

UNS Exterior Spatial Sound Events dataset for urban monitoring

Siniša Suzić*, Irene Martín-Morató†, Nikola Simić*, Charitha Raghavaraju†,
Toni Heittola†, Vuk Stanojev*, Dragana Bajovic*

*Faculty of Technical Sciences, University of Novi Sad, Serbia
{sinisa.suzic, nikolasimic, vukst, dbajovic}@uns.ac.rs

†Computing Sciences, Tampere University, Finland
{irene.martinmorato, charitha.raghavaraju, toni.heittola}@tuni.fi

Abstract—This paper presents the UNS-Exterior Spatial Sound Events 2023 (UNS-ESSE2023) dataset, which is targeted for applications related to monitoring urban environments. The dataset comprises spatial recordings collected outdoors in real acoustic environments by playing ambience and target sound samples with eight speakers placed circularly around the microphone array. The target sound events are three anomaly sounds (*gunshot*, *boom (explosion)*, and *shatter*), specifically selected as examples of unexpected sound events in the context of monitoring urban spaces. The dataset is evaluated using the sound event detection and localization baseline system from the DCASE2023 challenge. The model was fine-tuned for the dataset to introduce a benchmark for sound event localization and detection in exterior acoustic environments. Comparisons are also made with similar outcomes from the STARS22 dataset, a reference dataset for the SELD task in interior conditions. Results are presented using information about the different levels of signal-to-noise ratio and ambience sound pressure levels, showcasing the complexity of the dataset.

Index Terms—Spatial sound events dataset, sound event localization and detection

I. INTRODUCTION

Sound Event Localization and Detection (SELD) is a combined task that has to simultaneously perform sound source localization (SSL) and sound event detection (SED). Localization involves identifying the directional attributes of sound events, which can range from simple characteristics like left-to-right or right-to-left direction to more complex ones like azimuth and elevation, denoting the direction of arrival of a recognized sound event. These directional characteristics are determined with respect to a reference microphone’s position.

In recent years, different datasets with spatial sound events for SELD research have been released, showing the increase in popularity of this task. In [1], the dataset was generated synthetically by convolving sound event recordings with spatial room impulse responses. Another example of synthetically generated database for the purposes of SELD task using wearable microphone array is described in [2]. However,

This work was funded by the European Union’s Horizon 2020 research and innovation program MARVEL under grant agreement No 957337. This publication reflects the authors’ views only. The European Commission is not responsible for any use that may be made of the information it contains. The authors wish to thank CSC-IT Centre of Science Ltd., Finland, for providing computational resources.

synthetic data is not able to reproduce the complexity or real-life scenarios. In [3], SECL-UMons is presented, a dataset that uses non-synthetic recordings. The dataset consists of 11 target classes, recorded in two different rooms, with a total duration of 5.24 hours. The positions and ordering of target events were fixed and defined by a recording script, failing to capture the variability and diversity of sounds in a realistic scene. In [4], STARSS22 dataset approximates to a more real scenario, where the scenes were acted by the humans but were not strongly scripted. Actors acted naturally, without strong instructions or guidelines on which sounds to play. The dataset includes temporal and spatial annotation of each sound event present, consisting of 13 target classes, recorded in 11 rooms, with a total duration of about 4h 52min.

The collection of such datasets is time-consuming and expensive due to several factors, multiple sources and locations have to be considered, while also respecting relevant privacy regulations and following ethical guidelines. The annotation process is the one factor that increases the overall time needed to create a SELD dataset. Once the audio data has been recorded for multiple sources and locations, human annotators have to manually annotate temporal, position, and class-wise information for each recorded audio file. The overall complexity of such data collection tasks results in a lack of a real-life outdoor SELD datasets.

Due to the increased popularity of smart cities, applications such as monitoring in urban spaces have gained interest in recent years [5]. In these applications, static cameras with or without microphones are typically used for monitoring crowded areas and public events. In comparison with video monitoring, acoustic monitoring has certain advantages: It does not require a direct line of sight to the monitored area, it captures relevant information in low visibility conditions such as during the night time, and the number of sensors needed to cover large areas is lower than that required for video monitoring [6]. Acoustic monitoring in urban environments has been proposed, for example, for noise monitoring [7], sound classification [8], and anomalous sound detection applications [9], [10].

This paper focuses on the SELD-based acoustic monitoring used for detecting and localizing anomalous events in public spaces. We present a novel dataset, UNS-Exterior Spatial

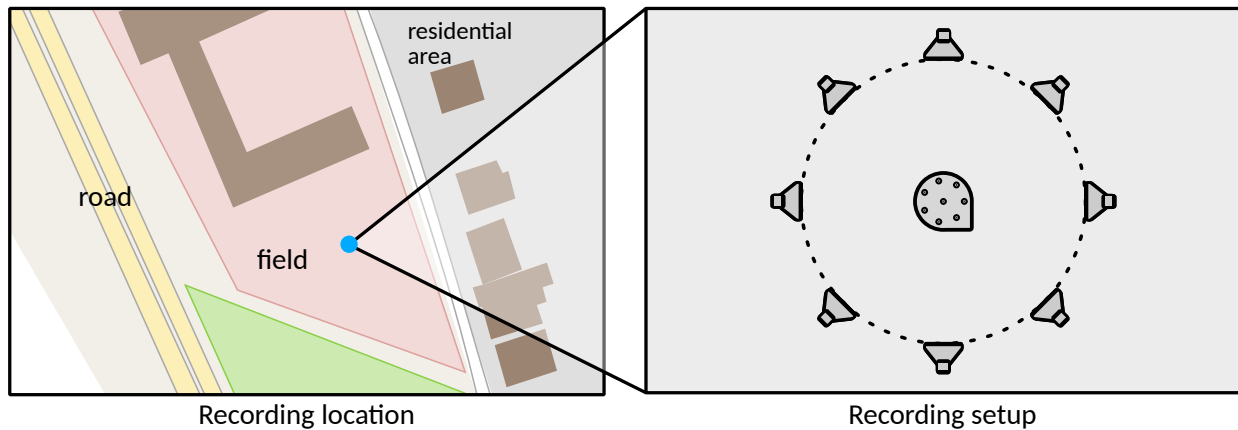


Fig. 1. Illustration of the recording location and speakers' positions in relation to the recording device.

Sound Events 2023 (UNS-ESSE2023) [11], with spatial sound events, where audio recordings have been collected outdoors, providing temporal, position, and class-wise information for each recording. The three sound classes are selected to reflect anomalous and hazardous events: gunshots, boom (explosion), and glass shatters. Similar sound classes have been used in previous research focusing on audio surveillance and detection of hazardous events [12]. UNS-ESSE2023 is the first spatial sound event dataset we are aware of containing recordings done outdoors in a real acoustic environment. The dataset is provided with a fixed cross-validation setup, and evaluated with baseline system to provide a good starting point for later studies.

The paper is organized as follows: Section II presents dataset collection process, explaining the recording setup, data preparation, and post-processing. Section III presents the model architecture used to analyse our dataset and the experiment setup. Section IV includes the evaluation results and discussion. Finally, Section V presents conclusions and future work.

II. DATABASE

The UNS-Exterior Spatial Sound Events 2023 database consists of recordings of anomalous events collected outdoors by using professional equipment of the audio-visual equipment rental services Studio Berar¹ (Novi Sad, Serbia).

A. Recording setup

The recording setup consists of eight speakers circularly placed around the recording device in equidistant positions, as illustrated in Figure 1, right panel. Each speaker was placed on top of a stand of 120 cm in height. We used the JBL VP7212MDP speakers, each connected to an audio console MIDAS M32. The audio console connected to the speakers was further connected to a laptop with a Digital Audio Workstation (DAW) to control playing the sound. The speakers were utilized to reproduce the sound events as well as the background sound, as detailed in the next subsection.

The recording device, shown in Figure 2, was developed by Infineon Technologies AG. The device consists of an 8-channel microphone array and enables wireless transmission of sound streams to the dedicated computer or recording directly to the SD card which can be installed to the device. The latter was used for this database recording. The microphone array consists of seven IM69D130 omnidirectional MEMS microphones, circularly and equidistantly positioned with an inter-microphone distance of 7 cm with one in the center of the array.

The overview of the recording setup is shown in the left panel of Figure 1. The map illustrates the recording conditions, where the recording field was located next to a busy road, which can introduce some extra realistic noise to the recordings.

Two recording scenarios were selected, with the only difference being the distance between the speakers and the microphone array (5 m or 10 m). Three recording sessions per scenario were produced, resulting in a total of six recording sessions, each lasting approximately 75 minutes.

B. Data preparation

Target sounds were extracted from the FSD50k dataset [13], and three anomalous events were selected: gunshot, boom (explosion), and shatter. Since audio files in FSD50k are a bit longer than the target event itself, the files were first semi-automatically processed to extract only audio parts of interest. In the first step of this process, the power of the signal for 5 ms windows was calculated. Signal power values were then used to detect the events comparing them with some heuristically determined threshold. Since this plain algorithm could be error-prone, all files were then examined, and appropriate labels were re-set manually.

A corresponding audio mixture was created for every audio file extracted from the FSD50k dataset. The audio mixture is represented as 8-channel audio whose length is 10 s. The target event was positioned in one randomly chosen audio channel in the middle of the 10-second segment. The audio in the remaining seven channels was randomly extracted from the

¹<https://studioberar.com/kontakt/>



Fig. 2. Recording device. Microphone positions are marked with red circles.

FSD50k dataset from samples labelled “chatter”. These seven channels will be named *ambience channels* in the rest of the text. For every ambience channel, the appropriate gain was applied. The gain value was calculated so that the signal-to-noise ratio (SNR) between the target event and the audio in the corresponding ambience channel represents some randomly chosen value in the -10 to 10 decibels range. Before obtaining the gain values, all the channels were equalized to the same energy. Finally, a linear 40 ms fade-in and fade-out window was applied to the ambience channels to avoid sudden signal bursts. Once the final track for recording was prepared, the neighbouring mixtures were separated with a 1-second silence. For each recording setup, the procedure described in section II-A was applied, i.e., each target event is being recorded in both positions but with different ambience noise tracks and their corresponding gain values.

C. Post-processing

The recording device saves recorded audio in chunks of length 233.017 s. The maximal length of the chunk is determined by the maximum file size allowed by the device, which is set to 256 MB by the manufacturer. After the chunk concatenation, we noticed that there was a slight time delay between the original tracks reproduced on the speakers and the recorded ones, which increases over time. For this reason, original labels could not be used to extract audio mixtures directly, and in order to achieve mixture extraction, a semi-automatic procedure was applied.

The algorithm to automatically label the mixtures had the following steps:

- 1) Extract recorded mixture using the expected timestamps, which are extended for 10 ms on both sides. Expected timestamps are calculated using the original time stamps as well as the outputs from processing the previous mixture.
- 2) Calculate correlation between a recorded mixture and the original mixture signal.

- 3) Based on the maximum correlation value, determine the offset between the two signals.
- 4) Update the expected timestamps for the next recorded mixture.

The suggested algorithm performed well in the majority of cases except for some segments where the presence of background noise was substantial. The automatically produced labels are manually validated and corrected when needed.

We also provide the Sound Pressure Level (SPL) value for the background noise at the recording position captured by an additional calibrated microphone placed at the same location as the microphone array. The SPL values were monitored with DAW, however the workstation did not allow to export the values and they had to be recorded from the screen and extracted using an optical character recognition library [14]. These values were calculated in the middle of the region between two consecutive mixtures.

D. Database content

The generated dataset consists of a total of 2392 mixtures. The mixtures are split into train, validation, and test following the splits from FSD50k [13], ensuring that the same events are not present in two or more splits. The number of sound events per split is shown in Table I. Each mixture is 10 seconds long with a 48kHz sampling rate. The total length of the dataset is 6.6 hours. Together with the audio files, spatiotemporal annotation of the sound events is provided. These annotations include temporal onset and offset, azimuth, source distance, SNR level and SPL values for background ambience noise.

TABLE I
DATASET CONTENT PER SETS IN CROSS-VALIDATION SETUP.

Sound events	Train	Validation	Test
Boom	280	54	74
Gunshot	588	108	268
Shatter	700	128	192
Total	1568	290	534

III. EXPERIMENTS

The baseline system used in sound event localization and detection task in DCASE Challenge 2023² was selected for our experiments to set a standard benchmark. The system is evaluated using the provided cross-validation setup.

A. Baseline system

The architecture of the baseline system uses three convolutional neural network (CNN) blocks to learn high-level representations, followed by two bidirectional recurrent layers as in [15]. Two multi-head self-attention (MHSA) blocks are added as proposed in [16] to capture the temporal relationships. The complete architecture of the system can be seen in Figure 3.

Given the nature of the data, where eight microphones are available, the configuration of the dataset loading parameters is

²<https://github.com/sharathadavanne/seld-dcase2023>

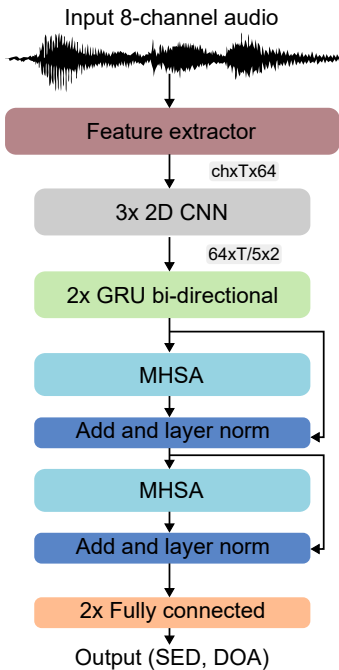


Fig. 3. Baseline system architecture.

MIC (from microphone signals) instead of FOA (ambisonics). The input features are mel-band spectrograms concatenated to the generalized cross-correlation (GCC) sequences for MIC as spatial features to identify the time difference of arrival of a signal.

The frame sequence is mapped to a Single Activity-Coupled Cartesian Direction of Arrival Representation (ACCDOA) [17] sequence output, which encodes both SED and direction of arrival (DOA). The Multi-ACCDOA encoding did not help improving the performance, which is expected, since there are no multiple sources.

B. Experimental set up

Different model configurations were examined, and the multiple parameters were fine-tuned in order to improve the performance of the benchmark system. Given the recording setup, we present an ablation study of the conditions of different SNR levels for the events in the mixtures and the effect of Sound Pressure Level (SPL) before and after the recording.

The performance of sound event detection is assessed following the joint SELD metrics defined in [18]. F1-score (20°) and Error Rate (20°) metrics are used, calculated in a location-dependent manner by applying a spatial threshold for true positives. True positives are identified when an event is correctly detected and localized within a 20-degree range of the ground truth. The evaluation of localization performance is conducted on a per-class basis, employing metrics such as localization error (LE_{CD}) and localization recall (LR_{CD}).

IV. RESULTS AND DISCUSSION

The system was trained and tested five times to deal with the non-deterministic results produced with GPU cards. The results presented are the average performance from the five independent trials.

Table II shows the results obtained with the baseline SELD system described in Section III-A. For comparison, we show the results using the STARSS23 dataset from the baseline system from DCASE2023 Challenge. The architecture of the model is the same, however the configuration setup is different given the differences of the dataset. We can see that even though the number of classes for the UNS-ESSE2023 is smaller (3 vs. 13 sound event classes), the outdoor conditions and the different levels of SNR make detecting and localizing events more challenging.

TABLE II
EXPERIMENT RESULTS.

Dataset	ER (20°)	F-score (20°)	LE_{CD} ($^\circ$)	LR_{CD}
STARSS23 (indoor dataset)	0.62	27.8%	27.0	44.3%
UNS-ESSE2023 (outdoor dataset)	0.71	39.1%	22.7	49.6%

The effect of the signal-to-noise ratio between the target event and ambience when detecting and localizing events is shown in Table III. The results show how the model's performance is highly affected by the different levels of SNR. Expectedly, the best detection and localization performance was achieved with the highest SNR levels.

TABLE III
INFLUENCE OF SIGNAL-TO-NOISE RATIO (SNR) BETWEEN THE TARGET EVENT AND THE AMBIANCE ON THE PERFORMANCE.

SNR [dB]	ER (20°)	F-score (20°)	LE_{CD} ($^\circ$)	LR_{CD}
[-10, -6]	0.91	21.1%	37.4	36.0%
[-6, -2]	0.85	28.6%	34.4	42.9%
[-2, 2]	0.70	39.4%	19.4	45.8%
[2, 6]	0.60	48.4%	18.9	59.9%
[6, 10]	0.51	54.8%	14.0	63.3%

As was mentioned in section II-C, the values of SPL have been collected before and after each mixture was recorded. To better understand the effect of background noise levels, table IV shows the detection and localization metrics for four ranges in the SPL levels. The SPL value for a mixture was calculated as the average value between before and after SPL values. The distribution of the SPL levels is concentrated around 72 dB. Ranges were selected to have roughly an equal amount of examples per range.

The table illustrates the effect of the environment conditions affecting the performance of the model. Having a busy road next to the recording scenario, increase the SPL values considerably, resulting in a more complex task. The effect of the source to microphone distance (5 vs 10 meters) was also

TABLE IV
INFLUENCE OF SOUND PRESSURE LEVEL OF THE NATURAL AMBIENCE DURING THE DATA COLLECTION ON THE PERFORMANCE, WHILE SNR BETWEEN -2 AND +2 DB.

SPL [dB]	ER (20°)	F-score (20°)	LE _{CD} (°)	LR _{CD}
[60,69]	0.56	53.0%	10.6	56.2%
[69,72]	0.65	42.3%	13.8	49.6%
[72,76]	0.74	36.7%	21.8	46.2%
[76,93]	0.80	27.0%	39.5	38.9%

evaluated, resulting in no difference on the performance of the model overall for all SNR conditions.

V. CONCLUSIONS

In this paper, we present the UNS Exterior Spatial Sound Events 2023 (UNS-ESSE2023) dataset, the first dataset with spatiotemporal information recorded in a field outdoors. The dataset includes SNR and SPL information for each of the recordings, and was recorded using two target to source distances (5 and 10 m). The benchmark model, based on the architecture of the baseline system used in DCASE22023, obtains a 39.1% F-score, showing a reasonable performance for the new exterior dataset. The complexity of the dataset is reflected in the degradation of the performance when only audio clips with high values of SNR are analyzed. Applications for monitoring urban spaces will benefit from the released dataset, where anomalous events have to be detected when different types of noise and level of it are present.

REFERENCES

- [1] Sharath Adavanne, Archontis Politis, and Tuomas Virtanen, "A multi-room reverberant dataset for sound event localization and detection," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, New York University, NY, USA, October 2019, pp. 10–14.
- [2] Kento Nagatomo, Masahiro Yasuda, Kohei Yatabe, Shoichiro Saito, and Yasuhiro 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)*. IEEE, 2022, pp. 156–160.
- [3] Mathilde Brousmiche, Jean Rouat, and Stéphane Dupont, "SECL-UMons database for sound event classification and localization," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 756–760.
- [4] Archontis Politis, Kazuki Shimada, Parthasaarathy Sudarsanam, Sharath Adavanne, Daniel Krause, Yuichiro Koyama, Naoya Takahashi, Shusuke Takahashi, Yuki Mitsufuji, and Tuomas Virtanen, "STARSS22: A dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events," in *Proceedings of the 8th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022)*, Nancy, France, November 2022, pp. 125–129.
- [5] Dragana Bajovic, Arian Bakhtiarnia, George Bravos, Alessio Brutti, Felix Burkhardt, Daniel Cauchi, Antony Chazapis, Claire Cianco, Nicola Dall'Asen, Vlado Delic, et al., "MARVEL: Multimodal extreme scale data analytics for smart cities environments," in *2021 International Balkan Conference on Communications and Networking (BalkanCom)*. IEEE, 2021, pp. 143–147.
- [6] Marco Crocco, Marco Cristani, Andrea Trucco, and Vittorio Murino, "Audio surveillance: A systematic review," *ACM Computing Surveys (CSUR)*, vol. 48, no. 4, pp. 1–46, 2016.
- [7] Jakob AbeBer, Marco Gotze, Stephanie Kuhnlenz, Robert Grafe, Christian Kuhn, Tobias ClauB, and Hanna Lukashevich, "A distributed sensor network for monitoring noise level and noise sources in urban environments," in *2018 IEEE 6th International Conference on Future Internet of Things and Cloud (FiCloud)*, 2018, pp. 318–324.
- [8] Sangeeta Srivastava, Dhrubojyoti Roy, Mark Cartwright, Juan P. Bello, and Anish Arora, "Specialized embedding approximation for edge intelligence: A case study in urban sound classification," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 8378–8382.
- [9] Alex Morehead, Lauren Ogden, Gabe Magee, Ryan Hosler, Bruce White, and George Mohler, "Low cost gunshot detection using deep learning on the raspberry pi," in *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, 2019, pp. 3038–3044.
- [10] N. P. García-de-la Puente, F. Fuentes-Hurtado, L. Fuster, V. Naranjo, and G. Piñero, "Deep learning models for gunshot detection in the albufera natural park," in *2023 31st European Signal Processing Conference (EUSIPCO)*, 2023, pp. 206–210.
- [11] Siniša Suzić, Nikola Simić, and Dragana Bajovic, "UNS-Exterior Spatial Sound Events 2023 (UNS-ESSE2023)," Mar. 2024, [Online]. Available: <https://doi.org/10.5281/zenodo.10792703>.
- [12] Kuba Lopatka, Jozef Kotus, and Andrzej Czyzewski, "Detection, classification and localization of acoustic events in the presence of background noise for acoustic surveillance of hazardous situations," *Multimedia Tools and Applications*, vol. 75, pp. 10407–10439, 2016.
- [13] Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra, "FSD50k: An open dataset of human-labeled sound events," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 30, pp. 829–852, dec 2021.
- [14] R. Smith, "An overview of the tesseract OCR engine," in *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, 2007, vol. 2, pp. 629–633.
- [15] 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*, vol. 13, no. 1, pp. 34–48, 2019.
- [16] Parthasaarathy Sudarsanam, Archontis Politis, and Kostantinos Drossos, "Assessment of self-attention on learned features for sound event localization and detection," in *Proceedings of the 7th Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE2021)*, Online, November 2021.
- [17] Kazuki Shimada, Yuichiro Koyama, Naoya Takahashi, Shusuke Takahashi, and Yuki Mitsufuji, "ACCDOA: Activity-coupled cartesian direction of arrival representation for sound event localization and detection," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 915–919.
- [18] Archontis Politis, Annamaria Mesaros, Sharath Adavanne, Toni Heittola, and Tuomas 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, 2020.