INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Assessment of von Mises–Bernoulli Deep Neural Network in Sound Source Localization

Katsutoshi Itoyama[1], *Yoshiya Morimoto*[1], *Shungo Masaki*[1], *Ryosuke Kojima*[2], *Kenji Nishida*[1], *Kazuhiro Nakadai*[1,3]

[1]Tokyo Institute of Technology, Japan
[2]Kyoto University, Japan
[3]Honda Research Institute Japan, Co., Ltd., Japan

{itoyama, morimoto, masaki, nishida, nakadai}@ra.sc.e.titech.ac.jp
kojima.ryosuke.8e@kyoto-u.ac.jp

Abstract

This paper addresses the properties and effectivenss of the von Mises–Bernoulli deep neural network (vM-B DNN), a neural network capable of learning periodic information, in sound source localization. The phase, which is periodic information, is an important cue in sound source localization, but typical neural network cannot handle periodic input values properly. The vM-B DNN has been theoretically revealed to be able to handle periodic input values and its effectiveness has been shown in a simple case study of sound source localiation using artificial sinusoids, but it was not in the case of speech signals. We conducted both numerical simulation and actual environment experiments. We compared a sound source localization method using vM-B DNN with those using ordinary neural networks, and showed that the vM-B DNN outperforms other methods under various conditions.

Index Terms: sound source localization, deep neural networks, von Mises distribution

1. Introduction

Sound source localization is one of the most important functions in speech and audio signal processing for dialogue systems and environment understanding. A microphone array, which is a device consisting of multiple synchronous microphones, can be used to achieve robust localization methods such as multiple signal classification (MUSIC) [1] and steered-response power phase transform (SRP-PHAT) [2]. MUSIC-based source localization methods use a transfer function that represents the relationship of sound propagatiion between a sound source and a microphone array in terms of both phase and amplitude. The transfer function is usually obtained either by calculating the geometric relationship between the sound source and the microphone array or by measuring impulse responses from the source to the array. The performance of sound source localization depends on how the transfer function obtained in advance matches that of the actual environment.

On the other hand, in recent years, deep learning approaches have been applied to the field of sound source localization. Takeda et al. [3] proposed a modified MUSIC method that introduces a subband-based hierarchical neural network instead of using a spatial spectrum calculation block in MUSIC. Their method achieved better performance than the original MUSIC. Nelson et al. [4] proposed a sound source localization method using a 19-layer neural network and reported better localization performance than the original MUSIC method. Since ordinary neural networks can only input real numbers, their sound source localization was limited to inputting only amplitude information, whereas the original MUSIC method input complex numbers that included both phase and amplitude information. Nevertheless, the performance of their sound source localization methods was still better than the original MUSIC method.

The phase, which is a periodic quantity of 2π or 360°, is an important cue for source localization methods such as MUSIC and SRP-PHAT. On the other hand, most neural networks are designed with the assumption that the input is an ordinary real number that does not have periodicity. There are several approaches to overcome the limitation that only real numbers can be used as inputs. One is to extend the neural network to allow input of complex numbers, which is called complex valued neural networks [5, 6, 7]. The other is simpler, that is, decomposing the complex numbers into real and imaginary parts and inputting each of them into the neural network [8]. These method do not include an explicit model for extracting both phase and amplitude information from the input, and there is no guarantee that these two types of information will be properly extracted and effectively trained. To solve this problem, a new activation function has been proposed by replacing the simple sigmoid function [9]. This activation function is inspired by a von Mises distribution that is a probabilistic distribution for a periodical value. They showed the validity of their proposed activation function through a discussion using restricted Boltzmann machines but the effectiveness of their method has been shown only on a simple case study. Thus we focus on the comparison of several neural networks using their activation function or a typical activation function through both numerical simulation and experimental evaluation.

The contributions of this study are the following:

1. A detailed analysis of the properties of the vM-B DNN using simulated data.
2. A new database of actual recordings was created to verify the effectiveness of vM-B DNN-based source localization method.

Section 2 reviews previous work on sound source localization and described vM-B DNN [9] which solves the problem of learning periodic information. Section 3 describes the sound source localization method using vM-B DNN, Section 4 gives a detailed assessment of the source localization method, and Section 5 concludes this paper.

2. Related Work

This section reviews previous work on sound source localization, and in particular discusses the properties of the methods based on deep learning. Then we discuss von Mises–Bernoulli DNN (vM-B DNN) [9], which solves the problem of learning

http://dx.doi.org/10.21437/Interspeech.2021-1050

periodic information, that was a challenge for many deep learning methods.

2.1. Sound Source Localization

Sound source localization (SSL) is a technique that estimates the direction of arrival (DOA) or location of sound sources from the acoustic signal recorded by a microphone array. One of the widely used SSL algorithms is steered-response power phase transform (SRP-PHAT) [2]. SRP-PHAT searches the source DOA that maximizes the steered-response power, like the output power of steered delay-and-sum beamformer, to various directions from the microphone array observation. Another widely used SSL algorithm is multiple signal classification (MUSIC) [1, 10]. MUSIC method exploits the orthogonality between the eigenvectors corresponding to the sound source and those corresponding to other noises in the spatial correlation matrix of the observed signal, and shapens the peaks of the directional likelihood function. The likelihood obtained by the MUSIC method is often referred to as the MUSIC spectrum.

In recent years, deep learning-based SSL methods have been reported. The network structures include simple fully connected network [3], convolutional neural network (CNN) [4, 11, 12], and recurrent neural network (RNN) [13]. In many cases, DOA estimation is formulated as a classification problem. For 2D source localization, it is formulated as a 72-class classification problem in which 360° of azimuth angle is divided into 5° increments or a 360-class problem with 1° increments. Inputs to the network are often amplitude/power spectra, phase spectra, and complex spectra with real and imaginary representations.

2.2. von Mises–Bernoulli DNN

While deep learning-based techniques have achieved good performance in source localization, the challenge is that many networks that take the phase spectrum as input do not handle the phase properly. The phase is usually represented as a value in the range of 0° to 360°, and many network inputs this value as it is or convert it to a range such as 0 to 1. Adopting such a phase representation, the angle between 1° and 359° is calculated to be 358°, even though geometically it should be 2°.

The von Mises–Bernoulli DNN (vM-B DNN) [9] is a neural network that can properly handle periodic information including phase. The vM-B DNN was devised with reference to the von Mises distribution [14], one of the probabilistic distributions that can represent periodic numbers such as directions, and its application to the restricted Boltzmann machine (RBM). The vM-B DNN is characterized such that the input layer is represented by

$$y_i = f\left(\sum_j \left(A_{ij}\cos x_j + B_{ij}\sin x_j\right) + c_i\right), \tag{1}$$

where $\boldsymbol{x}$ and $\boldsymbol{y}$ are the input and output the input layer, $\boldsymbol{A}$, $\boldsymbol{B}$ are the weight parameters, and $\boldsymbol{C}$ is the bias parameter. $f(\cdot)$ is an arbitrary activation function such as sigmoid and rectified linear unit (ReLU).

Equation (1) is derived as follows. Since the phase has a period of 2π, the output of a node that takes a single phase input must satisty

$$f(x) = f(x + 2n\pi), \quad n = 0, \pm 1, \pm 2, \ldots. \tag{2}$$

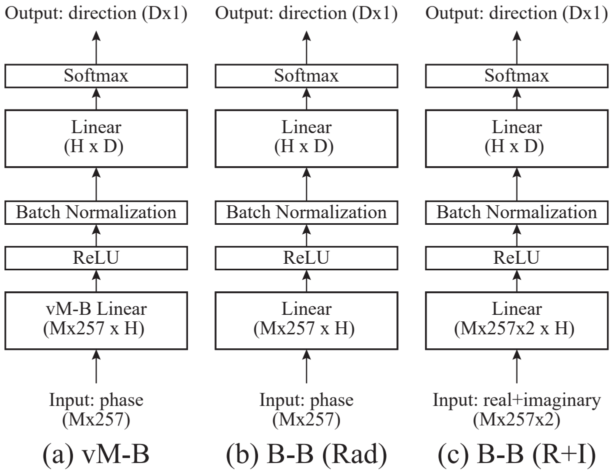


Figure 1: *Network structure for vM-B DNN and B-B DNNs.*

The general type of a function satisfying this condition is that

$$f(x) = g\left(\sum_{n=-\infty}^{\infty} \left(a_n\cos(nx) + b_n\sin(nx)\right)\right), \tag{3}$$

and simplifying this equation to the case where only $n = 0, 1$, we get

$$f(x) = g(a_0 + a_1\cos x + b_1\sin x). \tag{4}$$

This has the same form as the equation (1), which shows that it is a reasonable expression for dealing with periodic inputs.

Nakadai et al. [9] showed that vM-B DNNs are more effective than general fully connected networks in source localization through numerical simulation experiments, but the scope of their evaluation was limited. In this study, we aim to clarify the properties of vM-B DNN through detailed analysis and to demonstrate its effectiveness through source localization evperiments in real environements.

3. Sound Source Localization Method

This section describes the sound source localization method we have developed to assess the vM-B DNN. In this paper, we refer to the general fully connected network as Bernoulli-Bernoulli (B-B) DNN, as opposed to vM-B DNN, following the report by Nakadai et al. [9].

We use the three-layer fully connected network shown in Figure 1. The sizes of the input, hidden, and output layers are N, H, and D, resepctively. The activation functions for the input layer is ReLU, and for the hidden layer is softmax. Batch normalization is applied before linear operations on the hidden layer. The input features for the vM-B DNN are the phase spectrum. The input features for the B-B DNN are the real and imaginary representation of the complex spectrum or the phase spectrum. Each condition is labeled as *vM-B*, *B-B_R+I*, and *B-B_Rad*. Thus, a total of three different conditions were compared. The output of the network is a one-hot vector representing the source direction.

The acoustic signal observed by the M-channel microphone array is transformed into a complex spectrogram by the short-time Fourier transform (STFT). Frames with a volume below a predetermined threshold are excluded from the spectrogram because such frames do not contribute to the localization

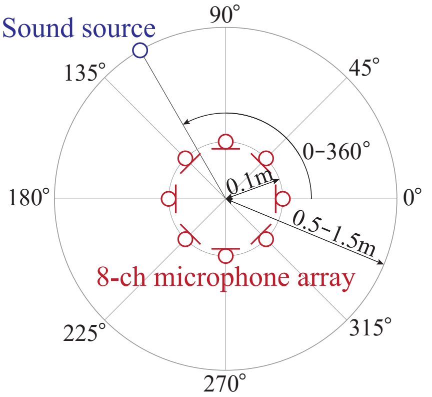


Figure 2: *2D virtual recording environment. The microphone array and the sound source were both placed at the same plane in the space.*

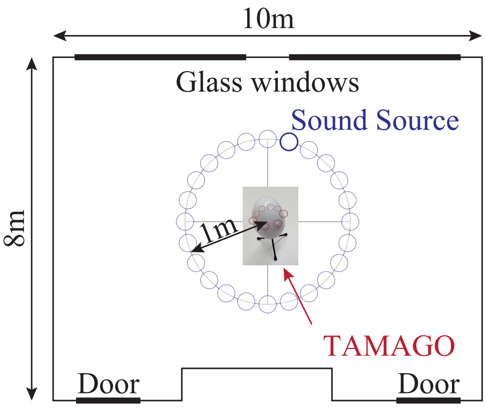


Figure 3: *Plain view of the actual recording environment.*

Figure 4: *Scene of the actual recording. TAMAGO and the sound source (loudspeaker) were both placed at a height of* 1 m *above the floor.*

Table 1: *Experimental condition.*

Parameter	Value
Size of the input layer N	2056 for vM-B and B-B_Rad 4112 for B-B_R+I
Size of the hidden layer H	256
Size of the output layer D	72
STFT window	512 points of hann window
STFT hop size	160
Volume threshold	−50 dB
Optimizer	Adam
Batch size	8192
Epochs (maximum)	100

of sound sources. For the frames with sufficient volume, the real ang imaginary parts of the complex spectrum or the phase spectrum are extracted and used as input features for the networks.

4. Evaluation

In order to clarity the effectiveness of vM-B DNN in sound source localization, experiments were conducted using speech recorded in a simulation environment and a real environment. In the simulation environment, the source localization performance was measured while changing the signal-to-noise ratio. In the real environment, the performance was evaluated using the speech signal recorded in a meeting room with reverberation and background noise. The accuracy of the source localization was used for criterion. We use 10-fold cross validation to train and test the network, thus the final accuracy shown below is the average over the 10 validation runs.

The parameter settings for feature extraction and network training are shown in Table 1. All networks were implemented using Pytorch framework. A categorical cross-entropy [15] was used as the loss function.

4.1. Dataset

The environment of the numerical simulation is shown in Figure 2. An eight-channel circular planar microphone array with a radius of 0.1 m is placed in a two-dimensional virtual space with no reflections or reverberations. The sampling frequency of the microphone array is 16 kHz. A sound source is placed in a uniformly random direction between 0° to 360°, at a uniformly random distance between 0.5 m to 1.5 m. A total of 13,000 utterances from “parallel100” and “nonpara30” of the JVS corpus [16], a dataset of speech utterances of 100 professional speakers, were used as sound sources. The sound source signals were played back after down-sampling to 16 kHz. The transfer function between the sound source and the microphone array is calculated geometrically. The microphone array records the sum of the source signal, which is convolved with the transfer function, and the white noise, which is independent of each microphone. The signal-to-noise ratio (SNR) was set to $+\infty$, 40, 30, 20, 10 and 0 dB. The total duration of the recorded signal is 26 h for each SNR condition.

The environment for the actual recordings is shown Figure 3. An eight-channel circular microphone array, TAMAGO[1], was placed in the center of a 10 m×6 m×4 m room at a height of 1 m. The sampling frequency of the microphone array is 16 kHz. The sound source was placed in a circle with a radius of 1 m around the microphone array, and in 72 directions in increments of 5°, with the front of the microphone array at 0°. The source signals are text readings by 20 male and 20 female speakers each, randomly selected from JNAS [17]. 40 utterances were played back from each direction and recorded by the array. The total duration of the recorded signal is 4.5 h.

Table 2: *Experimental result of numerical simulation.*

	SNR/dB					
	$+\infty$	40	30	20	10	0
Model	Accuracy/%					
vM-B (Proposed)	100	98.8	96.6	69.3	46.2	22.4
B-B_Rad	100	98.4	94.8	62.4	31.4	13.2
B-B_R+I	99.9	96.5	90.2	55.8	25.2	8.4

Table 3: *Experimental result using actual recordings.*

Model	Accuracy/%
vM-B (Proposed)	91.7
B-B_Rad	80.4
B-B_R+I	72.3

4.2. Experimental Results

The result of the numerical simulation is shown in Table 2. For all SNRs, vM-B shows the highest source localization accuracy. The result of the experiment using actual recordings is shown in Table 3. As same as the simulation experiment, vM-B shows the highest accuracy. These results indicate the effectiveness of the vM-B DNN for speech localization. Despite the fact that B-B_Rad utilizes only phase, i.e., poorer information than B-B_R+I, B-B_Rad has a higher accuracy than B-B_R+I in all SNRs and datasets. This result suggests that phase is the mandatory cue for speech localization and amplitude is the secondary (and sometimes unnecessary) due to the sparseness of the speech signals.

Figure 5 shows the accuracies for training and validation subsets on the real environment dataset. For the training dataset, the accuracy continues to increase, but for the validation dataset, the accuracy converges at around 20 epochs, indicating that the networks were sufficiently trained.

5. Conclusion

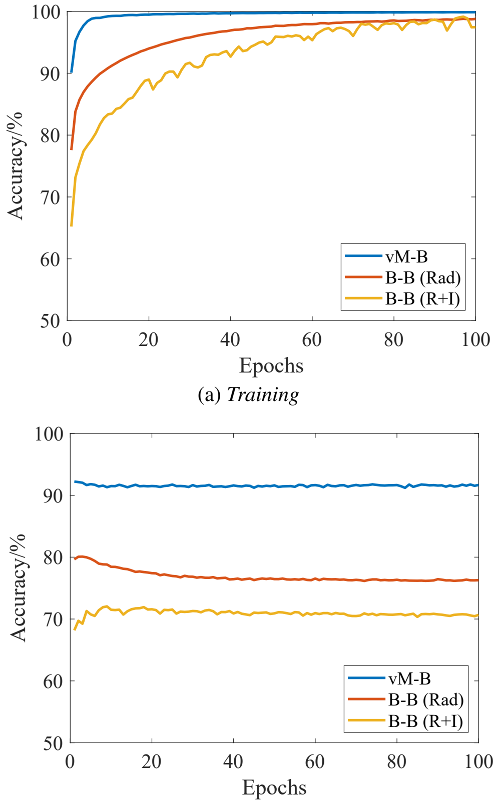
This paper presented the effectiveness of the von Mises–Bernoulli deep neural network (vM-B DNN) in sound source localization. Two datasets were constructed by recording speech signals with a microphone array in a simulated virtual environment and a actual environment. We compared the vM-B DNN method with a conventional neural network method for source localization, and showed that the vM-B DNN method was effective for both datasets. There are a lot of issues for practical use such as multiple sound sources, dynamically-changing environment, efficient training mechanism, and so on. The performance evaluation in the real world by introducing real robots and systems also remain as crucial future work.

6. Acknowledgements

This work was supported by JSPS KAKENHI Grant No. JP19K12017, JP19KK0260 and JP20H00475.

7. References

[1] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.



(a) *Training*

(b) *Validation*

Figure 5: *Accuracy for (a) training and (b) validation subsets on the real environment.*

[2] J. H. DiBiase, "A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays," Ph.D. dissertation, Brown University, 2000.

[3] R. Takeda and K. Komatani, "Sound source localization based on deep neural networks with directional activate function exploiting phase information," in *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 405–409.

[4] N. Yalta, K. Nakadai, and T. Ogata, "Sound source localization using deep learning models," *Journal of Robotics and Mechatronics*, vol. 29, no. 1, pp. 37–48, 2017.

[5] W.-H. Yang, K.-K. Chan, and P.-R. Chang, "Complex-valued neural-network for direction-of-arrival estimation," *Electronics Letters*, vol. 30, no. 7, pp. 574–575, 1994.

[6] H. Tsuzuki, M. Kugler, S. Kuroyanagi, and A. Iwata, "An approach for sound source localization by complex-valued neural network," *IEICE Trans. on Information and Systems*, vol. 96, no. 10, pp. 2257–2265, 2013.

[7] K. Terabayashi, R. Natsuaki, and A. Hirose, "Ultrawideband direction-of-arrival estimation using complex-valued spatiotemporal neural networks," *IEEE Trans. on Neural Networks and Learning Systems*, vol. 25, no. 9, pp. 1727–1732, 2014.

[8] W. Ma and X. Liu, "Phased microphone array for sound source localization with deep learning," arXiv:1802.04479, Tech. Rep., 2018.

[9] K. Nakadai, S. Masaki, R. Kojima, O. Sugiyama, K. Itoyama, and K. Nishida, "Sound source localization based on von-Mises-Bernoulli deep neural network," in *2020 IEEE/SICE International Symposium on System Integration (SII)*, 2020, pp. 658–663.

[10] K. Nakamura, K. Nakadai, F. Asano, Y. Hasegawa, and H. Tsujino, "Intelligent sound source localization for dynamic environ-

[1] https://www.sifi.co.jp/product/microphone-array/

ments," in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2009, pp. 664–669.

[11] W. He, P. Motlicek, and J. Odobez, "Deep neural networks for multiple speaker detection and localization," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 74–79.

[12] D. Salvati, C. Drioli, and G. L. Foresti, "Exploiting cnns for improving acoustic source localization in noisy and reverberant conditions," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 103–116, 2018.

[13] W. Xue, Y. Tong, C. Zhang, G. Ding, X. He, and B. Zhou, "Sound event localization and detection based on multiple DOA beamforming and multi-task learning," in *INTERSPEECH 2020*, 2020, pp. 5091–5095.

[14] K. Mardia and P. E. Jupp, *Directional Statistics*. Wiley, 1999.

[15] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning*. Springer, 2009.

[16] S. Takamichi, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, , and H. Saruwatari, "JVS corpus: free Japanese multi-speaker voice corpus," arXiv:1908.06248, Tech. Rep., 2019.

[17] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," *Journal of the Acoustical Society of Japan (E)*, vol. 20, no. 3, pp. 199–206, 1999.