circular microphone array based on sound field interpolation," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2021, pp. 1016–1020.

[21] R. Gao, C. Chen, Z. Al-Halah, C. Schissler, and K. Grauman, "Visualechoes: Spatial image representation learning through echolocation," in *Proc. of the European Conf. on Computer Vision (ECCV)*, 2020.

[22] F. Antonacci, J. Filos, M. R. P. Thomas, E. A. P. Habets, A. Sarti, P. A. Naylor, and S. Tubaro, "Inference of room geometry from acoustic impulse responses," *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 20, no. 10, pp. 2683–2695, 2012.

[23] H. R. Vaezi Joze, A. Shaban, M. L. Iuzzolino, and K. Koishida, "Mmtm: Multimodal transfer module for cnn fusion," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 13286–13296.

[24] T. N. Tho Nguyen, D. L. Jones, K. N. Watcharasupat, H. Phan, and W. Gan, "Salsa-lite: A fast and effective feature for polyphonic sound event localization and detection with microphone arrays," in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 716–720.

[25] A. Savitzky and M. J. E. Golay, "Smoothing and differentiation of data by simplified least squares procedures.," *Analytical Chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964.

[26] Q. Kong, Z. Wu, Z. Deng, M. Klinkigt, B. Tong, and T. Murakami, "MMAct: A large-scale dataset for cross modal human action understanding," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[27] 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 *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 316–320.

[28] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proc. IEEE Int. Conf. on Learn. Represent. (ICLR)*, 2015.

[29] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[30] 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 Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 684–698, 2021.