Smart Structures and Systems, *Vol. 29, No. 6 (2022) 777-783*
https://doi.org/10.12989/sss.2022.29.6.777

Point-level deep learning approach for 3D acoustic source localization

Soo Young Lee[1], Jiho Chang[2] and Seungchul Lee*[1,3]

[1]*Department of Mechanical Engineering, Pohang University of Science and Technology (POSTECH), Pohang, Gyeongbuk 37673, South Korea*
[2]*Korea Research Institute of Standards and Science (KRISS), Daejeon 34113, South Korea*
[3]*Graduate School of Artificial Intelligence, Pohang University of Science and Technology (POSTECH), Pohang, Gyeongbuk 37673, South Korea*

(*Received October 15, 2021, Revised January 31, 2022, Accepted April 26, 2022*)

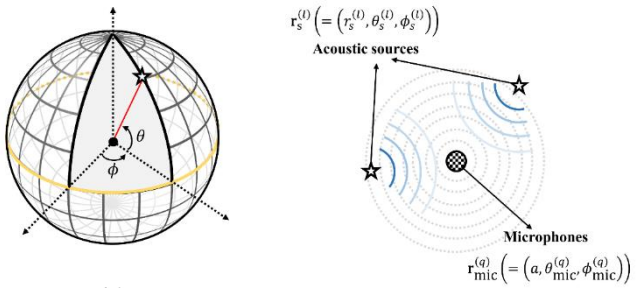
Abstract. Even though several deep learning-based methods have been applied in the field of acoustic source localization, the previous works have only been conducted using the two-dimensional representation of the beamforming maps, particularly with the planar array system. While the acoustic sources are more required to be localized in a spherical microphone array system considering that we live and hear in the 3D world, the conventional 2D equirectangular map of the spherical beamforming map is highly vulnerable to the distortion that occurs when the 3D map is projected to the 2D space. In this study, a 3D deep learning approach is proposed to fulfill accurate source localization via distortion-free 3D representation. A target function is first proposed to obtain 3D source distribution maps that can represent multiple sources' positional and strength information. While the proposed target map expands the source localization task into a point-wise prediction task, a PointNet-based deep neural network is developed to precisely estimate the multiple sources' positions and strength information. While the proposed model's localization performance is evaluated, it is shown that the proposed method can achieve improved localization results from both quantitative and qualitative perspectives.

Keywords: 3D acoustic source localization; 3D spherical beamforming; deep learning

1. Introduction

Acoustic source localization (ASL) refers to localizing and characterizing the sound sources by estimating their positions and strengths. While the ASL based on microphone arrays has been studied for decades, various techniques have been proposed for use in a variety of fields, including identifying noise sources of the mechanical system (Pillai and Burrus 1989), monitoring gas or liquid leaks (Kassab *et al.* 2019), and capturing talkers' speech signals (Brandstein and Ward 2013). Among several ASL approaches, beamforming methods have been explored to perform numerous ASL tasks due to their advantages of visualizing the acoustic source distribution map using a microphone array system (Bai *et al.* 2013).

While the spherical microphone array system can measure acoustic source signals in three-dimensional space, effective methods for processing the three-dimensional information of acoustic source distribution are required to improve ASL performance. Although the spherical microphone array can produce the acoustic source's direction-of-arrival (DOA) information in omnidirectional space, the relevant studies in the current literature are limited by using two-dimensional representations of spherical beamforming maps. An example of the 2D representation of the spherical beamforming map is illustrated in Supplementary Fig. 1. As shown in the figure, the two-dimensional representation of the spherical beamforming map is inherently distorted due to the mapping of three-dimensional information to two-dimensional space. Flattening a three-dimensional source distribution into a two-dimensional space causes severe spatial distortion in polar regions of the spherical beamforming map, which makes the task more challenging. In addition, another difficulty arises in the two-dimensional representation in that the 0° azimuthal region should be treated the same as the 360° azimuthal regions.



(a) (b)

Fig. 1 3D acoustic source localization scene: (a) a schematic plot that describes the acoustic source's direction-of-arrival (DOA) location in (θ, ϕ) (b) multiple acoustic sources and the microphone sensor system

To overcome the abovementioned challenges, this paper proposes a 3D deep learning (DL) approach for achieving high resolution and accuracy of multiple ASL in a spherical

*Corresponding author, Professor
E-mail: seunglee@postech.ac.kr

ISSN: 1738-1584 (Print), 1738-1991 (Online)

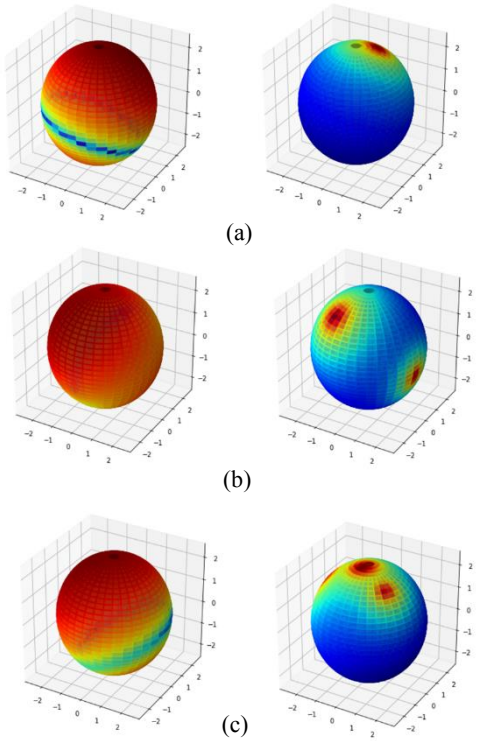


Fig. 2 Examples of 3D spherical beamforming maps and their 3D spherical target maps according to the number of acoustic sources N at 500 Hz: (a) N=1 (b) N=2 (c) N=3

microphone array system. To represent the DOA information of multiple acoustic sources in three-dimensional space (see Fig. 2), a novel method is first presented by defining a spherical target map that contains the sources' positions and strengths. The proposed spherical target map is defined based on the distance between a 3D scan grid and a source position, where multiple sources' arbitrary positions out of the grids can be expressed. While the multiple ASL task is converted into a 3D map-to-map point-wise prediction for the spherical target map, the PointNet-based localization network is then proposed to precisely estimate the output target map. Both quantitative and qualitative results are evaluated by our proposed method and a comparative model, i.e., fully convolutional autoencoder with regular convolution operations (denoted as a RegularCNN), to examine the prediction capability of the proposed method in comparison with the other 2D-based method.

The following summarizes the major contributions of this work in comparison to previous studies.

- *Multiple* acoustic source localization in a spherical microphone array system is presented. While the previous work (Lee *et al.* 2021a) proposed a data-driven method based on the spherical beamforming map, its application was restricted to *single* acoustic source localization. Since the previous work assumes only a single source, the method's scalability is limited for more diverse cases with more than one acoustic source. On the other hand, this study proposes a method for localizing multiple acoustic sources by establishing the spherical target map that spatially represents multiple sources' positions and strengths in a spherical microphone array system.
- *Three-dimensional* representation learning of the spherical beamforming map is presented. The earlier work (Lee *et al.* 2021b) considered the planar microphone array, which results in *a two-dimensional* beamforming map. On the other hand, this study expands the microphone array system from a 2D planar type to a 3D spherical type, capturing 3D information without any distortion is needed. In this study, the 3D representation-based deep localization neural network is proposed to achieve the three-dimensional ASL in a spherical microphone array.

The remainder of this paper is organized as follows. In Section 2, the related works on model-based and data-driven approaches are described. Section 3 provides the proposed spherical target map and the 3D localization neural network. The experimental settings, results, and discussion are presented in Section 4. Finally, Section 5 concludes the paper.

2. Related works

2.1 Model-based approaches

As model-based approaches, beamforming methods have been mainly developed in two stages, i.e., conventional beamforming (CB) and deconvolution algorithms. First, the conventional beamforming (CB) algorithm has been designed to produce the sources' spatial distribution based on phased microphone array and scanning vectors, while it is known to be relatively robust to external noise (Castellini and Martarelli 2008). However, the CB algorithm has the disadvantage of yielding poor spatial resolution at low frequencies as the resolution is proportional to the wavelength. On the other hand, deconvolution methods have been developed to increase the spatial resolution by iteratively deconvolving a CB map to examine the beamformer's response (Lylloff *et al.* 2015). Even though these deconvolution methods contribute to improving spatial resolution to a certain extent, they intrinsically entail high computational costs and restricted spatial resolution. Also, occasionally, the iteration process of the deconvolution methods does not converge, but it is also known to bring about ghost source problems (Lylloff and Fernandez-Grande 2018).

2.2 Data-driven approaches

Recently, deep learning (DL) has been introduced to the field of acoustic source localization. Unlike the conventional approaches with the model-based methods, the DL-based method utilizes a data-driven approach via deep-constructed neural networks to describe nonlinear phenomena from the learned representation of the features. With the DL algorithms' progress and their applicability, several studies using the DL-based methods have shown promising results for acoustic source localization, while the previous studies are mainly classified into two approaches: grid-based and grid-free methods. For example, as a grid-

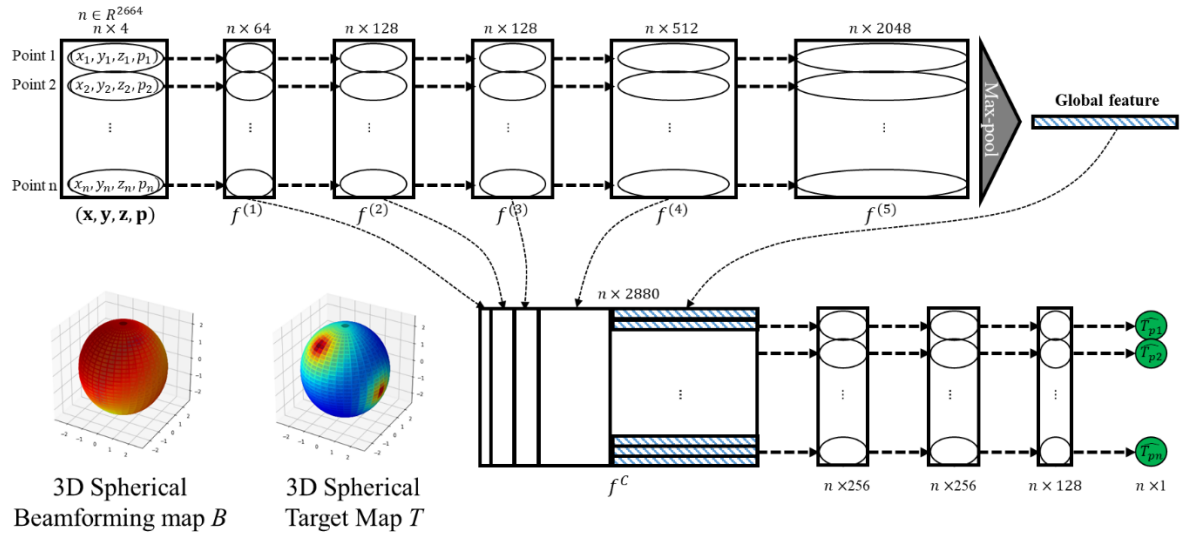


Fig. 3 Architectural description of the PointNet-based 3D acoustic source localization network

based approach, Ma and Liu (2019) suggested a convolutional neural network (CNN) for localizing multiple acoustic sources into a 10×10 grid dimension of the source distribution. Using simulation data from the phased microphone array, the authors demonstrated promising localization results for the acoustic sources on the grids. Xu *et al*. (2020) also presented a grid-based method by employing densely connected CNN to construct a source distribution map from a cross-spectral matrix (CSM). In contrast, as a grid-free approach, Kujawski *et al*. (2019) developed the CNN prediction model with three output neurons to estimate the single point source's strength and coordinates.

It should be noted that the aforementioned DL-based approaches merely consider a planar microphone array system, where the source distribution maps are extracted on a planar grid system with a regular two-dimensional array. On the contrary, the spherical microphone array system with a three-dimensional sphere grid is considered in this study, while it can be utilized to describe the acoustic sources' distribution in the omnidirectional situation, e.g., the 3D real world. While a recent work (Lee *et al*. 2021a) suggested a data-driven ASL method in an omnidirectional environment, the application scope of the study was limited to single-source cases, making it difficult to apply in practical situations where multiple acoustic sources may exist, which motivated to this study.

3. Proposed method

This section proposes a novel representation method to define output target maps spatially for multiple acoustic sources in a spherical microphone array system. Subsequently, the details on the PointNet-based deep localization neural network are provided.

3.1 3D spherical target map

A target function is proposed to generate a clean 3D output map (denoted as a spherical target map) that can spatially represent multiple sources without assuming the sources are located on spherical grids. The spherical target map can be constructed by the multiple acoustic sources' information, i.e., positions and strengths, in the form of a narrow main lobe and no side lobes for each source. This distance-based target function can be described as follows.

$$f(R) = \frac{\epsilon}{R^N + \epsilon}\,, \tag{1}$$

where R denotes the distance between a source and a grid of the spherical target map, and ϵ is a constant to avoid singularity. From Eq. (1), the spherical target function has the maximum value of unity when R is zero. In this study, the values of ϵ and N are determined such that the spherical target function decreases 13 dB when the interval between a source and a grid increases by grid size Δx. Note that the grid size Δx is set to be 5°, which entirely generates 2,664 points (37 points in a polar angle direction and 72 points in an azimuthal angle direction). Each point is comprised of a 4-dimensional vector of its Cartesian coordinate values (x, y, z) as well as the strength p. Several examples of the 3D spherical beamforming maps and their 3D spherical target maps are described in Fig. 2. As a result, the spherical microphone array system's 3D acoustic source localization task can be formulated as a point-level strength prediction task by our proposed representation of a spherical target map.

3.2 PointNet-based localization network

While the proposed target map expands the multiple acoustic source localization into a 3D map-to-map point-level prediction task, the PointNet-based deep neural network (Qi *et al*. 2017) is mainly utilized to estimate the multiple sources' positions and strengths precisely. The PointNet is the point set-based deep neural network that has been introduced to leverage the deep learning framework for the point cloud representation. Fig. 3 describes the proposed network architecture for the above-mentioned 3D

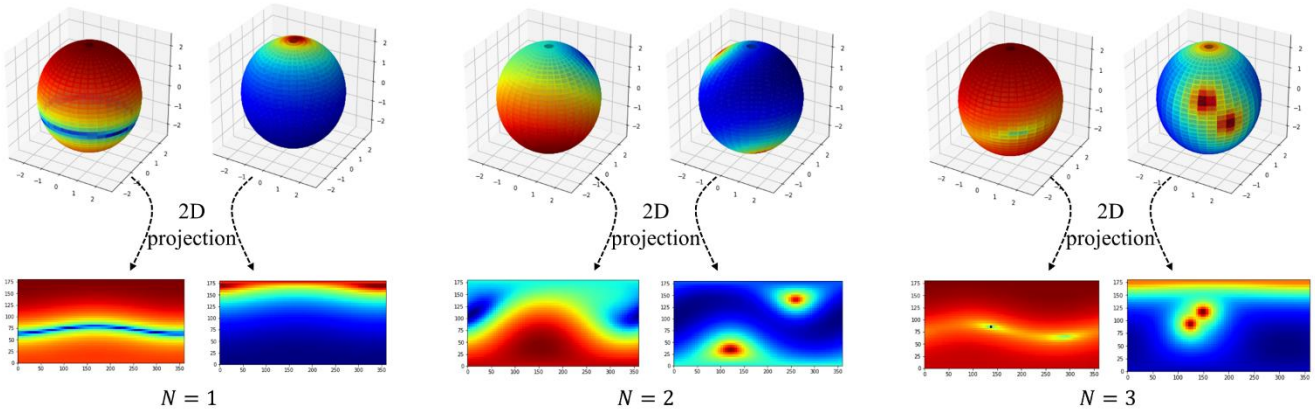


Fig. 4 Several examples of the 3D spherical beamforming maps, the 3D spherical target maps and their 2D projected maps that are flattened into the 2D space

acoustic source localization task. On the whole, the localization network is analogously constructed compared to the original PointNet for the segmentation task, except that the joint alignment networks, i.e., T-Net, are excluded at the early part of the network. It should be noted that these alignment networks can be removed since the 3D spherical beamforming representation does not suffer from any transformation. Given point-level input (x, y, z, p) that contains positional and strength information of the 3D spherical beamforming map, the former part of the localization network is trained to extract higher-level point-wise features via successive multi-layered perceptrons (MLPs) as

$$f_j^{(l)} = \varphi\left(\sum_i \omega_{ij}^{(l)} f_i^{(l-1)} + b_j^{(l)}\right),$$

where $f_j^{(l)}$ is the output feature of j-th neuron in the l-th MLP layer, $f_i^{(l-1)}$ is the input feature of i-th neuron in the former MLP layer. $\omega_{ij}^{(l)}$ and $b_j^{(l)}$ denote the trainable weights and biases for the j-th neuron in the l-th MLP layer. φ denotes the nonlinear activation function at each MLP layer, where a rectified linear unit (ReLU) $\varphi(\cdot) = \max(0, \cdot)$ is utilized. While each MLP layer is shared among every point elements to extract l-th layer's point-wise features $f^{(l)}$ (Qi *et al.* 2017), the feature extraction process consists of five MLP layers, and then the max-pooling operation is conducted to capture the global characteristics of the point-level features (see Fig. 3). Subsequently, the global features after the max-pooling operation are concatenated to per-point features from earlier MLP layers $f^{(l)}$ in order to combine the point-level hierarchical context from several MLP layers and the global contextual information from the max-pooling layer. Finally, the combined feature f^C is propagated to the three MLP layers and an output layer. As a result, the proposed network is trained to learn the point-wise signal strengths of the 3D spherical target map T. As for the loss function, the point-wise mean squared error (MSE) is applied to minimize point-level strength (dB) errors between the actual spherical target map T and the predicted spherical target map $\hat{T}$. The loss function L is expressed as:

$$L(T, \hat{T}) = \frac{1}{I}\sum_i (T_i - \hat{T}_i)^2.$$

Based on the neural network formulation and loss function, the proposed model is trained via a backpropagation algorithm, while trainable parameters are updated based on a mini-batch gradient descent algorithm with AdamOptimizer (Kingma and Ba 2014). The model is implemented using Python 3.7.8/PyTorch 1.9.0 framework and trained using an NVIDIA GeForce 2080 Ti.

4. Results and discussion

This section gives experimental results and discussion of multiple acoustic source localization. First, the data set used in this study is briefly described. Based on the data set comprised of various cases with multiple acoustic sources, both quantitative and qualitative results for the proposed model and the comparative method are presented.

4.1 Experimental description

The spherical microphone array set-up with regularly positioned 12 microphones is adopted in this study. As previously mentioned, the spatial resolution of the spherical beamforming map is set to be 5°, resulting in a total of 2,664 points. As for the frequency of the acoustic source, 500 Hz is considered. Besides, the number of acoustic sources N is set up to be three, where N is randomly selected from one to three during the data generation process. Each source's position, strength magnitude, and phase are also arbitrarily selected. Through the simulation-based data generation process, a total of 30,000 samples for the entire data set are extracted, where they are separated into 22,950 samples for the training set, 2,550 samples for the validation set, and 4,500 samples for the test set, respectively. Several examples corresponding to the different number of acoustic sources are visualized in Fig. 4. It is worth mentioning that we mainly visualize the 3D spherical beamforming maps and 3D spherical target maps in 2D projected maps to compare the proposed model's localization performance with the 2D-based method, i.e.,

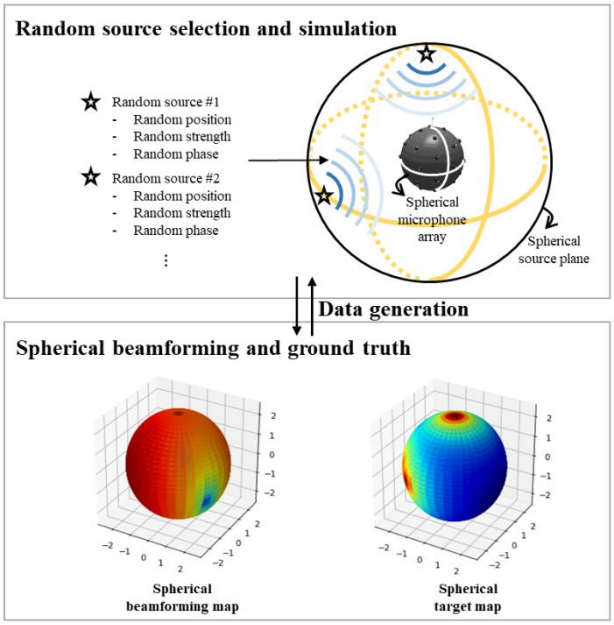


Fig. 5 A schematic diagram that shows the data generation process of random acoustic source localization simulation

RegularCNN. An overall description of the data generation process is depicted in Fig. 5.

Based on the generated dataset, architectural components and hyper-parameters of the proposed model and RegularCNN are optimized. To prevent excessive complexity of searching space in the entire optimization process, network structures of the proposed model and RegularCNN are first searched through manual search. While similarly controlling the total number of parameters of the two models, we manually find the network architectures in which the learning convergence of two models is guaranteed.

Table 1 Quantitative comparisons between the baseline and the proposed model

Model	Averaged strength error (dB)	Averaged distance error (°)	Averaged 2D-SSIM
RegularCNN	14.13	11.01	0.924
Proposed model	6.81	3.74	0.965

In this process, the architectural parameters of the proposed model include the layer depth and the number of neurons, while those of RegularCNN include the layer depth, the convolution kernel size, and the number of channels. The searched network structures are presented in Supplementary Table 1. In addition, 100 iterations of random search are conducted to find the best combination of hyper-parameters of batch size, learning rate, and L2 regularization coefficient. Searching spaces for considered hyperparameters are summarized in Supplementary Table 2. As a result, each model that shows the lowest loss value to the validation dataset through the overall optimization process mentioned above is selected.

4.2 Localization of multiple acoustic sources

Qualitative comparisons between the baseline model, i.e., RegularCNN, and our proposed model are described in Fig. 6. In this study, note that the 3D beamforming maps and the 3D target maps are represented with the 2D projected maps for visually comparing the proposed model's prediction performance with the baseline model. While B and T denote input spherical beamforming map and the ground-truth spherical target map, respectively, $\hat{T}_{reg}$

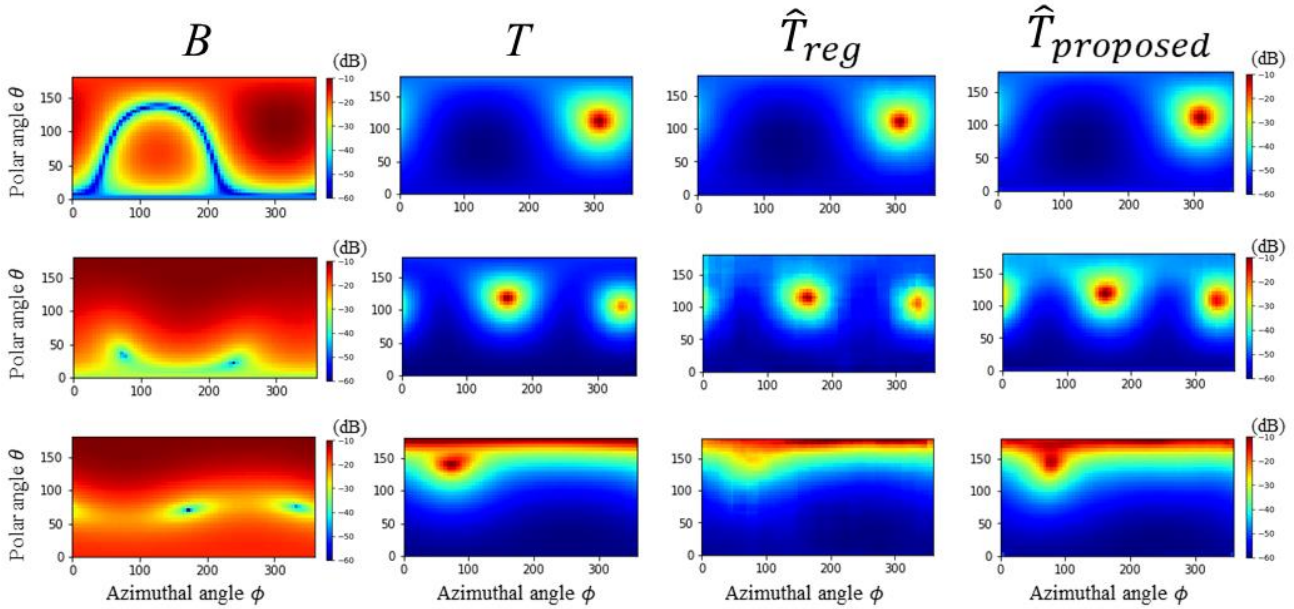


Fig. 6 Qualitative comparison between the comparative model and the proposed model. Each column represents projected spherical beamforming maps B, projected spherical target maps T, predicted spherical target maps of RegularCNN $\hat{T}_{reg}$, and predicted spherical target maps of our proposed model $\hat{T}_{proposed}$. Each row represents the evaluation case according to the number of acoustic sources (N = 1, 2, 3)

and $\hat{T}_{proposed}$ represent the spherical target maps predicted by the RegularCNN and the proposed model, respectively. As shown in Fig. 6, it is observed that both models succeed in estimating a single source's location as well as its strength value. However, the performance of the RegularCNN is shown to deteriorate as the number of acoustic sources increases. On the other hand, we observe that our proposed model utilizing the 3D point-level deep learning approach yields more precise spatial distribution than the RegularCNN even in $N = 3$ case, resolving two adjacent acoustic sources located at the polar region near $\theta = 180^{\circ}$.

Quantitative results for the localization performance between the baseline and the proposed model are also summarized in Table 1. To assess the model's capability for acoustic source localization, we consider three different evaluation metrics, i.e., strength error, distance error, and 2D structural similarity (SSIM) index. While the strength error and the distance error measure absolute differences between the actual and the predicted acoustic sources in terms of the strength magnitude and angular distance, the 2D-SSIM can be used to measure spatial similarity between the actual spherical target map and the predicted spherical target maps. As shown in Table 1, it is observed that our proposed model can achieve better localization performance than the baseline model both in strength error, distance error, and 2D-SSIM. It is shown that the proposed model reaches a mean strength error of 6.81 dB and a mean distance error of 3.74°, showing more than 50% performance improvement over the existing approach. It is also found that approximately 0.041 increase in the 2D-SSIM value, where the higher SSIM value signifies the localization model's more improved performance in spherical target map prediction. As a result, this confirms that the proposed approach can achieve improved accuracy for estimating the acoustic sources' positions and strengths.

Qualitative and quantitative comparisons indicate that the proposed method has superiority in the following parts. The most prominent part is that even if the same dataset is used, the performance of acoustic source localization can be enhanced depending on which representation learning strategy is adopted. To overcome the spatial distortion inherent in the 2D spherical beamforming map, this work proposed a method for learning the 3D representation itself without distortion, resulting in a considerable performance increase. The proposed method was shown to improve localization performance, particularly for multiple acoustic source scenarios. Considering that the larger the number of sources, the greater the number of overlapping source distributions and distortion, the proposed method is expected to contain advantages when much more potential sources exist or the degree of distortion is more severe. In addition, it is also noted that the proposed method achieved localization improvement even with fewer trainable parameters. As previously described in Supplementary Table 1, the total number of parameters of the proposed model is 45,976 less than that of the comparative model. In this wise, the fact that the performance was enhanced even with a smaller amount of parameters has the advantage of reducing the model complexity, which makes it more useful for practical applications.

5. Conclusions

This study proposed a 3D representation deep learning approach to attain high resolution and preciseness for multiple acoustic source localization of the spherical array system. First, a spherical target function was suggested to spatially represent multiple sources' positional and strength information within the 3D spherical target map, where the acoustic source localization task is converted into the point-wise prediction one. Besides, a PointNet-based deep localization neural network was developed to achieve the point-level 3D spherical target map prediction with high accuracy. It was shown that the proposed model could improve both quantitative and qualitative results compared to the other 2D CNN-based method using regular convolution operations with equirectangular projected beamforming and target maps. Regarding future work directions, further verification using a variety of real-world experiments will be investigated. In addition, we intend to further identify sources' distance to maximize the benefits of our proposed 3D representation learning approach.

Acknowledgement

The research was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korea Government Ministry of Science and ICT (MSIT) (No. 2020R1A2C1009744), in part by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (No. 2019-0-01906, Artificial Intelligence Graduate School Program (POSTECH)), in part by the Korea Institute of Energy Technology Evaluation and Planning (KETEP) Grant funded by the Korean Government [Ministry of Trade, Industry, and Energy (MOTIE)] under Grant 20206610100290, and in part by the Fundamental Research Program of the Korea Research Institute of Standards and Science.

Reference

Bai, M.R., Ih, J.G. and Benesty, J. (2013), *Acoustic Array Systems: Theory, Implementation, and Applications*, John Willey & Sons.

Brandstein, M. and Ward, D. (2013), *Microphone Arrays: Signal Processing Techniques and Applications*, Springer Science & Business Media.

Castellini, P. and Martarelli, M. (2008), "Acoustic beamforming: Analysis of uncertainty and metrological performances", *Mech. Syst. Signal Pr.*, **22**, 672-692. https://doi.org/10.1016/j.ymssp.2007.09.017.

Kassab, S., Michel, F. and Maxit, L. (2019), "Water experiment for assessing vibroacoustic beamforming gain for acoustic leak detection in a sodium-heated steam generator", *Mech. Syst. Signal Pr.*, **134**, 106332. https://doi.org/10.1016/j.ymssp.2019.106332.

Kingma, D.P. and Ba, J. (2014), "Adam: A method for stochastic optimization", arXiv preprint arXiv:1412.6980.

Kujawski, A., Herold, G. and Sarradj, E. (2019), "A deep learning method for grid-free localization and quantification of sound sources", *J. Acoust. Soc. Am.*, **146**, EL225-EL231.

https://doi.org/10.1121/1.5126020.

Lee, S.Y., Chang, J. and Lee, S. (2021a), "Deep learning-enhanced single point sound source localization for spherical microphone array", *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, **263**, 2279-2283.

Lee, S.Y., Chang, J. and Lee, S. (2021b), "Deep learning-based method for multiple sound source localization with high resolution and accuracy", *Mech. Syst. Signal Pr.*, **161**, 107959. https://doi.org/10.1016/j.ymssp.2021.107959.

Lylloff, O. and Fernandez-Grande, E. (2018), "Noise quantification with beamforming deconvolution: effects of regularization and boundary conditions", *The 7th Berlin Beamforming Conference*, Berlin, March.

Lylloff, O., Fernandez-Grande, E., Agerkvist, F., Hald, J., Roig, E.T. and Andersen, M.S. (2015), "Improving the efficiency of deconvolution algorithms for sound source localization", *J. Acoust. Soc. Am.*, **138**, 172-180. https://doi.org/10.1121/1.4922516.

Ma, W. and Liu, X. (2019), "Phased microphone array for sound source localization with deep learning", *Aerosp. Syst.*, **2**, 71-81. https://doi.org/10.1007/s42401-019-00026-w.

Pillai, S.U. and Burrus, C.S. (1989), *Array Signal Processing*, Springer.

Qi, C.R., Su, H., Mo, K. and Guibas, L.J. (2017), "PointNet: Deep learning on point sets for 3D classification and segmentation", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, July.

Xu, P., Arcondoulis, E.J.G. and Lie, Y. (2020), "Deep neural network models for acoustic source localization",. *The 8th Berlin Beamforming Conference*, Berlin, March.