ВЕСТНИК ТОМСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА

2025 Управление, вычислительная техника и информатика № 72

Tomsk State University Journal of Control and Computer Science

Original article
UDC 004.032.26, 004.048
doi: 10.17223/19988605/72/11

Evaluating the generalization ability of deep learning models for sound source localization

Ghiath M. Shahoud[1], Evgeny D. Agafonov[2]

[1, 2] Siberian Federal University, Krasnoyarsk, Russian Federation

[1] ghiathlovealaa@gmail.com
[2] evgeny.agafonov@mail.ru

Abstract. In this paper, the generalization ability of deep learning models used to solve the sound source localization problem with a spatial resolution of 10° is evaluated when the configuration settings are changed. The generalization ability of the models was evaluated in a closed reverberant environment using an orthogonal microphone array. Two models were considered: SI-GCC-CNN, which is based on combining the features of sound intensity and generalized cross-correlation - phase transform as input data for convolutional neural networks, and SI-CNN, which is based on feeding the features of the sound intensity into the convolutional neural network. Simulation and modeling results show that the SI-GCC-CNN model is effective in its generalization ability and outperforms the SI-CNN model, achieving an improvement in localization accuracy by 22,1% when changing the size of the room, by 15,6% when changing the location of the microphone array and by 32% when changing the distance between the source and the center of the microphone array.

Keywords: generalization ability; deep learning models, sound source localization; reverberant environment; orthogonal microphone array; sound intensity; generalized cross-correlation – phase transform; convolutional neural networks.

***For citation*:** Shahoud, G.M., Agafonov, E.D. (2025) Evaluating the generalization ability of deep learning models for sound source localization. *Vestnik Tomskogo gosudarstvennogo universiteta. Upravlenie, vychislitelnaja tehnika i informatika – Tomsk State University Journal of Control and Computer Science*. 72. pp. 107–113. doi: 10.17223/19988605/72/11

Научная статья
doi: 10.17223/19988605/72/11

Оценка обобщающей способности моделей глубокого обучения для локализации источника звука

Джиах Михаил Шаход[1], Евгений Дмитриевич Агафонов[2]

[1, 2] Сибирский федеральный университет, Красноярск, Россия

[1] ghiathlovealaa@gmail.com
[2] evgeny.agafonov@mail.ru

Аннотация. Оценивается обобщающая способность моделей глубокого обучения, используемых для решения задачи локализации источника звука с пространственным разрешением 10°, при изменении настроек конфигурации. Обобщающая способность моделей оценивалась в замкнутой реверберирующей среде с использованием ортогональной микрофонной решетки. Были рассмотрены две модели: SI-GCC-CNN, которая основана на объединении признаков интенсивности звука и обобщенной кросс-корреляции – фазового преобразования в качестве входных данных для сверточных нейронных сетей, и SI-CNN, которая основана на подаче признаков интенсивности звука в сверточную нейронную сеть. Результаты моделирования и имитации показывают, что модель SI-GCC-CNN эффективна по своей обобщающей способности и превосходит модель

SI-CNN, достигая улучшения точности локализации на 22,1% при изменении размера помещения, на 15,6% при изменении местоположения микрофонной решетки и на 32% при изменении расстояния между источником и центром микрофонной решетки.

Ключевые слова: обобщающая способность; модели глубокого обучения; локализация источника звука; реверберирующая среда; ортогональная микрофонная решетка; интенсивность звука; обобщенная кросс-корреляция – фазовое преобразование; сверточные нейронные сети.

***Для цитирования*:** Шаход Д.М., Агафонов Е.Д. Оценка обобщающей способности моделей глубокого обучения для локализации источника звука // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. 2025. № 72. С. 107–113. doi: 10.17223/19988605/72/11

Introduction

The problem of sound source localization (SSL) can be defined as estimating the direction of acoustic sources or objects that reflect acoustic signals, which can be applied in various complex environments. SSL is an active research topic in the field of acoustic signal processing using microphone arrays, it has many practical applications in the fields of engineering and technology, such as automatic camera tracking for teleconferencing, human-robot interaction, hearing aids, and remote speech recognition. SSL is also of paramount importance in geophysics and non-destructive testing of materials.

To acquire acoustic signals with further analysis of their characteristics, microphone arrays are used, which consist of a set of microphones located in space in a certain way to obtain spatial information about the acoustic source. The spatio-temporal information obtained from the microphone array can be used to estimate various source parameters (direction, distance).

Initially, the problem of SSL has been solved using traditional signal processing methods, such as time difference of arrival (TDOA) [1], delay-and-sum beamformer (DAS) [2], multiple signal classification (MUSIC) [3] and generalized cross-correlation - phase transform (GCC-PHAT) [4]. However, these methods have drawbacks due to the complexity of the acoustic characteristics of the environment, especially in the presence of noise and echoes [5]. In recent years, with the advent and development of deep learning (DL) methods and deep neural networks (DNNs) and their widespread use in the field of acoustic applications, a new vector for the development of SSL has been outlined.

The main advantage of SSL based on DL methods is the inclusion of information on acoustic characteristics in the learning process, while traditional methods are based only on spatial information [4]. As a result, data-driven methods such as DL could outperform traditional methods by dealing with large amounts of data, real or simulated. On the other hand, they are less able to generalize than traditional methods [5].

Designing DNNs for a specific application often requires exploring (and possibly combining) different architectures and tuning their hyperparameters. This has been the case with SSL in the last decade, and the evolution of DL-based SSL methods has followed the general evolution of DNNs towards more complex architectures or new efficient models. In other words, the DNN architectures used in SSL were often inherited from other applications (related or more distant fields) simply because they have been shown to work well with acoustic signals [6]. The literature relied on the same methodology, where different models were often combined (in parallel and/or sequentially), such as convolutional neural networks (CNN) [7], recurrent neural network (RNN) [8], convolutional recurrent neural network (CRNN) [9] and residual neural networks (ResNet) [10].

The effectiveness of DL-based SSL model is determined by its ability to generalize various aspects of the configuration (for example, the distance between the source and the microphone array, noise levels, reverberation time, etc.), i.e. the ability to correctly classify new test data with features that differ from those obtained during training and for different configuration settings. The ability of these models to generalize in noisy and reverberant environments using small-sized microphone arrays remains a challenging task.

The use of a sound intensity (SI) vector as input features for the DL-based SSL model was first proposed in [11], where superior performance has been demonstrated compared to traditional methods. SI as input features for CNN has proven its ability to work under noise and reverberation conditions when using small-sized microphone arrays [12], however, this deep model has not been tested for its ability to generalize when

changing some modeling conditions, such as room size and source distance. Although GCC-PHAT features do not give good localization results when working with small-sized microphone arrays, they have proven to be capable of generalization [13], because they depend on spatial information, while SI features depend on physical characteristics of sound (pressure and particle velocity).

In [14], a deep model with a spatial resolution of 10° was proposed to solving the SSL problem in a closed reverberant environment by integrating SI and GCC-PHAT features as input data for CNNs to utilize the advantages of these features.

In this paper, the ability of the proposed model in [14] to generalize when the configuration settings are changed will be tested.

1. Evaluation metric

In order to evaluate the effectiveness of the models, the localization accuracy is used as a performance measure, which is defined as:

$$PA(\%) = \frac{N_p}{N_s} \times 100, \tag{1}$$

where N_s represents the total number of source directions being evaluated and N_p is the number of source directions correctly recognized. The direction of the source is considered to be correctly recognized if the predicted direction is within the spatial resolution of the model, that is, the deviation of the predicted direction from the actual direction is within $\pm\theta_0$ for spatial resolution θ_0 [13].

2. Evaluating the Model's Ability to Generalize

The generalization ability of both the proposed model in [14] and the SI-CNN model [12] is evaluated. In the model SI-CNN, an improved feature extraction scheme based on SI estimation was proposed by decoupling the correlation between sound pressure and particle velocity components in the whitening construction, and feeding these features into CNN, which in turn estimates the direction of the source.

The SI-CNN model was trained and validated under the same simulation conditions and on the same training and validation dataset [14]. The training sample size was 6000 samples for each of the two models, the validation dataset size was 1000 samples. The ability of trained models to generalize when changing the configuration settings (modeling conditions) that were assumed when training the models will be considered.

Three settings will be changed (room size, microphone array location and distance between the source and the center of the microphone array) and the trained models will be re-evaluated on a new dataset that is generated taking into account the change in modeling conditions.

2.1. Changing the room size (y-dimension)

20 different room sizes are considered while maintaining the same shape, where the y-dimension of the previously defined room varies from 4 m to 20 m and in 19 equal steps. The other two dimensions x and z change while maintaining constant ratios with the y-dimension. Here attention is drawn to the need to vary RT60 corresponding to each size, according to the Sabin formula [15], as RT60 increases with the increase in room size, and therefore small-sized rooms reverberate less than large-sized rooms. The room size ranges from (6,67, 4, 1,78) m to (33,33, 20, 8,89) m. To create a test dataset, 200 sentences are randomly taken from the TIMIT test database and 200 random directions are generated. For each room size, 10 test samples are generated, and a total of 200 test samples are generated. The performance of the pertained models is evaluated on test samples corresponding to each room size.

A graph of the localization accuracy of both the proposed model and the SI-CNN model when changing the size of the room is presented in Fig. 1.

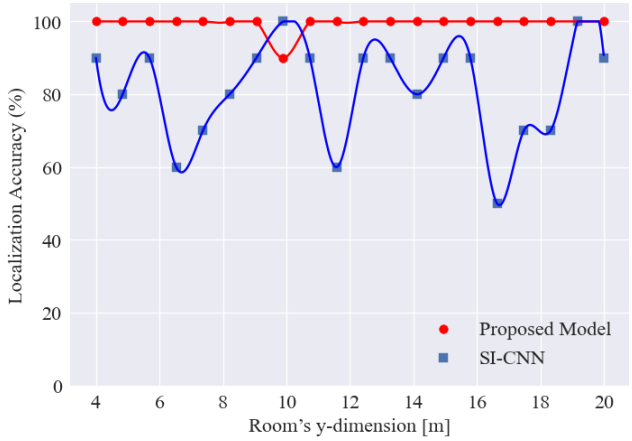

Fig. 1. Evaluating the ability of both the proposed model and SI-CNN model to generalize when changing the room size

From the simulation results presented in Fig. 1, it is clear that the proposed model proved to be effective in generalization when changing the size of the room and a localization accuracy with an average value of 99,5% was achieved, while the SI-CNN model achieved an average accuracy of 81,5%.

2.2. Changing the microphone array location (Center distance)

20 different locations of the center of the microphone array in the room are considered, with the location of the center varying from (7,5, 4,5, 1,5) m (the first location adopted in the modeling process) to (12,86, 6,86, 0,14) m and with 19 equal steps. To create a test dataset, 200 sentences are randomly taken from the TIMIT test database and 200 random directions are generated. For each center location, 10 test samples are generated, for a total of 200 test samples.

A graph of the models' localization accuracy when changing the microphone array location is shown in Fig. 2 (x-axis represents the distance between each center location of the microphone array and the first location).

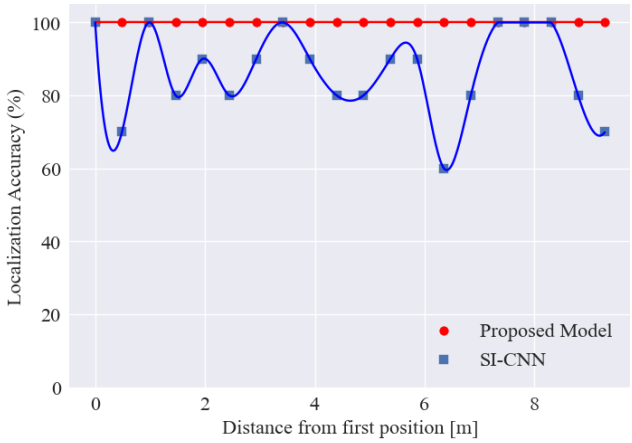

Fig. 2. Evaluating the ability of both the proposed model and SI-CNN model to generalize when changing the microphone array location

The proposed model achieved a localization accuracy of 100%, while the SI-CNN model achieved an average accuracy of 86,5%.

2.3. Changing the distance between the source and the center of the microphone array (Source distance)

20 different distances between the acoustic source and the center of the microphone array are considered, with the distance varying from 2,1 m to 4,4 m and in 19 equal steps. To create a test dataset, 200 sentences are

randomly taken from the TIMIT test database and 200 random directions are generated. For each distance, 10 test samples are generated, a total of 200 test samples are generated.

A graph of the models' localization accuracy when changing the distance between the source and the center of the microphone array is shown in Fig. 3. The proposed model achieved a localization accuracy of 99%, while the SI-CNN model achieved an average accuracy of 75%.

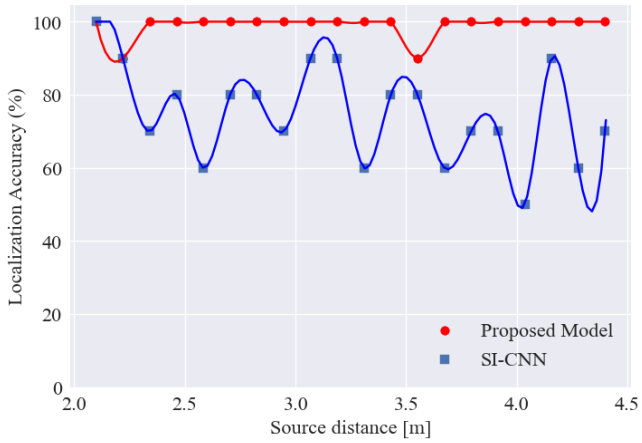

Fig. 3. Evaluating the ability of both the proposed model and SI-CNN model to generalize when changing the distance between the source and the center of the microphone array

Table shows the values of the localization accuracy metric for both the SI-CNN model and the proposed model with a spatial resolution of 10° when changing the room size, the location of the microphone array and the distance between the source and the center of the array. The average localization accuracy for each of the two models is calculated, and then the improvement rate in localization accuracy is calculated.

Localization accuracy of SI-CNN model and proposed model when changing the configuration settings

y-dimension (m)	SI-CNN (%)	Proposed (%)	Center distance (m)	SI-CNN (%)	Proposed (%)	Source distance (m)	SI-CNN (%)	Proposed (%)
4	90	100	0	100	100	2,1	100	100
4,84	80	100	0,49	70	100	2,22	90	90
5,68	90	100	0,98	100	100	2,34	70	100
6,53	60	100	1,47	80	100	2,46	80	100
7,37	70	100	1,95	90	100	2,58	60	100
8,21	80	100	2,44	80	100	2,71	80	100
9,05	90	100	2,93	90	100	2,83	80	100
9,89	100	90	3,42	100	100	2,95	70	100
10,74	90	100	3,91	90	100	3,07	90	100
11,58	60	100	4,4	80	100	3,19	90	100
12,42	90	100	4,89	80	100	3,31	60	100
13,26	90	100	5,37	90	100	3,43	80	100
14,11	80	100	5,86	90	100	3,55	80	90
14,95	90	100	6,35	60	100	3,67	60	100
15,79	90	100	6,84	80	100	3,79	70	100
16,63	50	100	7,33	100	100	3,92	70	100
17,47	70	100	7,82	100	100	4,04	50	100
18,32	70	100	8,31	100	100	4,16	90	100
19,16	100	100	8,8	80	100	4,28	60	100
20	90	100	9,28	70	100	4,4	70	100
Average	81,5	99,5	Average	86,5	100	Average	75	99

The simulation results shown in the table demonstrate that the proposed model is highly effective in its generalization ability and outperforms the SI-CNN model, achieving an improvement rate of 22,1% in localization accuracy when changing the room size, 15,6% when changing the location of the microphone array, and 32% when changing the distance between the source and the center of the microphone array.

Conclusion

The generalizability of both the proposed model in [14] and the SI-CNN model based on the feeding of SI features into CNN was evaluated when changing the configuration settings. The simulation results demonstrated that the proposed model is highly effective and outperforms the SI-CNN model, and also achieves better performance in generalization ability when changing settings, especially the distance between the source and the center of the array. An improvement rate in localization accuracy was achieved by 22,1% when changing the size of the room, by 15,6% when changing the location of the microphone array and by 32% when changing the distance between the source and the center of the microphone array.

Finally, after the effectiveness of the proposed model in generalization has been proven, future work will be to extend the proposed model to be able to localize multiple sound sources. This can be achieved through integration between the proposed model and a model for separating multiple sound sources, where a method will be applied to separate the sound sources, and then the proposed model will be applied to each source to estimate its direction.

References

1. Zhu, N. & Reza, T. (2019) A modified cross-correlation algorithm to achieve the time difference of arrival in sound source localization. *Measurement and Control*. 52(3-4). pp. 212–221. DOI: 10.1177/0020294019827977
2. Chiariotti, P., Martarelli, M. & Castellini, P. (2019) Acoustic beamforming for noise source localization – Reviews, methodology and applications. *Mechanical Systems and Signal Processing*. 120. pp. 422–448.
3. Zhong, Y., Xiang, J., Chen, X., Jiang, Y. & Pang, J. (2018) Multiple Signal Classification-Based Impact Localization in Composite Structures Using Optimized Ensemble Empirical Mode Decomposition. *Applied Sciences*. 8(9). pp. 1447.
4. Desai, D. & Mehendale, N. (2022) A Review on Sound Source Localization Systems. *Archives of Computational Methods in Engineering*. 29(7). pp. 4631–4642. DOI: 0.1007/s11831-022-09747-2
5. Shahoud, G.M. & Agafonov, E.D. (2024) Analysis of Approaches and Methods to Acoustic Sources Localization. *Journal of Siberian Federal University. Engineering & Technologies*. 17(3). pp. 380–398.
6. Grumiaux, P.A., Kitić, S., Girin, L. & Guérin, A. (2022) A survey of sound source localization with deep learning methods. *Journal of the Acoustical Society of America*. 152(1). pp. 107–151.
7. Nguyen, T.N.T., Gan, W.S., Ranjan, R. & Jones, D.L. (2020) Robust source counting and DOA estimation using spatial pseudo-spectrum and convolutional neural network. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 28. pp. 2626–2637. DOI: 10.1109/TASLP.2020.3019646
8. Nguyen, T.N.T., Nguyen, N.K., Phan, H., Pham, L., Ooi, K., Jones, D.L. & Gan, W.S. (2021) A general network architecture for sound event localization and detection using transfer learning and recurrent neural network. *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* pp. 935–939. DOI: 10.1109/ICASSP39728.2021.9414602
9. Adavanne, S., Politis, A., Nikunen, J. & Virtanen, T. (2018) Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. *IEEE Journal of Selected Topics in Signal Processing*. 13(1). pp. 34–48.
10. He, W., Motlicek, P. & Odobez, J.M. (2019) Adaptation of multiple sound source localization neural networks with weak supervision and domain-adversarial training. *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* pp. 770–774.
11. Perotin, L., Serizel, R., Vincent, E. & Guérin, A. (2018) CRNN-based joint azimuth and elevation localization with the Ambisonics intensity vector. *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*. pp. 241–245. DOI: 10.1109/IWAENC.2018.8521403
12. Liu, N., Chen, H., Songgong, K. & Li, Y. (2021) Deep learning assisted sound source localization using two orthogonal first-order differential microphone arrays. *Journal of the Acoustical Society of America.* 149(2). pp. 1069–1084. DOI: 10.1121/10.0003445
13. Li, Q., Zhang, X. & Li, H. (2018) Online direction of arrival estimation based on deep learning. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 2616–2620. DOI: 10.1109/ICASSP.2018.8461386
14. Shahoud, G.M. & Agafonov, E.D. (2024) A combined model for localizing acoustic sources using deep learning technology. *Vestnik Tomskogo gosudarstvennogo universiteta. Upravlenie, vychislitelnaya tekhnika i informatika – Tomsk State University Journal of Control and Computer Science*. 68. pp. 100–111. DOI: 10.17223/19988605/68/11
15. Ciaburro, G. & Iannace, G. (2021) Acoustic characterization of rooms using reverberation time estimation based on supervised learning algorithm. *Applied Sciences*. 11(4). Art. 1661. DOI: 10.3390/app11041661

Information about the authors:
Shahoud Ghiath M. (Post-Graduate Student, Siberian Federal University, Krasnoyarsk, Russian Federation). E-mail: ghiathlovealaa @gmail.com

Agafonov Evgeniy D. (Doctor of Technical Sciences, Professor, Siberian Federal University, Krasnoyarsk, Russian Federation). E-mail: eagafonov@sfu-kras.ru

Contribution of the authors: the authors contributed equally to this article. The authors declare no conflicts of interests.

Информация об авторах:

Шаход Джиах Михаил – аспирант кафедры систем автоматики, автоматизированного управления и проектирования Института космических и информационных технологий Сибирского федерального университета (Красноярск, Россия). E-mail: ghiathlovealaa@gmail.com

Агафонов Евгений Дмитриевич – доктор технических наук, профессор кафедры систем автоматики, автоматизированного управления и проектирования Института космических и информационных технологий Сибирского федерального университета (Красноярск, Россия). E-mail: evgeny.agafonov@mail.ru

Вклад авторов: все авторы сделали эквивалентный вклад в подготовку публикации. Авторы заявляют об отсутствии конфликта интересов.

Received 21.04.2025; accepted for publication 02.09.2025

Поступила в редакцию 21.04.2025; принята к публикации 02.09.2025