Proceedings of the 12th RSI International Conference on Robotics and Mechatronics (ICRoM 2024), Dec. 17-19, 2024, Tehran, Iran

Sound Source Localization in a 3D Multi-Surface Environment Using Learning-Based Methods

Soroush Esfahanian
Human and Robot Interaction Lab,
Electrical and Computer Engineering,
University of Tehran,
Tehran, Iran
soroushesfahanian@ut.ac.ir

Keyhan Rayati
Human and Robot Interaction Lab,
Electrical and Computer Engineering,
University of Tehran,
Tehran, Iran
keyhan.rayati@ut.ac.ir

Mehdi Tale Masouleh
Human and Robot Interaction Lab,
Electrical and Computer Engineering,
University of Tehran,
Tehran, Iran
m.t.masouleh@ut.ac.ir

***Abstract*—This study presents a three-dimensional localization of a ball drop in a multi-surface environment using cost-effective data acquisition devices, and proposes two learning-based methods to improve the baseline classical localization method based on GCC-PHAT-estimated TDOA values. The first proposed method employs a ball drop surface classification executed by Random Forest and XGBoost models to enhance localization performance, whereas the second method conducts region classification by using both basic and multi-branch neural network models, achieving improved and more efficient localization performance. A dataset of real-world sparse ball drop samples was collected in a controlled multi-surface environment for experimental execution, where results demonstrate that the two proposed methods improve the localization performance across various scenarios, with the region classification method significantly outperforming the baseline method, achieving a 17% reduction in the Mean Euclidean Distance error metric. Furthermore, the surface and region classification methods reduce the localization time compared to the baseline method by 50% and 97%, respectively. The obtained results reveal the potential of combining simple learning-based models with affordable microphone sensors for achieving an accurate localization of sparse audio samples in a multi-surface environment.**

***Index Terms*—Sound Source Localization, TDOA, Surface Classification, Region Classification, Multi-surface Localization, Sparse Audio Data, Learning-Based Models, Deep Neural Networks, GCC-PHAT**

I. Introduction

Sound Source Localization (SSL) refers to locating a sound source's relative position in a certain environment. This area of study has received a considerable amount of attention. Various fields, including speaker localization [1], audio surveillance [2], robotics [3], marine detection [4], military applications [5], etc., widely utilize this technology. Several classical algorithms have been proposed for SSL tasks throughout time, which can be classified into three major categories, namely, (a) Time Difference of Arrival (TDOA)-based methods, (b) methods based on maximizing the Steered Power Response (SPR), and (c) methods based on spectral estimations [6]–[8].

TDOA-based methods use the time difference in which the sound signal reaches different microphones to locate the sound source. These methods usually consist of two main steps. At first, the TDOA values between each pair of microphones are calculated. In the second step, the calculated TDOA values are employed by geometrical equations to locate the sound source. Generalized Cross-Correlation algorithm with PHAT normalization (GCC-PHAT) [9] is commonly used to measure the time difference between two recorded signals captured by a pair of microphones, producing relatively accurate and robust TDOA estimations. TDOA-based methods benefit from robustness to environment variations such as different microphone geometries, and can generate accurate TDOA estimations in low-noise environments, along with having an acceptable computational cost. However, multiple studies have shown that the presence of reverberation and noise in high values can have a significant impact on the performance of TDOA-based approaches [8], [10], [11]. To overcome this issue and develop models more robust to reverberation and low Signal-to-Noise Ratio (SNR) values, and through the impressive breakthrough of artificial intelligence models, several studies started to involve learning-based models and especially deep learning-based models in SSL tasks [5], [10]. A general overview of

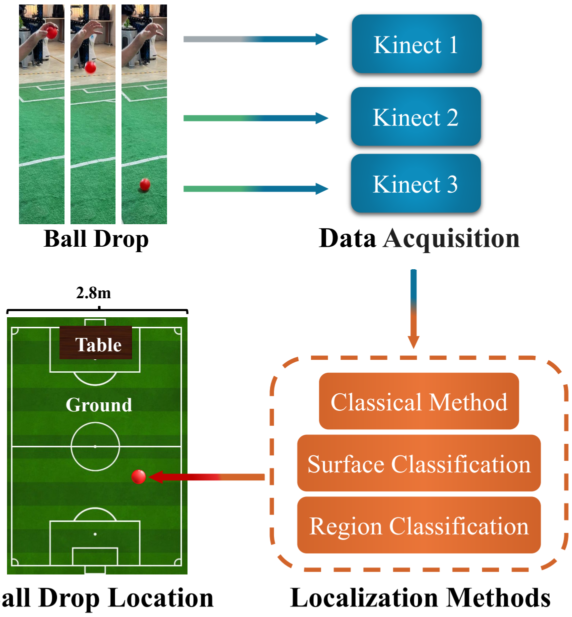


Fig. 1. An overview of the performed ball drop localization.

979-8-3315-2973-4/24/$31.00 ©2024 IEEE

deep learning models proposed for sound source localization is provided in [5].

The majority of SSL studies focus on providing the Direction of Arrival (DOA) of the sound source relative to the microphones, a representation that leaves the distance to the source unknown, which is critical for real-world applications [5], [12]. Additionally, most of the localization studies perform a 2-Dimensional (2D) localization, while 3-dimensional (3D) localization is of great demand in real-world applications [13]. The authors of [14] were among the first to employ a Convolutional Neural Network (CNN) for 3D localization. Recently, the attention towards 3D localization have been increasing and more research have porposed methods in this field [7], [13]. In localization tasks, proper audio sample preparation is essential, and most studies use dense recordings of continuous audio divided into smaller, informative samples [1], [7], [8], [11], [13]–[16]. However, in scenarios where the audio signal is generated in an instance, like a ball drop or a hand clap, the learning process becomes more challenging. Furthermore, the quality and setup of the microphones are crucial to achieving accurate localization, but high-quality arrays are not readily available to all researchers, making improvements with affordable setups more important for practical applications.

This study utilizes the Microsoft Kinect for Xbox 360, which is commonly used for sound and image data collection because of its affordability. A review and comparison of various publications on sound source localization with Kinect sensors is provided, as presented in Table I. In [17], increasing the distance between the Kinect's microphones enhanced 2D localization precision. A real-time learning-based architecture for binaural localization was introduced in [18] to predict the DOA of sound sources. The disassembly of the Kinect device to enhance microphone range for superior 2D localization was investigated in [19], whereas [20] represents one of the initial studies employing Kinect for 2D sound localization in an office setting. In [12], two Kinect devices were utilized for 2D localization via an encoder-decoder neural network; however, performance has deteriorated when transitioning from simulated to real-world data due to limited numbers of real-world samples.

TABLE I

Overview of studies on sound source localization.D and S denote dense and sparse samples, respectively. C and LB represent classical and learning-based methods, respectively.

Work	Sample Format	# of Surfaces	# of Dimensions	Output Pattern	Method
[12]	D	1	2D	x, y	LB
[17]	D	1	2D	x, y	C
[18]	D	1	3D	DOA	LB
[19]	S	1	2D	x, y	C
[20]	D	1	2D	x, y	C
Ours	**S**	**2**	**3D**	x, y, z	**C&LB**

In this study, localization of a ball drop in a 3D multi-surface environment, using cost-effective data acquisition devices is conducted, and several localization methods, especially simple learning-based models, are investigated to enhance the baseline localization performance. Initially, a classical SSL approach based on TDOA values of microphone pairs is implemented. Subsequently, simple learning-based models are employed to better distinguish between surfaces on which the ball is dropped. Lastly, the ability of learning-based methods, including deep neural networks and simple learning-based models, to perform a region classification task is leveraged to enhance the overall performance. An overview of the performed localization is illustrated in Figure 1. Unlike the majority of the works conducted in this field, where a great number of dense recorded samples are utilized, real-world sparse samples of ball drops are employed in this study, in which the ball is dropped on two different surfaces. The main goal of this work has been to specify and expand the extent to which the mentioned localization can be enhanced, by employing various methods for localization. The primary contributions of this paper are as follows: (1) A 3D multi-surface Cartesian localization of a sparse audio signal, i.e., a ball drop, using cost-effective data acquisition sensors is performed, and two learning-based methods are proposed to improve the localization performance, resulting in a satisfactory localization at last. To the best knowledge of the authors, the application of sparse audio samples in localization using cost-effective sensors is rarely documented in the literature. (2) The use of simple learning-based models, utilizing TDOA values, for surface and region classification tasks is investigated, and despite the limited number of real-world data, the utilized models manage to achieve notable classification performances, thereby, increasing the localization accuracy. Due the limited number of real-world data available, as collecting real-world samples is often cumbersome in SSL studies [12], training more complicated models, especially deep neural networks, is challenging. Multiple studies [8], [10], [12], report specific declines in performance when attempting to generalize to real-world data using deep neural networks. Therefore, this study investigates the application of simple learning-based models alongside deep neural networks, as these models are often shown to outperform neural networks when working with limited number of samples. (3) A thorough comparison between several localization methods and the impact of various learning-based models on the localization performance, including both simple learning-based models and deep neural networks, is conducted, and the effect of multiple factors on the described localization is examined. In addition, a comparison between different microphone configurations is conducted, and the optimal microphone setting for the designated environment is specified.

The remainder of the paper is organized as follows. Section II offers a detailed elucidation of the localization methods utilized in this study. Section III outlines the experimental setup for the gathering of real-world samples related to the ball drops. Section IV presents the results gathered, compares various methodologies, and examines the influence of different factors on localization performance. Lastly, Section V concludes the study.

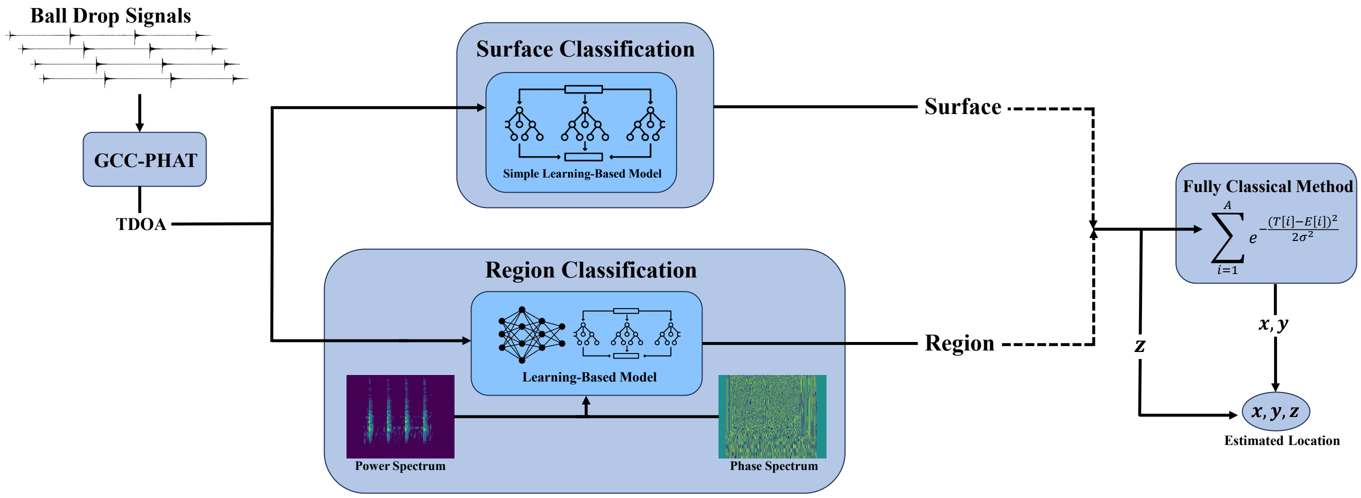


Fig. 2. An overview of the learning-based methods employed to enhance localization.

II. Methods

A. Fully Classical Method

In the first approach, sound source localization is conducted using a fully classical method. Two key steps are taken to ascertain the location of the ball drop: (1) TDOA estimation using the GCC-PHAT algorithm and (2) prediction of the sound source's cartesian coordinates using a Gaussian distribution likelihood algorithm. The initial stage involves calculating the TDOA values for each pair of microphones within a single Kinect device. For this purpose, the commonly employed GCC-PHAT algorithm is utilized. Let x_1 and x_2 represent signals recorded by a pair of microphones, each holding N discrete sample values. The GCC-PHAT algorithm is then defined as follows:

$$R(\tau) = \frac{1}{N} \sum_{i=0}^{N-1} \frac{X_1[i] X_2^*[i]}{|X_1[i] X_2^*[i]|} \tag{1}$$

where X_1 and X_2 represent the Discrete Fourier Transform (DFT) of x_1 and x_2, respectively. The $(.)^*$ operator represents the complex conjugate, and τ denotes the delay values between two captured signals. The TDOA value between the two captured signals is then equal to:

$$\text{TDOA}_{x_1,x_2} = \arg_\tau \big(\max\{R(\tau)\}\big) \tag{2}$$

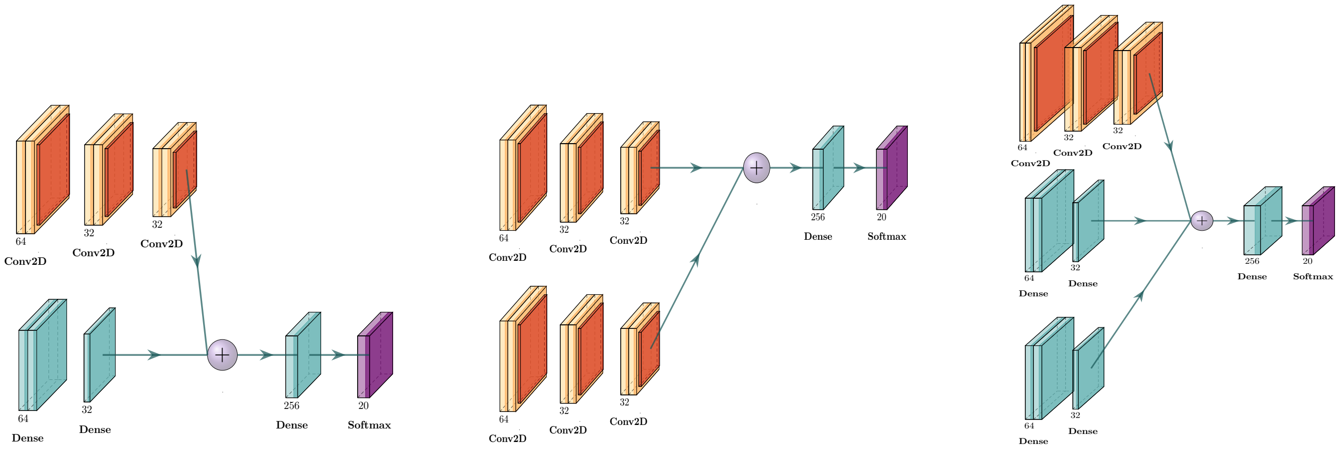
For each Kinect device, the TDOA values for each pair of microphones are computed. A single Kinect device contains four microphones, resulting in six TDOA values per Kinect for a single recorded sample. The obtained TDOA values are then fed into the so-called likelihood algorithm based on the Gaussian distribution to determine the position with the highest probability of being the location of the sound source. In order to accomplish this, a 3D grid array is iterated initially, where each element corresponds to an individual point in the provided environment, with the environment resolution set to one centimeter. To the end of improving the accuracy of localization and minimize overall computation, the locations which cannot be the potential source's location, i.e., those above the considered ground surface and under the table surface, are initialized with a negative value in the grid array. This indicates that the corresponding position is invalid and cannot be the predicted sound source's location. For the sake of predicting the ball drop's position, valid elements in the grid are iterated and their Likelihood based on the Gaussian distribution is calculated, where the true TDOA values, along with the GCC-PHAT-estimated TDOA values are employed. There are 6 different TDOA values for each Kinect device, for a total of 18 TDOA values per point in the grid. Let T and E be arrays of length 18, containing true and GCC-PHAT-estimated TDOA values for a given point (x, y, z), respectively. The Gaussian Likelihood for the position (x, y, z), $\mathcal{L}$, is calculated as follows:

$$\mathcal{L} = \sum_{i=1}^{A} e^{-\frac{(T[i]-E[i])^2}{2\sigma^2}} \tag{3}$$

where A represents the total number of TDOA values for each point, i.e, 18, and σ denotes the standard deviation of the distribution. The division by the denominator is omitted in the actual procedure for the sake of simplification, as it would not alter the point with the highest probability. Ultimately, the point with the highest obtained probability in the grid is returned as the source's estimated location.

B. Learning-based Surface Classification

This approach employs simple learning-based models to ascertain whether the ball has landed on the 'Ground' or 'Table' surface, as illustrated in Figure 2. Subsequently, the classical localization algorithm is only employed on the anticipated surface, enhancing accuracy and decreasing computation time by constraining the search area within the 3D grid. Models, namely Random Forest, Gradient Boosting, SVM, and



(a) Two-Branch MLP and CNN model

(b) Two-Branch CNN model

(c) Three-Branch MLP and CNN model

Fig. 3. Architectures of the multi-branch deep learning models employed for region classification.

XGBoost classifiers are employed for surface prediction, using GCC-PHAT-estimated TDOA values, which are relatively unpopular in the similar tasks [21]. The models were trained using the TDOA values of the provided train samples, and their performance was assessed on test samples. The procedure of collecting real-world samples is outlined in Section III.

C. Learning-based Region Classification

In the final approach, the environment was segmented into 20 discrete regions—18 ground sections and 2 regions for the the table. A visual representation of the division of the environment is depicted in Fig. 7(c). As depicted in Fig. 2, in this method, learning-based models are employed to predict the ball drop's region, after which the classical localization method iterates through the predicted region to ascertain the final position of the ball drop. This method substantially decreases the search area, enhancing both localization precision and computational efficiency relative to the prior methods. For this task, models similar to those used in the last method were employed. Additionally, four Deep Neural Network (DNN) architectures were constructed to evaluate their efficacy in this task. Nevertheless, owing to the limited number of real-world samples and sparsity of the recorded data, numerous DNN architectures from other SSL researches were not entirely applicable, and only models possessing characteristics appropriate for this context were chosen in this method.

The first model is a three-branch Multi-Layer Perceptron (MLP), with each branch corresponding to the TDOA data from a single Kinect device, where each branch comprises two dense layers. The next three models are influenced by the architecture proposed in [15]. The second model is consisted of an MLP branch along with a CNN branch, with TDOA data of the Kinect devices and power spectra of the samples, used as the input features for the corresponding branches. The third model includes two CNN branches which utilize the power and phase spectra of the samples as input features, respectively. Lastly, the fourth model is composed of two MLP branches, corresponding to the TDOA data of the second and third Kinect devices, respectively, along with a CNN branch fed with power spectra of the samples. The detailed architectures of the models inspired form the the architecture in [15] are depicted in Fig. 3. In all of the employed models, the branches are concatenated and fed into two final dense layers, employed for the prediction of the regions. The models underwent training utilizing optimization approaches and monitoring mechanisms to facilitate effective learning and avoid overfitting during the training process.

III. Experimental Setup

This section describes the environment setup for preparing real-world samples of the ball drops. Since numerous studies have reported a decline in the performance of SSL models when attempting to generalize from simulated data to real-world data [8], [10], [12], only real-world data samples are utilized in this study, employed for training and evaluation. As depicted in Figs. 1 and 4, a confined space of 2.82×4×0.75 meters within a relatively large laboratory room is dedicated for localizing the ball drops, with the bottom left corner of the environment serving as the origin of the coordinate system. A 1.1×0.7×0.75 table is placed at the top center of the dedicated area to form our second surface onto which the ball is dropped. A layer of artificial grass was used as the environment's ground. Initially, three Microsoft Kinect devices, namely K1, K2, and K3, were utilized to capture the audio signals and obtain the optimal configuration, with their centers positioned at (2.82, 2, 0.75), (1.4, 0, 0.05), and (0, 2, 0.75) meters, respectively. The same notation is used in the conducted experiments to refer to each corresponding Kinect device. Each Kinect device contains 4 synchronized small-scale linear microphones with a sampling rate of 16 kHz, placed with varying spacing, as illustrated in Figure 5. Note that in this study, the Kinect devices are not synchronized themselves, meaning the TDOA information between two separate Kinect

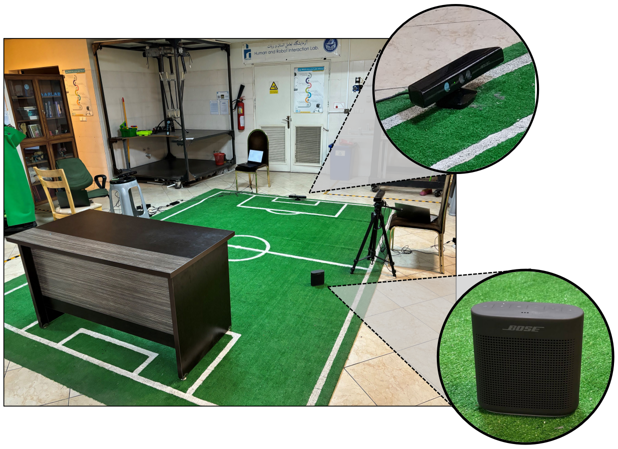


Fig. 4. Experimental setup of the environment.

devices cannot be used as additional localization information, which further complicates the localization task.

For real-world sample collection, the provided environment was first separated into ten sections, eight for the ground surface and two for the table. Subsequently, in each section, the ball was dropped at the center of the section for four times without bouncing, and the generated audio was recorded using a smartphone placed 5 cm away, yielding ten distinct ball drop recordings. This was done to ensure that the ball drops are more accurately represented in the final audio samples based on the location they are taken. In the second stage, coordinates with 20 cm distances were determined, starting at the origin, yielding a total of 315 points, 28 of which belong to the table and the rest to the ground. To play the recorded ball drop audio of each corresponding section from the specified locations, the Bose SoundLink Color Bluetooth® Speaker II was utilized. The provided setup for localization is depicted in 4.

The speaker was placed at each of the 315 specified locations and played the corresponding recorded audio. Among the 315 prepared samples, 25 were chosen as test samples, and 290 were used to train the learning-based models. The 25 test locations were handpicked to ensure a nearly uniform distribution of locations among the test samples, allowing for a better evaluation of each localization method. Additionally, in order to prepare the samples for feature extraction conducted during the pre-training phase of the DNN models, the recorded samples were initially denoised using the spectral subtraction algorithm and then zero-padded to achieve a length of 145,000 sample values, corresponding to 9.0625 seconds for each sample. The average SNR value of the samples and the RT60 reverberation time of the experimented environment was measured at 15.29 dB and 0.6 seconds, respectively.

IV. Results

A. Evaluation Metrics

In order to evaluate the SSL performance of the above methods and multiple microphone configurations, the Mean Euclidean Distance (MED) metric, based on the Multiple Object Tracking Precision (MOTP) metric, initially developed under the CHIL project [22], is adopted. Let $(x_{a_i}, y_{a_i}, z_{a_i})$ and $(x_{p_i}, y_{p_i}, z_{p_i})$ denote the actual and predicted Cartesian coordinates of the ball drop's location, in the recorded sample i, respectively. The utilized MED measure is then defined as:

$$\text{MED} = \frac{1}{N_r}\sum_{i=0}^{N_r}\sqrt{(x_{a_i}-x_{p_i})^2+(y_{a_i}-y_{p_i})^2+(z_{a_i}-z_{p_i})^2} \tag{4}$$

where N_r represents the total number of recorded ball drop samples, separated for evaluation. Furthermore, to better understand each model's performance in localization of the ball drops in each Cartesian axis, the Mean Absolute Error (MAE) metric is utilized to evaluate the localization error in each axis. It should be noted that the axes in the results are not relative to each Kinect device and represent the axes across the defined SSL environment.

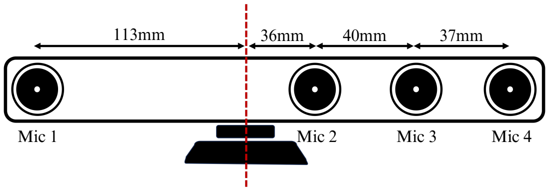


Fig. 5. Microphone positioning of the microsoft kinect device.

B. Results of the Fully Classical Method

In Table II, the first line of each row represents the localization results of the baseline TDOA-based classical method across various microphone configurations. The localization of the ball drops for the test samples is performed for each set of Kinect devices, employing a distinct combination of the three devices for localization each time. Initially, localization is conducted using solely one of the three devices. Subsequently, two Kinect devices are utilized, followed by the use of all three devices for the localization of the ball drop.

As reported in Table II, the localization accuracy is notably unsatisfactory when employing a single device exclusively, with the minimum MED of 118.14 cm, which is achieved when using K2 only. Interestingly, the MAE values of localization using a single Kinect device suggest that each device produces the smallest localization error along its parallel axis. In other words, the angular error of the estimated locations would be less than the corresponding MED error, indicating that each Kinect device would perform better in DOA estimation compared to the estimation of the precise Cartesian coordinates of the sound source. This is supported by the mathematical limitation that a single linear microphone array can only predict one single angle regarding the sound source [2].

The use of two Kinect devices for localization has significantly decreased the localization error, with the best SSL performance achieved by employing K2 and K3 as data acquisition devices. When using K2 and K3, the classical method has accurately identified the surface of the ball drop, yielding an MAE of zero along the Z axis. The use of all three devices for localization has resulted in a decline in localization

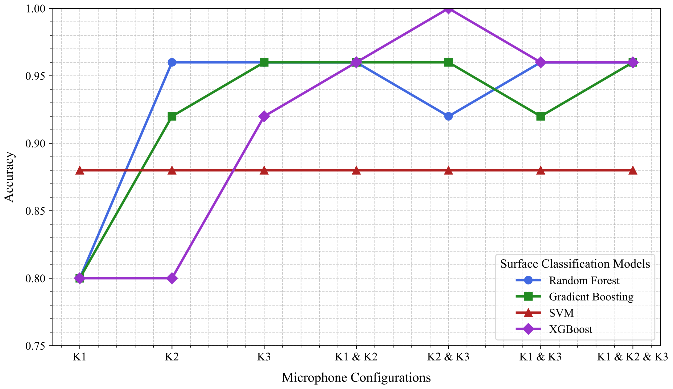


Fig. 6. Accuracy of the employed simple learning-based models for surface classification across various microphone configurations.

performance when compared to the utilization of K2 and K3 alone. Nonetheless, the localization performance has improved relative to other configurations where two Kinect devices were employed.

C. Results of the Learning-based Surface Classification method

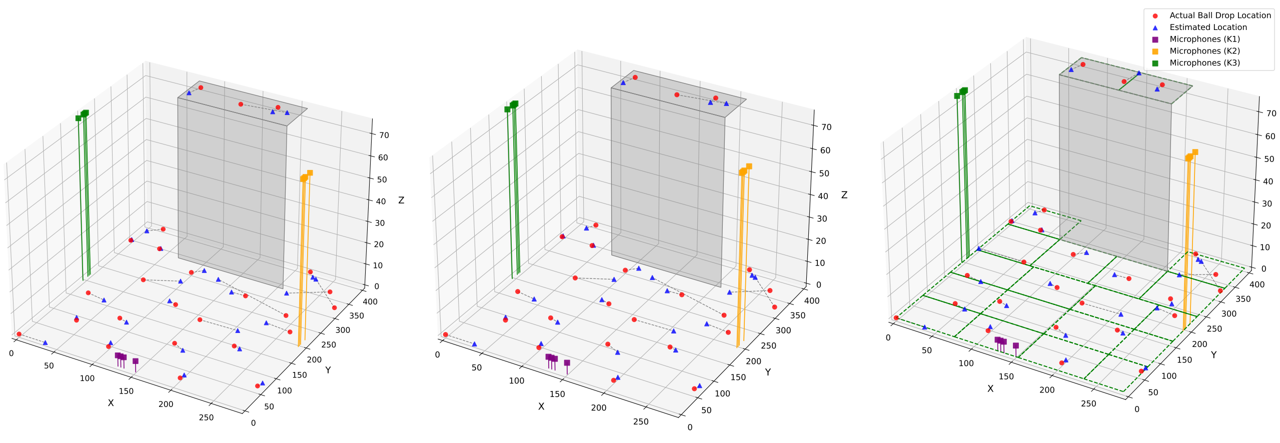
For the classification of the surface onto which the ball is dropped, four simple learning-based models, namely Random Forest, Gradient Boosting, SVM, and XGBoost classifiers, are employed. The results of the surface classification based on the obtained classification accuracy of each model across different microphone configurations are illustrated in Figure 6. Each model is trained on 290 real-world training samples, of which 25 belong to the table and the rest belong to the ground. In order to assess the models' accuracy, 25 training samples are provided, with 3 from the table class and 22 from ground class. As depicted in Fig. 6, the performance of the SVM classifier has remained consistent across different configurations, standing at 88% accuracy, which arises from the misclassification of all three table samples and failure in surface classification. The performance of the other three models has generally enhanced when utilizing more Kinect devices for classification, and the XGBoost model has achieved perfect surface classification of the test samples when employing the data from K2 and K3. As the employed simple learning-based models were able to classify the surface with notable accuracy, the performance of the more complex models, such as neural networks, is not investigated for this particular task.

TABLE II
LOCALIZATION PERFORMANCE OF THE FULLY CLASSICAL AND SURFACE CLASSIFICATION METHODS FOR DIFFERENT SETUPS OF KINECT DEVICES.

Configuration	MAE on X (cm)	MAE on Y (cm)	MAE on Z (cm)	MED (cm)
K1 only	43.24	98.56	15.0	119.03
K1 only + surface cls.	43.4	94.44	9.0	113.6
K2 only	94.4	54.28	15.0	118.14
K2 only + surface cls.	86.8	53.92	3.0	107.93
K3 only	106.4	49.36	9.0	123.26
K3 only + surface cls.	97.08	45.65	3.0	112.28
K1, K2	41.8	29.64	3.0	56.8
K1, K2 + surface cls.	41.8	29.64	3.0	56.8
K2, K3	**21.48**	**14.32**	**0.0**	**28.4**
K2, K3 + surface cls.	**21.48**	**14.32**	**0.0**	**28.4**
K1, K3	38.28	37.12	3.0	59.7
K1, K3 + surface cls.	38.28	37.12	3.0	59.7
K1, K2, K3	27.2	20.64	3.0	38.5
K1, K2, K3 + surface cls.	27.2	20.64	3.0	38.5

In order to evaluate the localization performance of the second approach, which utilizes the predicted surface, the best classification model for each microphone configuration according to Fig. 6 is employed as the surface classifier for the corresponding configuration. For models with the same accuracy for a single configuration, the model which generated the least MED error values is selected. The localization results using the second proposed method are reported in the second line of each row in Table II. The application of surface classification models has improved localization performance when using data from a single Kinect device, as the classification model has successfully identified the surface of certain samples for which the classical method failed to do so. Utilizing multiple Kinects has not improved localization through the surface classification model, as the number of misclassified surface samples remains unchanged when utilizing two or all three Kinect devices. However, this method reduces computation time for localization, as the classical algorithm iterates exclusively through the predicted surface of the classification model rather than the entire grid, leading to comparatively faster localization. The average localization time per each test sample using the K2 and K3 setup for the classical method was measured at 29.8 seconds, while the duration for the surface classification method was equal to 14.7 seconds, representing a 50% reduction in calculation time while preserving localization accuracy. As reported in Table II, utilizing K2 and K3 solely, has resulted in better localization compared to other examined configurations. Unexpectedly, the inclusion of K1 as the third data acquisition device has led to a decline in localization performance, regardless of the application of surface classification models.

In [17], the authors have reported that increasing the distance between the sensors of the Kinect device increases the localization accuracy. Furthermore, in [19], disassembling the Microsoft Kinect device and increasing the distance between its microphones to cover the entire width of the area, yielded satisfactory localization performance in a 2D environment. Consequently, it was anticipated that the TDOA values of the closely placed microphones could introduce bias in the localization algorithms, thereby decreasing the localization accuracy. Therefore, as the last step in determining the optimal configuration for ball drop localization within the specified environment, various combinations of TDOA values acquired from the closest microphone pair in each Kinect device in the K2 and K3 configuration, were eliminated to investigate their impact on the localization process. Each Kinect device contains 4 synchronized microphones, with the closest pair positioned within a distance of 3.7 cm, as depicted in Fig. 5. In this experiment, the TDOA value between the two closest microphones in each Kinect device is referred to as TClose$j$ for convenience, where j denotes the Kinect device's number. Table III represents the impact of eliminating each specified TDOA value from the calculations in each localization algorithm. The removal of TClose2 has led to a 0.5 cm



(a) Classical

(b) Surface classification

(c) Region classification

Fig. 7. Visualization of the result of each method's localization.

TABLE III

K2 & K3 SETUP MED ERROR VALUES (CM) CORRESPONDING TO THE REMOVAL OF TDOA OF THE CLOSEST MICROPHONE PAIRS.

Method	Classical	Surface cls. + Same Input	Surface cls. + Baseline input
All TDOAs included	28.4	28.4 XGBoost, 100%	28.4 XGBoost, 100%
−TClose2	30.82	30.82 XGBoost, Gradient Boosting, 96%	**27.9** XGBoost, 100%
−TClose3	36.69	34.07 Gradient Boosting, 96%	31.17 XGBoost, 100%
−TClose2, −TClose3	33.48	33.4 XGBoost, Gradient Boosting, 96%	30.59 XGBoost, 100%

TABLE IV

PERFORMANCE OF THE REGION CLASSIFICATION MODELS IN TERMS OF THE CLASSIFICATION ACCURACY AND THE RESULTED MED VALUE (CM).

Model	Region classification Accuracy (%)	MED
Simple learning-based models		
Random Forest	**92%**	**23.4**
Gradient Boosting	76%	28.9
SVM	64%	32.8
XGBoost	80%	26.1
Neural Networks		
3branch, MLP/MLP/MLP	64%	**27.86**
2branch CNN/MLP	48%	72.35
3branch CNN/MLP/MLP	72%	27.95
2branch CNN/CNN	**76%**	33.48

reduction in the MED error value when the classification model utilizes all TDOA data as input, compared to when no TDOA values are excluded from localization computations. The removal of TClose values from the input features of the surface classification models has reduced their accuracy in comparison to when all TDOA data are utilized as input features. In this experiment, the elimination of other TDOA values was also examined, but it did not enhance localization and is hence not reported to maintain clarity in presentation.

According to the obtained MED values in Tables II and III, the optimal configuration for multi-surface ball drop localization in the specified environment is achieved by employing K2 and K3 exclusively as data acquisition devices and utilizing the XGBoost classifier for surface classification, incorporating all TDOA values, while excluding TClose2 during the classical phase of the localization. The described optimal configuration will serve as the localization setting for the classical method's phase performed in the region classification method.

D. Results of the Learning-based Region Classification method

In the last method of the performed SSL, the designated environment is partitioned into 20 regions, and a region classification model is employed to determine the region wherein the ball has been dropped. The traditional method is then employed solely within the predicted region to identify the location with the highest Likelihood as the ball's drop location. As aforementioned, several learning-based models are employed for this task, including four multi-branch neural networks and four simple learning-based models, similar to those employed for surface classification. Table IV represents a comparison of the region classification models, where the properties of the obtained optimal configuration are utilized for applying the classical method within each predicted region. Among the employed models, the simple learning-based models have mostly surpassed the DNN models and the three-branch MLP model regarding classification accuracy, and the random forest model and the XGBoost classifier have produced a lower MED error value in comparison to the neural network models. Among the utilized models, the random forest stands out by achieving 92% accuracy in region classification, with an MED of 23.4 cm, which is 4.5 cm lower than the previously achieved minimum MED value, and offers the best localization performance among the methods employed in this study. The three-branch MLP model, utilizing the TDOA values from all three Kinect devices, has surpassed the multi-branch DNN models regarding the localization's MED value. However, the two-branch CNN/CNN model and the three-branch CNN/MLP/MLP model have achieved superior region classification accuracy. This is due to the fact that in certain correctly classified regions, the distance to the true source location from the position given by the classical method is greater than the distance to a position predicted in some of the adjacent misclassified regions; therefore, certain misclassifications have resulted in enhanced localization performance. The average computation time of localization per each test sample was measured at 0.71 seconds in the region classification method, which indicates a significant decrease in computation time compared to other two methods as the search area is restricted to a single region.

Having satisfactory localization accuracy and a notable low computation time, the region classification method stands out as the most efficient localization method examined in this study. The visual representation of the localization performed in each of the methods is depicted in Fig. 7. As illustrated in the foregoing figure, almost all of the samples, except those positioned on the ground at the right side of the table—where the audio signal is obstructed by the table before reaching K1 and K3—exhibit satisfactory localization. This indicates the efficacy of utilizing simple learning-based models for localization tasks, even with affordable sensors, as performed in this study.

V. Conclusion

This study conducted a three-dimensional localization of a ball drop in a multi-surface environment using small-scale, cost-effective sensors, and two learning-based methods, surface and region classification, were proposed to enhance the performance of the baseline method. Real-world data was gathered from drops on Table and Ground surfaces, and despite the limited number of data, the proposed methodologies improved both precision and computational time of the localization, with the region classification approach outperforming the other two methods, reducing the localization error and computation time of the baseline method by 17% and 97%, respectively, demonstrating the method's potential for application in real-world scenarios. Due to the limited number of real-world data, simple models such as random forest and XGBoost classifiers surpassed deep neural networks in both of the classification tasks, indicating the capability of simple learning-based models in such contexts, where the number of real-world samples is limited. In addition, the optimal microphone configuration of the designated environment was obtained, providing a reference for similar cases to be utilized in further experiments. As ongoing work, the authors aim to integrate simple learning-based models with deep neural networks to attain precise localization in real-world tasks, such as those in the field of robotics. Employing audio data instead of visual data [23] for localization tasks, may allow robots detect objects in low-visibility environments, therefore enhancing functionality.

References

[1] Fabio Vesperini, Paolo Vecchiotti, Emanuele Principi, Stefano Squartini, and Francesco Piazza. Localizing speakers in multiple rooms by using deep neural networks. *Computer Speech & Language*, 49:83–106, 2018.

[2] Marco Crocco, Marco Cristani, Andrea Trucco, and Vittorio Murino. Audio surveillance: A systematic review. *ACM Computing Surveys (CSUR)*, 48(4):1–46, 2016.

[3] Caleb Rascon and Ivan Meza. Localization of sound sources in robotics: A review. *Robotics and Autonomous Systems*, 96:184–210, 2017.

[4] Darryl Franck Nsalo Kong, Chong Shen, Chuan Tian, and Kun Zhang. A new low-cost acoustic beamforming architecture for real-time marine sensing: Evaluation and design. *Journal of Marine Science and Engineering*, 9(8):868, 2021.

[5] Pierre-Amaury Grumiaux, Sran Kitić, Laurent Girin, and Alexandre Guérin. A survey of sound source localization with deep learning methods. *The Journal of the Acoustical Society of America*, 152(1):107–151, 2022.

[6] DN Kumar. Study of sound source localization using music method in real acoustic environment. *International Journal of Electronics Engineering Research*, 9(4):545–556, 2017.

[7] Jiajun Yan, Wenlai Zhao, Yue Ivan Wu, and Yingjie Zhou. Indoor sound source localization under reverberation by extracting the features of sample covariance. *Applied Acoustics*, 210:109453, 2023.

[8] JAKUB PEKÁR. Sound source localization from a microphone array.

[9] Charles Knapp and Glifford Carter. The generalized correlation method for estimation of time delay. *IEEE transactions on acoustics, speech, and signal processing*, 24(4):320–327, 1976.

[10] Gabriel Jekateryńczuk and Zbigniew Piotrowski. A survey of sound source localization and detection methods and their applications. *Sensors*, 24(1), 2024.

[11] Ran Lee, Min-Seok Kang, Bo-Hyun Kim, Kang-Ho Park, Sung Q Lee, and Hyung-Min Park. Sound source localization based on gcc-phat with diffuseness mask in noisy and reverberant environments. *IEEE Access*, 8:7373–7382, 2020.

[12] Guillaume Le Moing, Phongtharin Vinayavekhin, Tadanobu Inoue, Jayakorn Vongkulbhisal, Asim Munawar, Ryuki Tachibana, and Don Joven Agravante. Learning multiple sound source 2d localization. In *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6. IEEE, 2019.

[13] Yunjie Zhao, Yansong He, Hao Chen, Zhifei Zhang, and Zhongming Xu. Three-dimensional grid-free sound source localization method based on deep learning. *Applied Acoustics*, 227:110261, 2025.

[14] Juan Manuel Vera-Diaz, Daniel Pizarro, and Javier Macias-Guarasa. Towards end-to-end acoustic localization using deep learning: From audio signals to source position coordinates. *Sensors*, 18(10):3418, 2018.

[15] Eric L Ferguson, Stefan B Williams, and Craig T Jin. Sound source localization in a multipath environment using convolutional neural networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2386–2390. IEEE, 2018.

[16] 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.

[17] Belgacem Douaer, Farid Ykhlef, and Fayçal Ykhlef. Experimental investigation into the influence of the distance between microphones for 2d real-time sound source localization using gcc-phat technique. In *Advances in Computing Systems and Applications: Proceedings of the 4th Conference on Computing Systems and Applications*, pages 354–362. Springer, 2021.

[18] Alessia Saggese, Nicola Strisciuglio, Mario Vento, and Nicolai Petkov. A real-time system for audio source localization with cheap sensor device. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–7. IEEE, 2017.

[19] Belgacem Douaer. A laboratory experiment for real-time sound source localization using microsoft kinect for xbox 360. 2022.

[20] Ling Pei, Liang Chen, Robert Guinness, Jingbin Liu, Heidi Kuusniemi, Yuwei Chen, Ruizhi Chen, and Stefan Söderholm. Sound positioning using a small-scale linear microphone array. In *International Conference on Indoor Positioning and Indoor Navigation*, pages 1–7. IEEE, 2013.

[21] Annamaria Mesaros, Toni Heittola, Tuomas Virtanen, and Mark D Plumbley. Sound event detection: A tutorial. *IEEE Signal Processing Magazine*, 38(5):67–83, 2021.

[22] Jose Velasco, Daniel Pizarro, and Javier Macias-Guarasa. Source localization with acoustic sensor arrays using generative model based fitting with sparse constraints. *Sensors*, 12(10):13781–13812, 2012.

[23] Keyhan Rayati, Amirhossein Feizi, Alireza Beigy, Pourya Shahverdi, Mehdi Tale Masouleh, Ahmad Kalhor, and Wing-Yue Geoffrey Louie. Real-time imitation of human head motions, blinks and emotions by nao robot: A closed-loop approach. In *2023 11th RSI International Conference on Robotics and Mechatronics (ICRoM)*, pages 794–800. IEEE, 2023.