electronics

MDPI

Article

Simultaneous Localization of Two Talkers Placed in an Area Surrounded by Asynchronous Six-Microphone Arrays

Toru Takahashi *, Taiki Kanbayashi and Masato Nakayama *

Graduate School of Engineering, Osaka Sangyo University, 3-1-1, Nakagaito, Daito 574-0013, Japan; s24mh01@ge.osaka-sandai.ac.jp

* Correspondence: takahashi@ise.osaka-sandai.ac.jp (T.T.); nakayama@ise.osaka-sandai.ac.jp (M.N.); Tel.: +81-6-875-3001 (M.N.)

Abstract: If we can understand dialogue activities, it will be possible to know the role of each person in the discussion, and it will be possible to provide basic materials for formulating facilitation strategies. This understanding can be expected to be used for business negotiations, group work, active learning, etc. To develop a system that can monitor speech activity over a wide range of areas, we propose a method for detecting multiple acoustic events and localizing sound sources using an asynchronous distributed microphone array arranged in a regular hexagonal repeating structure. In contrast to conventional methods based on sound source direction using triangulation with microphone arrays, we propose a method for detecting acoustic events and determining sound sources from local maximum positions based on estimation of the spatial energy distribution inside the observation space. We evaluated the conventional method and the proposed method in an experimental environment in which a dialogue between two people was simulated under 22,104 conditions by using the sound source signal convolving the measured impulse response.We found that the performance changes depending on the selection of the microphone array used for estimation. Our finding is that it is best to choose five microphone arrays close to the evaluation position.

Keywords: distributed microphone arrays; asynchronous processing; acoustic event detection; sound source localization; spatial energy distribution; beamforming

Academic Editor: Valeri Mladenov

Received: 10 January 2025
Revised: 30 January 2025
Accepted: 7 February 2025
Published: 12 February 2025

Citation: Takahashi, T.; Kanbayashi, T.; Nakayama, M. Simultaneous Localization of Two Talkers Placed in an Area Surrounded by Asynchronous Six-Microphone Arrays. *Electronics* **2025**, *14*, 711. https://doi.org/10.3390/electronics14040711

1. Introduction

In recent years, various application systems of sound information processing technology have been proposed [1–4], such as speech dialogue recognition, detection of animal activity sounds, and detection of drone flight sounds. Many systems like these use acoustic event detection and sound source localization in the front end. Therefore, if we can expand the observation space for acoustic event detection and sound source localization, it is expected that many of these applied systems can be applied to more sites than are currently available. In general, acoustic event detection often refers to detecting the time when the event occurred. However, acoustic event detection also involves detecting the spatial coordinates of the acoustic event that occurred. In other words, it is important to simultaneously detect the time of occurrence of an acoustic event and estimate the location of the sound source.

There are several methods to achieve acoustic event detection and sound source localization. The main ones are methods using distributed microphones, methods using a microphone array, and methods using distributed microphone arrays. When using distributed microphones, acoustic event detection and sound source localization can be

achieved simultaneously over a wide range of observation spaces, but the spatial resolution of sound source localization is low, and information that indicates only that an acoustic event was detected near the placed microphone is obtained. Therefore, ambiguity remains in spatial information. Although the method using a microphone array can realize sound source detection and sound source localization with high accuracy, it has the problem that the observation space is narrow and it is not possible to construct a system that targets a wide observation space. To solve these problems, methods using distributed microphone arrays can cover a wide space and also realize sound source localization with relatively high resolution. This method estimates the sound source arrival direction from two microphone arrays and localizes the sound source based on triangulation, but it is difficult to improve the estimation accuracy due to the limited accuracy of triangulation. We propose a new method for acoustic event detection and sound source localization using a distributed microphone array.

We describe a method for simultaneously detecting multiple acoustic events scattered within a space from the observed signals by distributing multiple microphone arrays and estimating their source positions. Microphone array technology is a technology that groups multiple microphone elements and uses them as one virtual microphone. This is a method that allows you to control the directivity of a virtual microphone using software control. The observation space targeted by our method is a wide area that is several tens of meters or more in each axis. Our method is an elemental technology for a system that analyzes and visualizes dialogue activities by distributing multiple microphone arrays on the ceiling of a large hall. The reason for installing the microphone array on the ceiling is to prevent it from affecting the propagation of sounds emitted during dialogue activities. Many such systems have been proposed in the last few decades [5–8], but the problem is that they do not cover a sufficient area. Figure 1 shows the microphone array arrangement in the system. The orange circle marks in the figure represent the microphone array. The microphone array is arranged in a triangular repeating structure on the ceiling. We propose a solution to the problem of detecting the presence or absence of people's speech and estimating the location of speech from multichannel audio signals observed with these microphone arrays. If this problem can be solved, it will be possible to estimate which conversation group a speaker belongs to based on the proximity of the speaker's utterance position.

The area of people's activities, i.e., the observation space, is divided into a repeating pattern of equilateral triangles with sides of 1.5 m, and microphone arrays are placed at the vertices of these equilateral triangles. As the idea of dividing the observation space is used in the field of sound source tracking to reduce the amount of calculation [9], we also use the idea. We show the evaluation results using acoustic event detection and sound source localization when two speakers simultaneously talk with each other inside a regular hexagon formed by six equilateral triangles around one microphone array position.

A large number of microphones are required to collect the acoustic events in such a wide observation space, to detect the events, and to localize their source position. The reason is that the distance range captured by a single microphone is limited. It is possible to capture acoustic events occurring within the observation space by distributing many microphones. However, the problem is that the system scale becomes too large as the number of microphones increases. The wiring length might be longer to distribute the microphones over a wide area. The longer the wiring between the microphone and the analog-to-digital converter, the more noise is picked up on the route. Connecting many microphones requires analog-to-digital converters with a large number of channels, which increases system complexity and installation costs. Furthermore, typical acoustic event detection and sound source localization methods are based on the assumption that observations are captured by multichannel synchronized recording. Due to the limitation

of the number of channels that can be implemented as a synchronous recording device, it is difficult to distribute a sufficient number of microphones over the wide observation space. The reason why our proposed method can handle a wide observation space is that it employs an architecture that can separate the microphone array part that directly handles the observed sound and the part that actually estimates the sound source position.

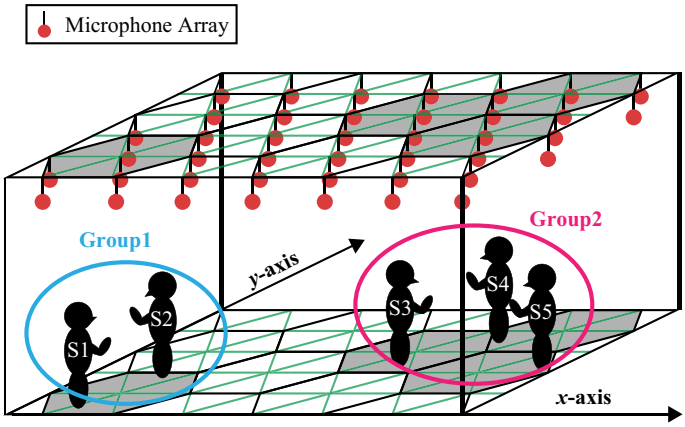

Figure 1. An overview of a system that visualizes dialogue activities by distributing multiple microphone arrays on the ceiling of a large hall.

In this article, we review related works first. Then, we describe the baseline method in Section 3. In Section 4, we describe the conventional method, which is an improved version of the baseline method. Then, we propose a new method for acoustic event detection and sound source localization based on spatial energy distribution in Section 5. In Section 6, we show the evaluation results of acoustic event detection accuracy and sound source localization error for the conventional method and the proposed method. We describe the case of one sound source and the case of two sound sources. We discuss our method in Section 7. Finally, we conclude this article.

2. Literature Review

In applications such as the one targeted by this research, there is a problem in that the number of microphones used for observation exceeds the number that can be achieved in synchronization. This problem becomes more serious as the observation space becomes wider. In this section, we introduce conventional research that uses a wide area as the observation space and conventional research that uses a narrow area as the observation space.

To attack this problem, several methods [1–4] have been proposed that use multiple microphone arrays to detect acoustic events occurring in a wide observation space and to localize sound sources. These studies deal with acoustic event detection and sound source localization problems over a wide observation space, such as outdoors. These methods do not require synchronized recording between multiple microphone arrays. These methods estimate the direction of arrival of the sound source [10–15] independently for each microphone array, and by collecting the information, the position of the sound source can be estimated based on the principle of triangulation. This feature increases the flexibility of the microphone array arrangement.

Sumitani, et al. [1] and Gabriel, et al. [2] attempted to detect acoustic events of bird calls. They installed two microphone arrays at two positions, used each microphone array to estimate the direction of the sound source, and then estimated the bird's position using the principle of triangulation. The authors also proposed a method [3] for estimating the position of a quadcopter based on the principle of triangulation after estimating the arrival direction of a sound source from two positions using a stereo microphone. However,

these methods have the problem that they only consider cases where there is one sound source in the observation space. The larger the observation space, the greater the number of acoustic events that occur there. And the greater the possibility that they might occur simultaneously, so the applicable scope of these conventional studies might be limited.

In contrast, many studies have been conducted on methods for detecting acoustic events and localizing sound sources that occur in a narrow space. It is not so hard to detect multiple acoustic events and localize multiple sound sources, when the observation space is narrow. One of the authors [16] also developed simultaneous localization of multiple sound sources, simultaneous sound source separation, and speech recognition in a narrow space of about 2 m radius using an 8-channel microphone array. This method, which implements three-speaker simultaneous speech recognition on the humanoid robot HRP-2, is implemented using multichannel signal processing with synchronous recording using a single 8-channel microphone array (you can see a demonstration [17] in Japanese). Multichannel signal processing with synchronous recording can utilize phase information between channels, so it can achieve high sound source localization and sound source separation capabilities in an overdetermined environment (i.e., an environment where the number of microphone elements on a microphone array is more than the number of existing sound sources in the observation space). However, systems that assume synchronous recording have the problem that even if it is desired to increase the number of microphone elements to cover a wide range of observation space, it is not easy to increase the number of microphone elements due to hardware constraints.

In this article, we propose a method that can detect multiple acoustic events simultaneously over a wide observation space and estimate the positions of their detected sound sources. The proposed method divides the observation space into small areas and observes the sound source while switching the microphone array responsible for observation for each divided area. This method does not require time synchronization between microphone arrays, so it has the feature that it can be easily distributedly arranged over a wide observation space.

Baseline methods based on triangulation can also be applied to observe divided small areas using multiple microphone arrays. In this article, we first describe an extension of the baseline method based on triangulation to apply it to multiple sound sources, and consider this method as a conventional method based on triangulation. The conventional method consists of a combination of localization and sound source detection using triangulation from three directions. The proposed method is based on the local maximum position of the spatial energy distribution.

The conventional method and baseline method are based on a two-step estimation framework of direction estimation and sound source position estimation. The disadvantage of the method is that two different criteria are applied in two-step estimation before determining an estimated position, which, in principle, accumulates estimation errors and increases the average localization error. In addition, some kind of processing is required to select the sound source that is considered to be the true sound source and remove spurious sound sources after estimating the sound source position. As such processing causes deterioration of the criterion for estimation, the accuracy of the sound source detection decreases.

The proposed method directly estimates the sound source position using a single criterion, without relying on the two-step framework that is problematic with triangulation-based methods. Our proposed method estimates the spatial energy distribution inside the observation space and finds the local maximum positions of the distribution as the sound source position. Our method treats the local maximum position as the sound source position. Since the local maximum value corresponds to the sound source being detected,

detection and sound source localization can be achieved simultaneously and using a single criterion. We show the effectiveness of our method by comparing the cumulative distribution of localization errors and the detected number of fake sound sources between a baseline method, conventional method, and the proposed method.

3. Triangulation-Based Method (Baseline Method)

Baseline sound source localization, as typified by Sumitani et al. [1], Gabriel et al. [2], Yamamoto et al. [4], and Takahashi et al. [3], estimates the sound source position based on the principle of triangulation from the sound observed by two microphone arrays. Given the sound source direction estimated by two microphone arrays, the intersection of two straight lines extending from the microphone array position to the sound source direction can be estimated as the sound source position. Sound source detection is performed using energy threshold processing. For threshold processing, there is a method that uses the energy of a sound source emphasis waveform obtained by beamforming [18,19] with one microphone array toward the estimated sound source position. There is also a method that uses the energy of the audio waveform collected by an element. Any microphone array can be triangulated as long as there are two microphone arrays. In the remainder of this text, we use an example of a 4ch microphone array as a specific system configuration.

In this work, we use a 4-channel circular microphone array with a diameter of 65 mm, as shown in Figure 2. Microphone elements are mounted at the positions indicated by the four orange circles in Figure 2. The four elements are arranged at equal angular intervals on the same circumference.

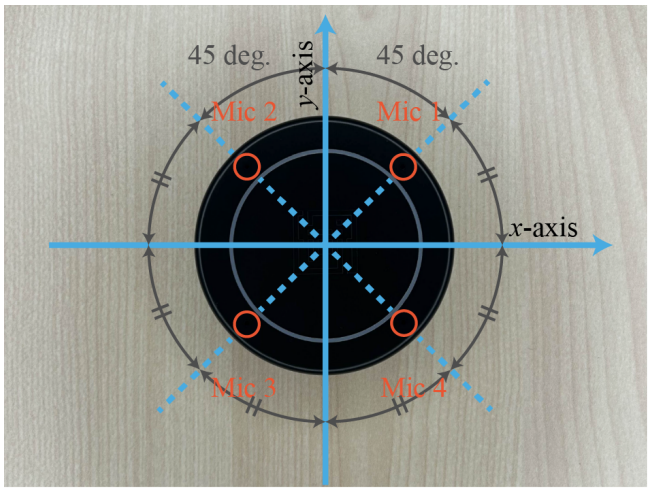

Figure 2. Microphone array with 4 elements (Seeed Studio ReSpeaker USB Mic Array v2.0 [20]).

Two microphone arrays are required to localize the sound source based on the principle of triangulation. We represent the coordinates of the two 4-channel microphone arrays, MA_1 and MA_2 as (x_1, y_1) and (x_2, y_2). In addition, we show the sound source directions estimated by these microphone arrays are θ_1 and θ_2. Figure 3 shows an overview of the localization of one sound source by triangulation with MA_1 and MA_2. The two orange circle marks show microphone arrays and the yellow green rectangle mark shows the localization position of a sound source. The yellow green rectangle mark is placed at the intersection of dotted lines from two microphone arrays. These two lines show the sound source direction estimated by the microphone array. The origin of the azimuth, which represents the sound source direction, is the positive direction of the y-axis in Figure 3. Also, the positive direction of the angle is counterclockwise.

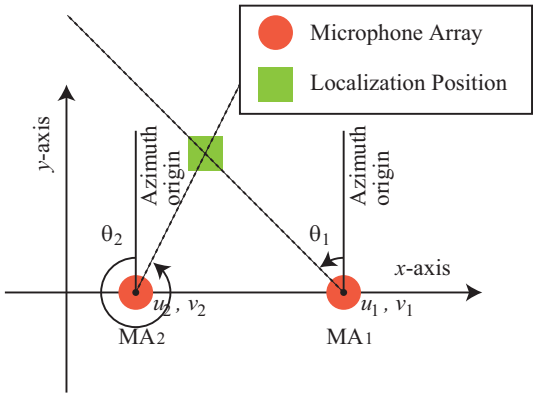

Figure 3. Sound source localization by triangulation with two-microphone arrays, MA_1 and MA_2.

Algorithms such as generalized cross-correlation with phase transform (GCC-PHAT) [10], multiple signal classification (MUSIC) [11], and steered response power (SRP) [12] can be used for sound source direction estimation. Some methods, such as GCC-PHAT, cannot estimate multiple sound sources simultaneously. MUSIC and SRP can simultaneously estimate the directions of multiple sound sources. However, compared to SRP, MUSIC is unsuitable for our objective, which uses a large number of microphone arrays, because it consumes a large amount of computational resources. SRP collects sound by beamforming in several predetermined directions, and estimates the direction of the sound source based on the relationship between the direction and the collected sound energy. This method uses the direction of the local maximum value in the directional distribution of energy as the direction of the sound source. In the case of multiple sound sources, multiple local maxima appear. Figure 4 shows an example of four local maxima appearing when two sound sources are observed by SRP based on the minimum variance distortionless response (MVDR) [18,19]. The horizontal axis shows the direction of beamforming. The vertical axis shows the output sound level of beamforming. A high output level can be obtained if the direction of the beamforming is close to the direction of the sound source. The steering step is 1 degree. Two sound sources are placed at 110 degrees and 172 degrees. Two local maxima appear at the directions corresponding to the true sound sources but the other two local maxima appear at 266 degrees and 355 degrees. Since the local maximum value of the true sound source often has a higher value, it is compatible with threshold processing to remove extra local maximum values. In this way, the direction of the sound source corresponds to the local maximum level obtained by scanning the observation space using beamforming.

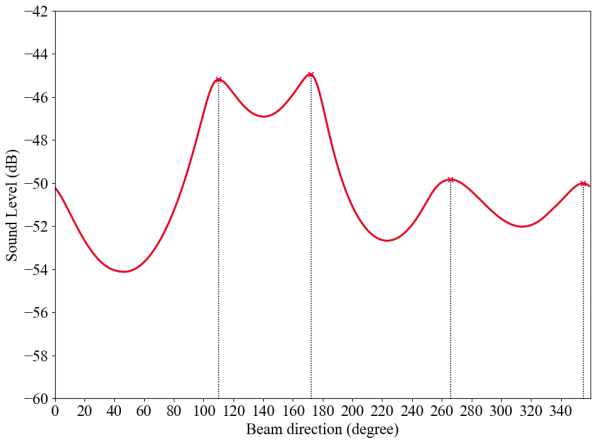

Figure 4. Simultaneous estimation of direction of arrival for two sound sources based on a circular microphone array.

4. Improved Version of Triangulation-Based Method (Conventional Method)

The sound source can be estimated based on triangulation when at least two microphone arrays are available. Since we distribute a large number of microphone arrays, it seems possible to improve the accuracy of acoustic event detection and sound source localization by triangulating with several microphone arrays. If the sound source is inside a triangle formed by three microphone arrays, stable localization is possible by triangulating with a pair of three-microphone arrays. Since sound source direction estimation involves errors, the sound source positions estimated by the three pairs of microphone arrays rarely match completely. An additional method is required to integrate these three estimation results.

To integrate the three estimated positions, we focus on a triangle formed with the three positions as vertices. The estimation results can be integrated by setting the center of gravity of this triangle as the new estimated position. This method can be easily extended to simultaneous localization of multiple sound sources. By selecting one straight line from each microphone array and selecting a triangle whose area is smaller than a threshold (ϕ) from among the triangles formed by these three straight lines, it is possible to determine the presence or absence of a sound source. We consider this method as a traditional method for triangulation-based acoustic event detection and sound source localization, and a competitor to our method proposed in this article. We call this method the conventional method.

We explain the details of the conventional method for one sound source. The three 4-channel microphone arrays (MA_m, $m = 1, 2, 3$) are placed at the vertices of an equilateral triangle area. In addition, we represent the sound source direction estimated by the mth microphone array MA_m as θ_m. Figure 5 shows an overview of localization of one sound source by triangulation with MA_1, MA_2, and MA_3. The orange circle marks show microphone arrays and the yellow green rectangle mark shows the localization position of a sound source. The yellow green rectangle mark is placed at the center of gravity of the triangle formed by the intersections of the three dotted lines corresponding to the sound source directions estimated by the three microphone arrays. The origin of the azimuth, which represents the sound source direction, is the positive direction of the y-axis in Figure 5. Also, the positive direction of the angle is counterclockwise.

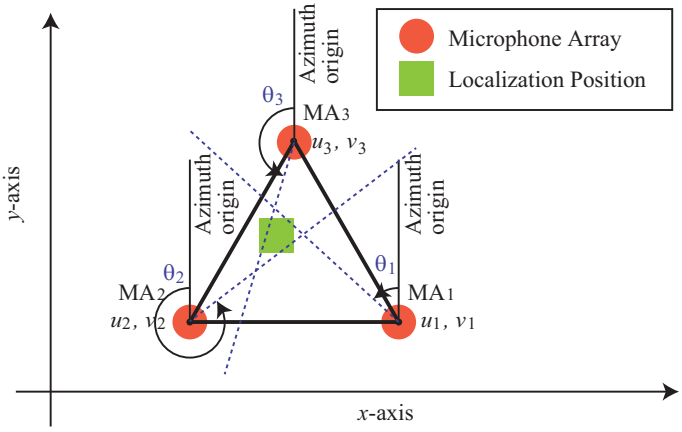

Figure 5. An overview of localization of one sound source by triangulation with MA_1, MA_2, and MA_3.

Next, we describe the details of the conventional method for two sound sources. In the case of one sound source, only one triangle can be formed from the estimated sound source direction, but in the case of two or more sound sources, multiple triangles can be formed from the estimated sound source directions. Therefore, we need to select appropriate triangles from multiple combinations. Figure 6 shows an overview of localization of

two sound sources by triangulation with MA_1, MA_2, and MA_3. We represent the set of sound source direction candidates estimated by the mth microphone array MA_m as $\boldsymbol{\theta}_m = \{\theta_m(1), \theta_m(2), \cdots, \theta_m(C_m)\}$. C_m is the number of sound source direction candidates detected by the mth microphone array MA_m. Figure 6 shows an example of $C_1 = 2$, $C_2 = 2$, $C_3 = 2$, $\boldsymbol{\theta}_1 = \{\theta_1(1), \theta_1(2)\}$, $\boldsymbol{\theta}_2 = \{\theta_2(1), \theta_2(2)\}$, $\boldsymbol{\theta}_3 = \{\theta_3(1), \theta_3(2)\}$. The orange circle marks show microphone arrays and the yellow green rectangle mark shows the localization position of a sound source. The yellow green rectangle mark is placed at the center of gravity of the triangle formed by the intersections of the three dotted lines corresponding to the sound source directions estimated by the three microphone arrays. The origin of the azimuth, which represents the sound source direction, is the positive direction of the y-axis in Figure 6. Also, the positive direction of the angle is counterclockwise. To avoid making this diagram too complicated, we have omitted the azimuth origin and curved arrows representing the angle of the sound source direction that was shown in Figure 5. In the example, two sound source directions are estimated from each microphone. Six straight lines can be drawn according to the estimated directions. We localize the sound source by finding a combination that can form two triangles and minimize the sum of their triangle areas.

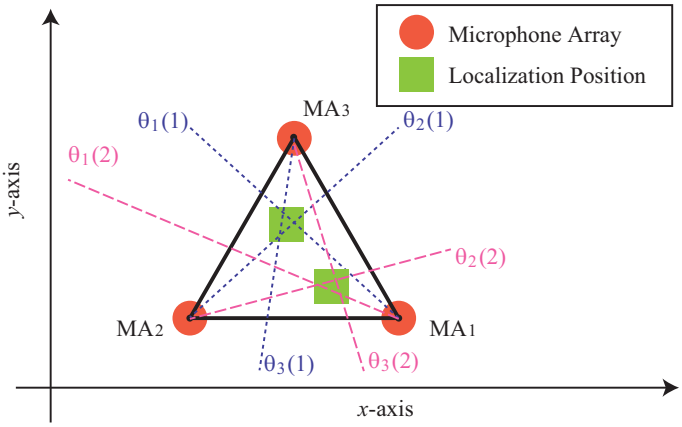

Figure 6. An overview of localization of two sound sources by triangulation with MA_1, MA_2, and MA_3.

5. Spatial Energy Distribution-Based Method (Proposed Method)

The method described in the previous section has a problem in that it uses a two-step estimation framework with different criteria, which limits the performance of acoustic event detection and sound source localization. In this section, we describe a proposed method that uses a single criterion to detect acoustic events and localize sound sources without relying on two-step estimation, which is a problem with triangulation-based methods.

The proposed method calculates the energy distribution in the observation space and uses the energy distribution to detect acoustic events and localize sound sources. This method assumes that an acoustic event occurs at the local maximum positions of the energy distribution, and estimates these positions as the sound source positions. Since the distributed microphone array is sparsely arranged in the observation space, it is not possible to directly measure any point in the observation space. Therefore, the problem to be solved by this method is formulated as the problem of estimating the energy distribution from the observed signals obtained from a distributed microphone array. The proposed method can be applied to one or more microphone arrays, but from the perspective of the application shown in Figure 1, we describe the procedure for estimating the spatial energy distribution using seven microphone arrays as an example.

We describe the logical processing scheme of the proposed method. If we can determine the spatial energy distribution in the observation space, we can know where the sound source is located in the observation space. This is because an acoustic event occurs from a certain position, and the acoustic energy is diffused and attenuated from there, so the problem of sound source localization boils down to the problem of finding the local maximum positions for a given distribution. We focused on dividing the observation space into small areas and finding the energy of each area. To obtain the energy distribution, we discretize the observation space. Although the spatial distribution is discrete, sound source localization can be achieved by making sufficiently fine divisions and checking which divided region has the maximum energy. Instead of directly measuring the energy of the divided small area, we propose a method for estimating its energy as the sum of the energy beam formed by several microphone arrays near the small area. This method consists of the following five steps:

Step 1: Perform the following procedure using all microphone arrays MA_m (for m = 0, 1, 2, 3, 4, 5, 6)
For all reference points R_r ($r = 1, 2, \cdots, 30$), calculate the energy $E(\mathrm{R}_r|\mathrm{MA}_m)$ of the output sound beamformed to the reference point R_r by the microphone array MA_m.

Step 2: Determine the contribution weights $W_m(\mathrm{R}_r)(m = 0,\ 1,\ 2,\ 3,\ 4,\ 5,\ 6)$ of microphone arrays $\mathrm{MA}_m(m = 0,\ 1,\ 2,\ 3,\ 4,\ 5,\ 6)$ for the energy estimation of reference point R_r.

Step 3: Calculate the estimated energy $E(\mathrm{R}_r)$ of reference point R_r by

$$\sum_{m=0}^{6} W_m(\mathrm{R}_r) E(\mathrm{R}_r|\mathrm{MA}_m) \tag{1}$$

Step 4: Find the local maximum points of $E(\mathrm{R}_r)$ regarding R_r.

Step 5: Output the local maximum point found in the step 4 as the position where the acoustic event occurred.

The microphone array is written as MA_m, $m = 0,\ 1,\ 2,\ 3,\ 4,\ 5,\ 6$, and the position of MA_m on the xy-plane is written as (x_m, y_m).

$$(x_m, y_m) = \begin{cases} (0,0), & m = 0 \\ (R\cos(2\pi(m-1)/6, R\sin(2\pi(m-1)/6)), & m = 1,\ 2,\ 3,\ 4,\ 5,\ 6 \end{cases} \tag{2}$$

where $R = 1.5$ m.

Figure 7 shows the arrangement of the microphone array and the position where the energy is estimated to assess the spatial energy distribution. Since it is not possible to estimate the energy distribution in a continuous space, the positions at which the energy is estimated are discretized into 37 positions. The energy distribution in this space is represented using 30 positions, excluding the 7 positions where the microphone array is placed. We call these 30 positions reference positions ($\mathrm{R}_r, r = 1, 2, \cdots, 30$). The orange circle marks show microphone arrays, and the green square marks show reference positions. A regular hexagon is made by connecting the MA_1, $\mathrm{MA}_2, \cdots, \mathrm{MA}_6$. The inside of this regular hexagon is the observation space.

If the energy of each reference position can be estimated, the energy distribution in the observation space can be estimated. The method for estimating the energy of each reference position is described later. Figure 8 shows an example of the spatial distribution of energy estimated from 30 reference positions. The energy distribution inside the regular hexagon in Figure 7 is represented by grayscale shading. The closer the color is to black, the higher the energy, and the closer it is to white, the lower the energy. The small regular hexagonal area with a red frame in Figure 8 shows the two positions of the local maximum values.

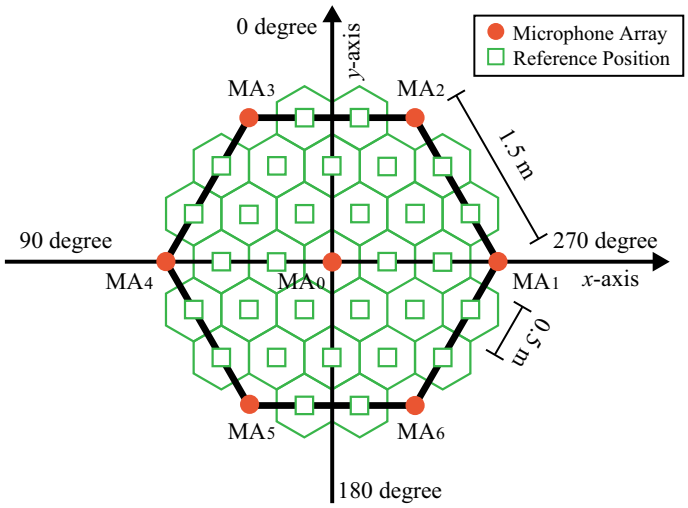

Figure 7. Layout of seven microphone arrays and reference positions.

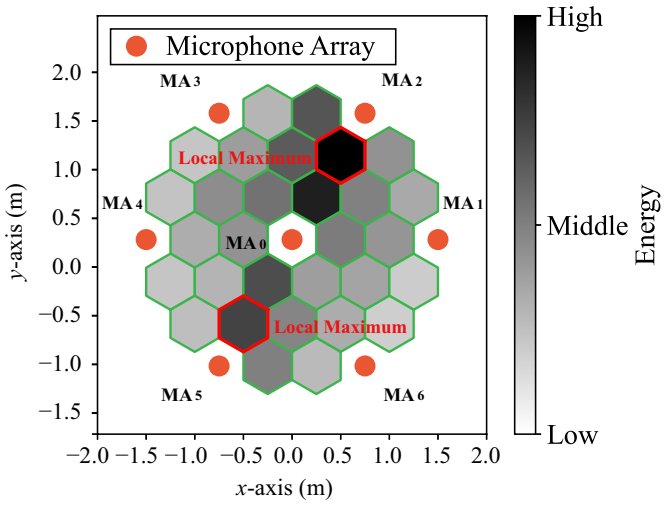

Figure 8. Heat map view of spatial distribution of energy.

In this method, the spatial resolution of sound source localization is determined by how densely the reference positions are placed inside a regular hexagon. In this study, we considered the area occupied by each adult when determining the distance between adjacent reference positions. The two talkers are assumed to be at least 0.5 m apart. Therefore, the reference position is the point that divides the regular hexagonal area into equilateral triangular areas with each side of 0.5 m, as shown in Figure 7.

We describe how to estimate the energy of each reference position. The estimated energy $E(\mathrm{R}_r)$ of the reference position R_r is defined as the following equation:

$$E(\mathrm{R}_r) = \sum_{m=0}^{6} W_m(\mathrm{R}_r) E(\mathrm{R}_r|\mathrm{MA}_m), \tag{3}$$

$$E(\mathrm{R}_r|\mathrm{MA}_m) = \frac{1}{T} \sum_{t=0}^{T-1} y^2(t|\theta_m), \tag{4}$$

where $\theta_m = \tan^{-1} \dfrac{y - y_m}{x - x_m}$ is the beamforming direction from the microphone array MA_m to the reference position $\mathrm{R}_r = (x, y)$, and W_m represents the amount of energy contribution of position (x, y) obtained by MA_m. W_m has the meaning of weight, and is $0 \leq W_m \leq 1$. Also, $E(\mathrm{R}_r|\mathrm{MA}_m)$ is the average energy per sample of the beamforming output waveform $y(t|\theta_m)$ when the beam is directed in the θ_m direction by MA_m, where t is the time index and T is the sample length of the observed signal. In our experiment, the minimum

variance distortionless response (MVDR) [18,19] is used for beamforming. A bandpass filter (1600 Hz–2500 Hz) is applied to the output waveform of beamforming by MVDR. This is because beamforming has low spatial resolution in the low-frequency range, and the effect of spatial aliasing is large in the high-frequency range.

We describe the MVDR filter. We represent the four channel observed signal at time t as $\boldsymbol{z}(t) = [z_1(t), z_2(t), z_3(t), z_4(t)]^T$, where ^{T} is a vector transpose. And the frame length is N and the frame shift length is S. The observed signal is represented in the frequency domain at f-frame as

$$\boldsymbol{Z}(f,k) = [Z_1(f,k), Z_2(f,k), Z_3(f,k), Z_4(f,k)]^T = \sum_{t=0}^{N-1} \boldsymbol{z}(t+fS|\theta) \exp\left(-\frac{2\pi tk}{N}\right). \quad (5)$$

Then, the MVDR filter coefficient $\boldsymbol{w}_{MV}(k|\theta)$ with its beam toward θ for frequency bin k is defined as

$$\boldsymbol{w}_{MV}(k|\theta) = [w_{MV,0}(k|\theta), w_{MV,1}(k|\theta), w_{MV,2}(k|\theta), w_{MV,3}(k|\theta)]^T, \quad (6)$$

$$= \frac{\boldsymbol{R}_k^{-1}\boldsymbol{a}(k|\theta)}{\boldsymbol{a}^H(k|\theta)\boldsymbol{R}_k^{-1}\boldsymbol{a}(k|\theta)}, \quad (7)$$

where $\boldsymbol{a}(k|\theta)$ is the steering vector of a micropohne array, $\boldsymbol{R}_k^{-1} = E[\boldsymbol{Z}(f,k)\boldsymbol{Z}(f,k)^H]$, and ^{H} is a Hermitian transpose. Since we cannot obtain this expected value, we introduce the following approximation:

$$\boldsymbol{R}_k^{-1} \simeq \frac{1}{L}\sum_{l=0}^{L-1} \boldsymbol{Z}(f+l,k)\boldsymbol{Z}(f+l,k)^H, \quad (8)$$

where L is empirically determined. In our experiment $L = 10$. The output signal of the MVDR filter with beam direction θ is given as $\boldsymbol{y}(t|\theta)$

$$y(fS+n) = \frac{1}{N}\sum_{k=0}^{N-1} Y(f,k|\theta)\exp\left(\frac{2\pi nk}{N}\right), (n = 0,1,\cdots,N-1), \quad (9)$$

$$Y(f,k|\theta) = \boldsymbol{w}_{MV}^H(k|\theta)\boldsymbol{Z}(f,k), (k = 0,1,\cdots,N-1). \quad (10)$$

There are implementation variations in the proposed method depending on how the weight $W_m(\mathrm{R}_r)$ is given. Since Yang et al. [21] succeeded in improving the accuracy of multiple sound source localization based on a Bayesian framework with time-delay of arrival (TDOA) by focusing on microphone arrays close to the sound source, we also propose weighting that preferentially uses microphones close to the sound source. For example, when estimating $E(\mathrm{R}_r)$ from three microphone arrays near the sound source, set $W_m(\mathrm{R}_r)$ to 1 for m of the three microphone array numbers near R_r, and set $W_m(\mathrm{R}_r)$ to 0 for the others. We propose a generalized method. We formulate the method of estimating $E(\mathrm{R}_r)$ from p microphone arrays in the neighborhood of R_r as the p-Neighbor method as follows:

$$W_m = \begin{cases} 1, & m \in \boldsymbol{I} \\ 0, & m \notin \boldsymbol{I} \end{cases}, \quad (11)$$

where a set $\boldsymbol{I}(= \{i_0, i_1, \cdots, i_{p-1}\})$ is the set of the top p-elements from the index sequence when the Euclidean distances d_ms between (x,y) and (x_m,y_m) are sorted in ascending order after computing the Euclidean distance $d_m = ((x-x_m)^2 + (y-y_m)^2)^{0.5}$ for $m = 0,1,2,3,4,5,6$.

The Neighbor method selects microphone array numbers that have a weight of 1.0 based on their proximity to the reference position $\mathrm{R}_r = (x,y)$. Çakmak et al. [22] showed that selecting the microphone pair near the sound source leads to improved estimation accuracy as a method for selecting the optimal microphone pair for TDOA estimation

from 23 synchronous microphones distributed in a conference room. We thought that this method could also be used to select between asynchronous microphone arrays. This selection method is suitable when the sound source is an omnidirectional sound source, such as a point source. However, sound sources are not always omnidirectional. For example, the radiation characteristics of human speech are known to be non-directional in the low-frequency range, but directional in the high-frequency range [23]. Since talkers and loudspeakers are directional sound sources, if microphone arrays are selected based on their proximity to the reference position $\mathrm{R}_r = (x, y)$, there is a risk of inappropriately selecting a microphone array located in a low radiational direction. When such a microphone array is selected, the value of $E(\mathrm{R}_r|\mathrm{MA}_m)$ is small to begin with, and its contribution to $E(\mathrm{R}_r)$ is small regardless of the weight $W_m(\mathrm{R}_r)$. We believe that it is easier to estimate the shape of the spatial energy distribution if microphone arrays with small contributions are not selected.

We also propose a method to estimate $E(\mathrm{R}_r)$ from the top p microphone arrays that are sorted in descending order of $E(\mathrm{R}_r|\mathrm{MA}_m)$. This method preferentially selects microphone arrays that observe higher energy when beamformed to $\mathrm{R}_r = (x, y)$. We call this method the p-Energy method and formulate it as follows:

$$W_m = \begin{cases} 1, & m \in \boldsymbol{J} \\ 0, & m \notin \boldsymbol{J} \end{cases} \tag{12}$$

where a set $\boldsymbol{J}(= \{j_0, j_1, \cdots, j_{p-1}\})$ is the set of the top p-elements from the index sequence when the observed energy e_ms beamforming toward $\mathrm{R}_r = (x, y)$ is sorted in descending order after computing the beamforming energy e_m for $m = 0, 1, 2, 3, 4, 5, 6$.

Threshold processing can be applied to the p-Neighbor method and p-Energy method. When $E(\mathrm{R}_r)$ takes a local maximum value inside a regular hexagon and $E(\mathrm{R}_r) > \phi$, we estimate that the sound source is placed at position $\mathrm{R}_r = (x, y)$. However, the determination of the threshold ϕ is often empirical. When the threshold is 0, the number of false detections of acoustic events is maximum. Increasing the threshold can reduce the number of false detections. However, at the same time, the number of positive detections also decreases. Furthermore, acoustic events, that could not be detected when the threshold is 0, are still not detected even if the threshold is changed. To newly detect such acoustic events, improving the estimation accuracy of spatial energy distribution is necessary.

6. Experiment

We evaluate the number of acoustic event detections and sound source localization errors by each method in this section. First, we describe an evaluation experiment in which a single directional sound exists. Then, we describe an evaluation experiment in which two directional sound sources exist.

The performance of acoustic event detection is generally evaluated by the number of correctly detected acoustic events in the evaluation dataset and the number of incorrectly detected acoustic events when no acoustic events exist. In addition, sound source localization performance is generally evaluated based on the distance difference between the estimated sound source position and the true sound source position in the evaluation dataset. It is not appropriate to directly apply these evaluation methods to a method that simultaneously evaluates acoustic event detection and sound source localization, so they are often used in combination for evaluation. Once the acoustic event detection process is performed, the sound source position of the detected sound source is estimated, and if the distance difference between the estimated position and the true sound source position is less than or equal to a certain value (ϵ), the acoustic event is considered to have been detected correctly. In other cases, it is considered not to have been detected. It is commonly

evaluated by the correct rate based on the number of acoustic events for the evaluation dataset. However, this evaluation has a problem in that it focuses on the performance of acoustic event detection. The problem is that the only evaluation indicator is that the localization error of the detected sound source is less than ϵ. Furthermore, ϵ is determined empirically. Therefore, we evaluate the number of correctly detected acoustic events for all ϵ. By extending the representation without changing the evaluation criteria from the common criteria, we can visualize the impact on the correct detection rate caused by increasing the acceptable localization error (ϵ) to increase the number of correctly detected acoustic events. The specific visualization procedure is to find the localization errors of all detected acoustic events and the cumulative distribution of these localization errors. Next, the cumulative distribution is plotted with the horizontal axis representing the error and the vertical axis representing the cumulative number of correctly detected acoustic events within the given localization error. The horizontal axis represents the acceptable localization error ϵ. In the following evaluation, the upper limit of the vertical axis is the number of acoustic events included in the evaluation dataset; so, the upper, middle, and lower ends of the vertical axis correspond to the correct detection rate of 100%, 50%, and 0%, respectively.

6.1. Localization Experiment for One Sound Source

Figure 9 shows the arrangement of the microphone arrays and the positions of the evaluation sound source. The orange circle marks show microphone arrays and the blue rectangle marks (23 positions) show sound source positions for evaluation. At the sound source position, a loudspeaker is installed facing each direction of 0, 45, 90, 135, 180, 225, 270 and 315 degrees, and we evaluate 184 conditions (= 23 positions × 8 directions).

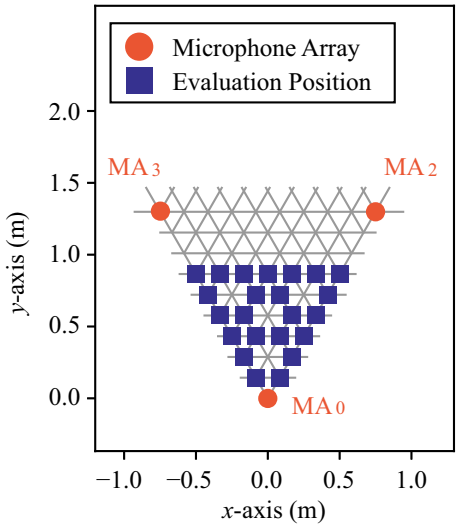

Figure 9. Layout of microphone arrays and sound source positions for evaluation.

The sound reproduced from the sound source at these positions was localized using the four-channel signal observed by the microphone array. Instead of evaluating the sound produced by actually reproducing the sound source signal from a loudspeaker at each position, we evaluated the sound obtained by convolving the sound source signal with the impulse response of each condition. The phonetically balanced sentences uttered by one male and one female from Japanese Newspaper Article Sentences (JNAS) [24] were used as the sound source signals, where the different sentences were used between a male and a female. These impulse responses were measured in advance from each position to each microphone array (i.e., these filenames are 'bm001a01.hs.raw' and 'bf001i01.hs.raw'). A time-stretched pulse (TSP) [25] signal with a sample length of 16,384 was used to measure

the impulse response. To measure the impulse response, we applied time synchronization processing 16 times to improve the signal-to-noise ratio (SNR). The Seeed Studio ReSpeaker USB Mic Array v2.0 (ReSpeaker) [20] was used for AD/DA conversion. Playback was performed through powered speakers (Genelec 8020DPM-1) connected to the stereo output of the ReSpeaker via an unbalanced-to-balanced converter (TASCAM LA-80MK2). The recording was performed using the ReSpeaker's onboard MEMS 4-channel microphone synchronized with the playback. The TSP playback was performed only on the left channel. The sampling rate for recording and playback was 16,000 Hz, and the number of quantization bits was 16 bit. Figure 10 shows the situation during TSP recording.

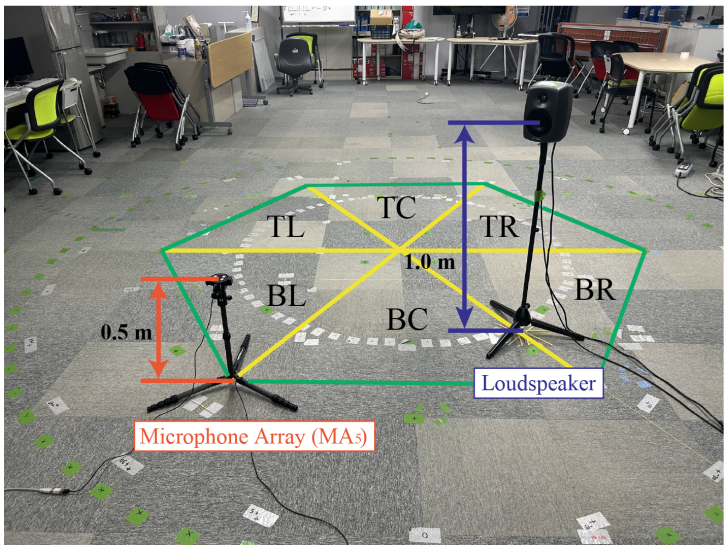

Figure 10. Impulse response measurement location and recording equipment.

As shown in the system overview in Figure 1, the microphone array was installed near the ceiling during system operation. To avoid deterioration of the sound source localization accuracy due to reflection from the ceiling, the microphone array was installed at a distance of about 0.5 m from the ceiling surface. The height of the ceiling of room 15806 was 2.5 m; so, assuming that the height of the lips of a person standing in the room is 1.5 m, the suitable height for the microphone array is 2.0 m from the floor. We considered the floor as the ceiling and measured the impulse response by placing a microphone array at a height of 0.5 m and a loudspeaker at a height of 1.0 m from the floor.

The impulse response was measured at Osaka Sangyo University, Building 15, Laboratory 806 (Room number 15806). Figure 11 shows the regular hexagonal area containing the measurement positions in the room superimposed on the bird's eye view of 15806. The background noise level in room 15806 was approximately $LA_{eq} = 44.6$ dB, and the reverberation time was approximately $T_{60} = 0.4$ s. The output level of the speech signal for evaluation was adjusted to be equivalent to $LA_{eq} = 65$ dB to 70 dB, which is the average Japanese conversation level.

Figure 12 shows the cumulative distribution of localization errors obtained from the sound source localization results for the evaluation dataset. As the horizontal axis is considered an acceptable error in sound source localization, the vertical axis corresponds to the number of acoustic events that could be localized within the acceptable error. Since the total number of acoustic events evaluated is 368 (=23 positions $\times$ 8 directions $\times$ 2 sentences), dividing the value on the vertical axis by 368 can be regarded as the acoustic event detection rate.

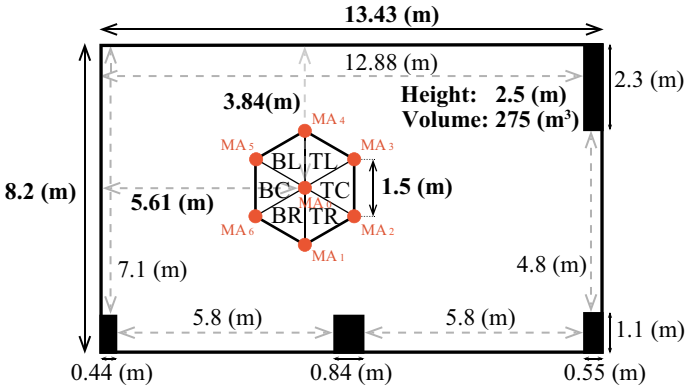

Figure 11. Our experimental environment (Osaka Sangyo University, Building 15, Laboratory 806).

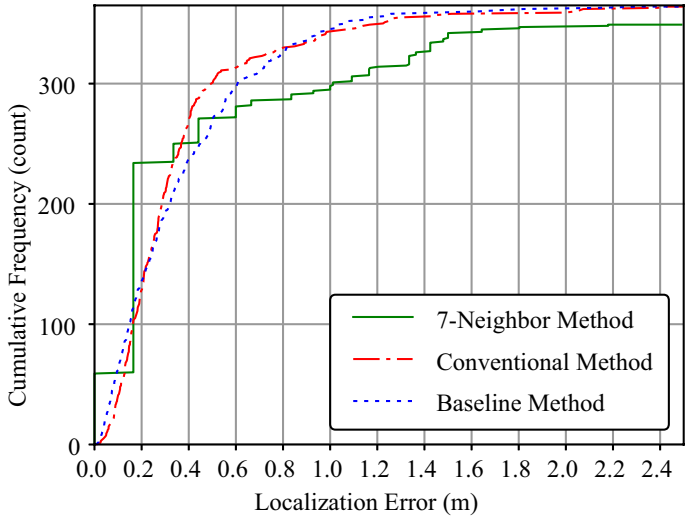

Figure 12. Cumulative distribution of localization error for one sound source.

The cumulative error distribution of the proposed method has a stepped graph shape. This is because the proposed method localizes based on processing that discretizes the observation space. This is also because the positions output as estimation results are also discretized. If the reference positions are arranged more densely, the result will be a finer step-like pattern than this. Although the number of acoustic events detected over 0.5 m is slightly lower than that of the baseline method and the conventional method, a similar trend appears.

This graph shows that the baseline method detects more acoustic events than the conventional method when the acceptable error is between 0.0 m and 0.2 m and over 0.8 m. However, in the vicinity of a practically important acceptable error of 0.5 m, the conventional method (under the condition: threshold $\phi = 0$ and setting the centroid of the triangle with the smallest area as the estimated position) detects more acoustic events. It can be confirmed that these methods have roughly equivalent performance. Therefore, as a representative sound source localization method based on triangulation, we choose the conventional method, which can be applied to two sound sources, instead of the baseline method, which can only be applied to one sound source.

6.2. Localization Experiment for Two Sound Sources

In this subsection, we compare the conventional method and the proposed methods based on the spatial energy distribution (p-Neighbor method and p-Energy method). We placed two directional sound sources within a hexagon and conducted an experiment to compare the performance of acoustic event detection and sound source localization. Two

sound source signals were generated in the same way as in the previous section, using one phonetically balanced sentence from one male and one female in JNAS [24]. Then, the evaluation sound was generated by mixing the two sound source signals. In this experiment, two sound sources existed at the same time; so, the first sound source was male speech and the second sound source was female speech.

Figure 13 shows the arrangement of the microphone array and the position of the evaluation sound source. The orange circle marks show the microphone arrays and the blue rectangle marks (63 positions) show the sound source positions for evaluation. These positions are called evaluation positions. The evaluation position is on the trisection point between the reference position and the adjacent reference position, and the reference position itself is also an evaluation position. The sound source signals for evaluation at these positions were created from the measured impulse responses as in Section 6.1.

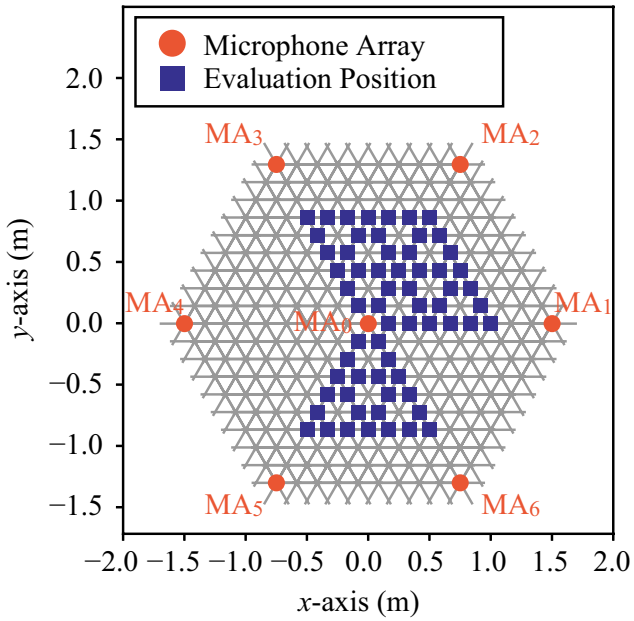

Figure 13. Layout of microphone arrays and sound source positions for evaluation.

We explain a variation in which two sound sources are placed. First, we divide a regular hexagon into six regular triangles. Then, we call each area Top Left (TL), Top Center (TC), Top Right (TR), Bottom Left (BL), Bottom Center (BC), and Bottom Right (BR), as shown in Figures 10 and 11. We can select one area for arranging each sound source from TL, TC, TR, BL, BC, and BR. There are ${}_6C_2 = 15$ combinations of arranging two sound sources in these six divided areas, but if we exclude geometrically similar combinations, there are four. We perform evaluation under comprehensive evaluation conditions by selecting two sound source positions from the pairs TC-TC, TC-TR, TR-BC, and TC-BC. However, candidates for sound source placement are limited to the positions marked by a blue rectangle inside of the TC, TR, and BC areas. The combinations where the distance between the two sound sources is not more than 0.5 m are excluded. For evaluation, we select three orientations of the loudspeaker among eight orientations, i.e., 0, 45, 90, 135, 180, 225, 270, and 315 degrees, at each sound source position. The selection method is the closest orientation toward the origin of the coordinate $(0, 0)$ from the sound source position and the two orientations adjacent to the closest orientation. There are 22,104 total evaluation conditions. Since there are two sound sources in one evaluation condition, the number of detections is 44,208 when all acoustic events are detected.

Figure 14 shows the results of the p-Neighbor method for $p = 3, 5, 7$. We can see how the cumulative distribution of localization errors changes depending on the number of

microphone arrays used to estimate spatial energy. When the acceptable error is less than 0.5 m, the number of detected acoustic events tends to increase as p increases from 3 to 7. In contrast, it can be seen that $p = 5$ is suitable over 0.5 m, especially over 1.0 m. It can also be seen that the number of detections of $p = 7$ saturates rapidly over 0.5 m.

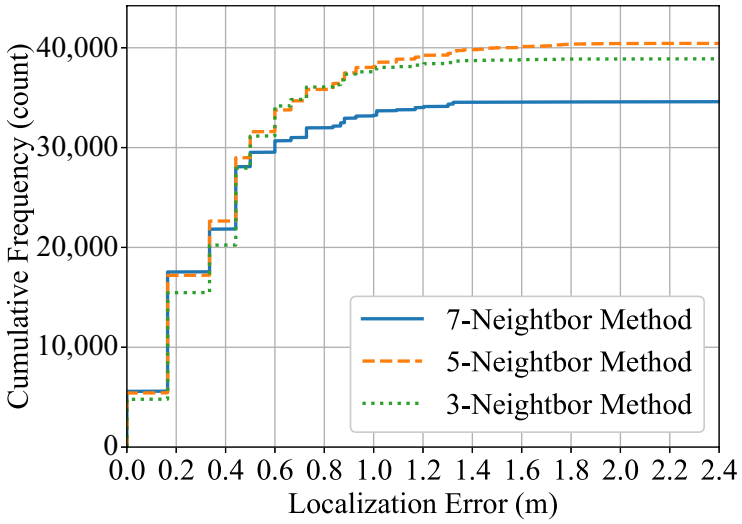

Figure 14. Cumulative distribution of error for two sound sources localized by the p-Neighbor method ($p = 3, 5, 7$).

Figure 15 shows the results of the p-Energy method for $p = 3, 5, 7$. We can see how the cumulative distribution of the localization errors changes depending on the number of microphone arrays used to estimate the spatial energy. When the acceptable error is less than 0.5 m, the number of detected acoustic events tends to increase as p increases from 3 to 7. In contrast, it can be seen that $p = 3$ is suitable over 0.5 m, especially over 1.0 m. It can also be seen that the number of detections of $p = 7$ saturates rapidly over 0.5 m.

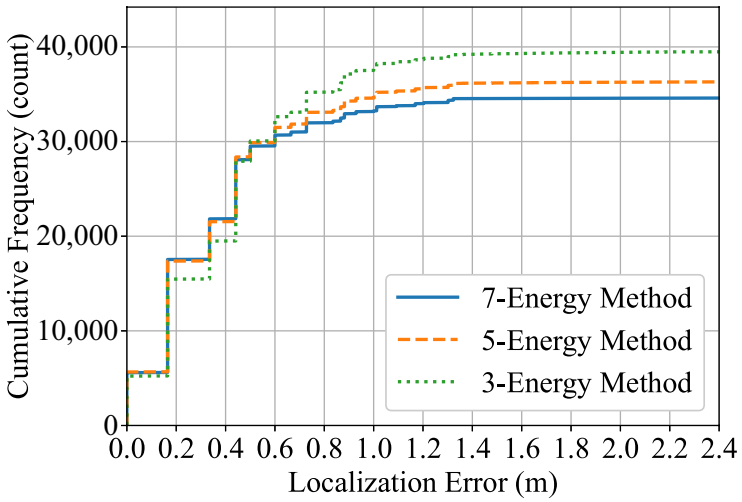

Figure 15. Cumulative distribution of error for two sound sources localized by the p-Energy method ($p = 3, 5, 7$).

From Figures 14 and 15, we confirm that $p = 3$ or $p = 5$ provides a well-balanced performance for both proposed methods. Finally, we compare the conventional method, 3-Neighbor method, 5-Neighbor method, 3-Energy method, and the 5-Energy method. The threshold value of the conventional method is compared with the previously optimized value of $\phi = 0.48$. Figure 16 shows the results. It can be confirmed that for any given

acceptable error, the proposed method has a higher ability to detect acoustic events than the conventional method. In particular, the 5-Neighbor method consistently ranks high in the number of acoustic event detections for all acceptable errors. There is no difference between the Neighbor method and the Energy method under the same p condition of less than 0.5 m. At a distance of 1.0 m or more, there is a difference in the trends, and the number of detected acoustic events is about the same for the 5-Neighbor method and the 3-Energy method.

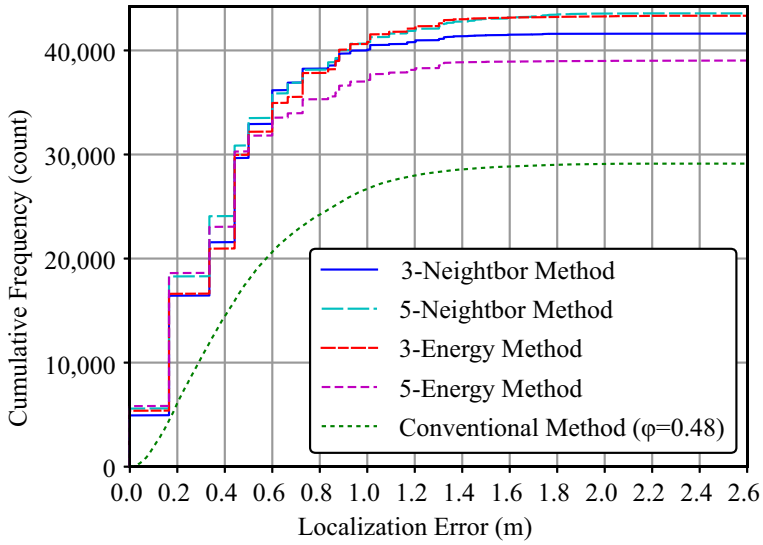

Figure 16. Cumulative distribution of error for two sound sources localized by conventional method, 3-Neighbor method, 5-Neighbor method, 3-Energy method, and 5-Energy method.

7. Discussion

In this article, we proposed a method for detecting acoustic events that occur in a wide space and estimating the location of the sound source. By detecting acoustic events and localizing sound sources in a wide space, it becomes possible to know the spatial distribution of the speaker positions. Since speakers who are spatially close to each other belong to the same dialogue group, each speaker within the observation space can be divided into dialogue groups using a point cloud clustering algorithm, such as the k-means algorithm. Although we described acoustic event detection and sound source localization for each sentence, it is possible to detect where a particular speaker conversed at each time and in which dialogue group by frame-by-frame-based processing acoustic event detection and sound source localization in units of several milliseconds to several seconds. In addition, it is possible to find out who speaks the most and at what point in time the speaker's utterances are leading their topic, which can be applied to facilitation support technology. Since it is possible to estimate this kind of information, it is expected that it will be applied to understanding complex dialogue activities in which multiple dialogues are occurring simultaneously. Since the proposed method does not involve speech recognition, it has the advantage of being applicable to applications that require protection of the privacy of utterances. Examples of multiple dialogues occurring at the same time include lecture halls where group work is conducted, panel discussions, and poster session halls. A method that covers a wide observation space, such as the one proposed in this article, is indispensable to realize such applications. We showed that an asynchronous distributed microphone array architecture is a promising method to detect acoustic events and localize sound sources in a wide observation space.

In this article, we evaluated a method that uses a wide space as the observation space by cutting out a part of that space into a regular hexagon and assuming that it does not

interfere with sound sources in adjacent hexagons. In the future, we believe that evaluations that take into account interference with adjacent areas will be necessary. Furthermore, in this evaluation, sound sources were detected on a per-utterance basis. To monitor dialogue activity in the temporal direction, it is necessary to detect acoustic events and localize sound sources on a segment-by-segment basis of every few milliseconds to several seconds. In this case, it is difficult to evaluate each elemental technology individually (e.g., acoustic event detection, sound source localization, target clustering, head-orientation estimation [26,27], etc.) as we have to solve the multiple target tracking problem [28,29].

Our evaluation was conducted in a room simulating a real environment at a sufficient distance from the wall. Therefore, in order to understand the basic properties of the algorithm, it is also considered necessary to evaluate it in a soundproof room or an anechoic room.

8. Conclusions

We proposed a method that simultaneously realizes acoustic event detection and sound source localization for multiple sound sources based on the spatial energy distribution using a single criterion. In an environment with two directional sound sources, we compared the acoustic event detection ability and localization error distribution of the proposed method and a method based on sound source direction estimation and triangulation (i.e., the conventional method). The effectiveness of the proposed method was confirmed by evaluating the number of events that could be detected within acceptable error. The proposed method used several microphone arrays in the observation space to estimate the spatial energy distribution, and we confirmed that the performance changes depending on the selection of the microphone array used for estimation. When we compared the case of selecting p microphone arrays near the evaluation position and the case of selecting p microphone arrays in descending order of the energy of the observed signal, we found that the best performance was obtained by selecting five microphone arrays near the evaluation position.

Author Contributions: Conceptualization, T.T. and M.N.; methodology, T.T.; software, T.T. and T.K.; validation, T.T., T.K. and M.N.; formal analysis, T.T.; investigation, T.T., T.K. and M.N.; resources, T.T.; data curation, T.T. and T.K.; writing—original draft preparation, T.T.; writing—review and editing, T.T., T.K. and M.N.; visualization, T.T.; supervision, T.T.; project administration, T.T.; funding acquisition, T.T. and M.N. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by JSPS KAKENHI Grant Number JP21H03488 and JP23K21691.

Data Availability Statement: The JNAS database used as the evaluation audio to reproduce this research is available from NII-SRC (https://research.nii.ac.jp/src/en/JNAS.html (accessed on 5 January 2025)). Scientific Python is used for multichannel signal processing. This software is available at the following link: https://scientific-python.org/.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Sumitani, S.; Suzuki, R.; Chiba, N.; Matsubayashi, S.; Arita, T.; Nakadai, K.; Okuno, G.H. An Integrated Framework for Field Recording, Localization, Classification and Annotation of Birdsongs Using Robot Audition Techniques—Harkbird 2.0. In Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 8246–8250.
2. Gabriel, D.; Kojima, R.; Hoshiba, K.; Itoyama, K.; Nishida, K.; Nakadai, K. 2D sound source position estimation using microphone arrays and its application to a VR-based bird song analysis system. *Adv. Robot.* **2019**, *33*, 403–414. [CrossRef]

3. Takahashi, T.; Fukuda, K.; Awatani, T.; Nakayama, M. Quadcopter Tracking by Acoustic Sensing Based on Two Stereo Microphones. In Proceedings of the 2023 IEEE 12th Global Conference on Consumer Electronics (GCCE), Nara, Japan, 10–13 October 2023; pp. 386–389.
4. Yamamoto, T.; Hoshiba, K.; Yen, B.; Nakadai, K. Implementation of a Robot Operation System-based network for sound source localization using multiple drones. In Proceedings of the 2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Macau, China, 3–6 December 2024.
5. Goseki, M.; Ding, M.; Takemura, H.; Mizoguchi, H. Combination of microphone array processing and camera image processing for visualizing sound pressure distribution. In Proceedings of the 2011 IEEE International Conference on Systems, Man and Cybernetics, Anchorage, AK, USA, 9–12 October 2011; pp. 139–143.
6. Bando, Y.; Otsuka, T; Itoyama, K.; Yoshii, K.; Sasaki, Y.; Kagami, S.; Okuno, H.G. Challenges in deploying a microphone array to localize and separate sound sources in real auditory scenes. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, 19–24 April 2015; pp. 723–727.
7. Masuyama, Y.; Bando, Y.; Yatabe, K.; Sasaki, Y.; Onishi, M.; Oikawa, Y. Self-supervised Neural Audio-Visual Sound Source Localization via Probabilistic Spatial Modeling. In Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, NV, USA, 25–29 October 2020; pp. 4848–4854.
8. Sato, G.; Shiramatsu, S.; Hoshino, M.; Watanabe, S.; Haibo, Y.; Mizumoto, T. LLM-based Structuring of Oral Discussion in Workshop to Support Collaboration among Local Government and Simulated Citizens. In Proceedings of the 30th International Conference on Collaboration Technologies and Social Computing, Barcelona, Spain, 11–14 September 2024; pp. 3–16.
9. Fallon, M.F.; Godsill, S.J. Acoustic Source Localization and Tracking of a Time-Varying Number of Speakers. *IEEE Trans. Audio Speech Lang. Process.* **2012**, *20*, 1409–1415. [CrossRef]
10. Knapp, C.H.; Carter, G.C. The Generalized Correlation Method for Estimation of Time Delay. *IEEE Trans. Acoust. Speech Signal Process.* **1976**, *24*, 320–327. [CrossRef]
11. Schmidt, R.O. Multiple Emitter Location and Signal Parameter Estimation. *IEEE Trans. Antennas Propag.* **1986**, *34*, 276–280. [CrossRef]
12. Ward, B. *Microphone Arrays*; Springer: Berlin/Heidelberg, Germany, 2001; pp. 157–180.
13. Jameel, M.M.; Ali, N.; Khan, M.; Wajid, M.; Usman, M. Finding Direction of Arrival using Uniform Circular Array of Microphones and Recurrent Neural Network. In Proceedings of the 2024 IEEE 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kamand, India, 24–28 June 2024; pp. 1–6.
14. Ge, Q.; Zhang, Y.; Wang, Y. A Low Complexity Algorithm for Direction of Arrival Estimation With Direction-Dependent Mutual Coupling. *IEEE Commun. Lett.* **2020**, *24*, 90–94. [CrossRef]
15. Gunjal, M.M.; Bazil Raj, A.A. Improved Direction of Arrival Estimation Using Modified Music Algorithm. In Proceedings of the 2020 IEEE 5th International Conference on Communication and Electronics Systems (ICCES), Virtual, 15–16 December 2020; pp. 249–254.
16. Takahashi, T.; Nakadai, K.; Komatani, K.; Ogata, T.; Okuno, G.H. Missing-feature-theory-based robust simultaneous speech recognition system with non-clean speech acoustic model. In Proceedings of the 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), St. Louis, MO, USA, 10–15 October 2009; pp. 2730–2735.
17. Available online: https://tallt.sakura.ne.jp/demo/3-SP-DEMO.wmv (accessed on 5 January 2025).
18. Van Veen, B.D.; Buckley, K.M. Beamforming: A versatile approach to spatial filtering. *Acoust. Speech Signal Process. Soc. Mag.* **1988**, *5*, 4–24. [CrossRef]
19. Capon, J. High-resolution frequency-wavenumber spectrum analysis. *Proc. IEEE* **1969**, *57*, 1408–1418. [CrossRef]
20. ReSpeaker. Mic Array v2.0—Far-Field w/4 PDM Microphones. Available online: https://wiki.seeedstudio.com/ReSpeaker_Mic_Array_v2.0/ (accessed on 5 January 2025).
21. Yang, J.; Zhong, X.; Chen, W.; Wang. W. Multiple Acoustic Source Localization in Microphone Array Networks. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *20*, 334–347. [CrossRef]
22. Çakmak, B.; Dietzen, T.; Ali, R.; Naylor, P.; Waterschoot, T. Microphone pair selection for sound source localization in massive arrays of spatially distributed microphones. In Proceedings of the 32nd European Signal Processing Conference (EUSIPCO), Lyon, France, 26–30 August 2024; pp. 251–255.
23. Chu, W.T.; Warnock, A.C.C. Detailed directivity of sound fields around human talkers. *NRC CNRC* **2002**, *IRC-RR-104*, 1–47.
24. Itou, K.; Yamamoto, M.; Takeda, K.; Takezawa, T.; Matsuoka, T.; Kobayashi, T.; Shikano, K.; Itahashi, S. JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research. *J. Acoust. Socity Jpn.* **1999**, *20*, 199–206. [CrossRef]
25. Suzuki, Y.; Asano, F.; Kim, H.Y.; Sone, T. An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses. *J. Acoust. Soc. Am.* **1995**, *97*, 1119–1123. [CrossRef]
26. Abad, A.; Segura, C.; Nadeu, C.; Hernando, J. Audio-based approches to head orientation estimation in a smart-room. In Proceedings of the Interspeech, Antwerp, Belgium, 27–31 August 2007; pp. 590–593.

27. Felsheim, R.C.; Brendel, A.; Naylor, P.A.; Kellermann, W. Head Orientation Estimation from Multiple Microphone Arrays. In Proceedings of the 2020 28th European Signal Processing Conference (EUSIPCO), Amsterdam, The Netherlands, 18–21 January 2021; pp. 491–495.
28. Blackman, S.S. *Multiple Target Tracking with Radar Applications*; Artec House: London, UK, 1986.
29. Granström, K.; Orguner, U. A phd filter for tracking multiple extended targets using random matices. *IEEE Trans. Signal Process.* **2012**, *60*, 5657–5671. [CrossRef]