Multi-Speaker Localization Based on von Mises-Bernoulli ViViT

1st Haruto Yokota
Dept. Systems and Control Engineering
School of Engineering
Institute of Science Tokyo
Tokyo, Japan

2nd Benjamin Yen
Dept. Systems and Control Engineering
School of Engineering
Institute of Science Tokyo
Tokyo, Japan
benjamin@ra.sc.eng.isct.ac.jp

3rd Kazuhiro Nakadai
Dept. Systems and Control Engineering
School of Engineering
Institute of Science Tokyo
Tokyo, Japan
nakadai@ra.sc.eng.isct.ac.jp

***Abstract*—This paper addresses the problem of multi-speaker sound source localization (SSL). Previously, we developed a neural network that integrates the von Mises Bernoulli (vM-B) distribution into a Residual Network (ResNet) architecture, enabling robust SSL by leveraging the periodicity of phase information to mitigate environmental noise. Building on this, we proposed a Video Vision Transformer (ViViT)-based method for SSL, which not only demonstrated superior robustness in real-world environments compared to vM-B ResNet but also achieved higher localization accuracy. However, both methods were limited to single-source localization. In this study, we extend these approaches by modifying the ViViT architecture to accommodate multi-source sound localization. Additionally, we incorporate the von Mises Bernoulli distribution into the ViViT framework to further enhance robustness against varying environmental conditions. Experimental results confirm the effectiveness of the proposed method.**

***Index Terms*—sound source localization, multiple sources, periodic phase information, Video Vision Transformer (ViViT)**

I. Introduction

Sound source localization (SSL) is crucial for robot audition [1] and computational auditory scene analysis [2]. Traditional SSL methods, such as multiple signal classification (MUSIC) [3], often struggle in the presence of modeling errors, including low signal-to-noise ratios (SNRs), reverberations and transfer function mismatches. To address these challenges, deep learning-based SSL methods have been actively studied in recent years. Among these, Vision Transformers (ViT) have shown excellent performance in various domains, leading to the development of the Video Vision Transformer (ViViT) [4], which extends ViT to handle spatiotemporal patterns in video data. Motivated by the hypothesis that ViViT' s ability to model temporal dependencies could benefit SSL as well, we previously proposed a ViViT-based model for single-speaker sound source localization [5]. However, our model was limited to single-speaker audio input and did not support multiple speakers. Moreover, neural networks typically do not assume periodic distributions as inputs [6], [7]. This characteristic has not been explicitly considered in conventional approaches, resulting in phase information being fed directly into models without adaptation. To address these limitations, this study extends our ViViT-based model [5] in two key ways. First, it incorporates the von Mises Bernoulli (vM-B) distribution into the transformer embedding function to better account for phase periodicity. Second, it introduces multi-label classification to enable the localization of multiple sound sources. Our key contributions are:

- Multi-source localization: Reformulating SSL as a multi-label classification problem by applying binary cross-entropy (BCE) loss, making the model more suitable for real-world environments.
- Phase periodicity handling: Introducing the vM-B distribution into the ViViT embedding function to improve robustness against environmental variations.
- Detailed performance evaluation: Ablation tests and comparisons with other methods were conducted under various acoustic conditions, including noise, reverberation, and multiple sound sources, demonstrating the effectiveness of the proposed method.

II. Related work

Current SSL methods face two major challenges: traditional subspace techniques like MUSIC [3] degrade significantly in low - SNR or reverberant conditions, and many deep learning approaches overlook the periodic nature of phase and/or temporal context of input signals [8], [8], [9], [9], [10]. Both aspects are vital for robust localization.

Deep learning methods for multiple-source localization, especially those from the SELD (Sound Event Localization and Detection) community, have made progress. SELDnet uses a CRNN to perform frame-wise multi-label sound event detection and concurrent DoA regression, tracking static and moving sound events [8]. Variants utilizing squeeze - excitation residual blocks or sequential CRNN ensembles have further improved localization performance [11], [12]. Self-attention [9] and ViT-based models [10] also capture temporal dependencies but still lack explicit modeling of phase periodicity and often assume fixed numbers of sources.

Phase periodicity, inherent in inter-channel phase cues (0 = 360 degrees), is frequently overlooked as networks usually process phase as linear values—introducing angular discontinuity [13]. Notably, [6] and [7] represent the first DNNs to explicitly incorporate this periodicity into their

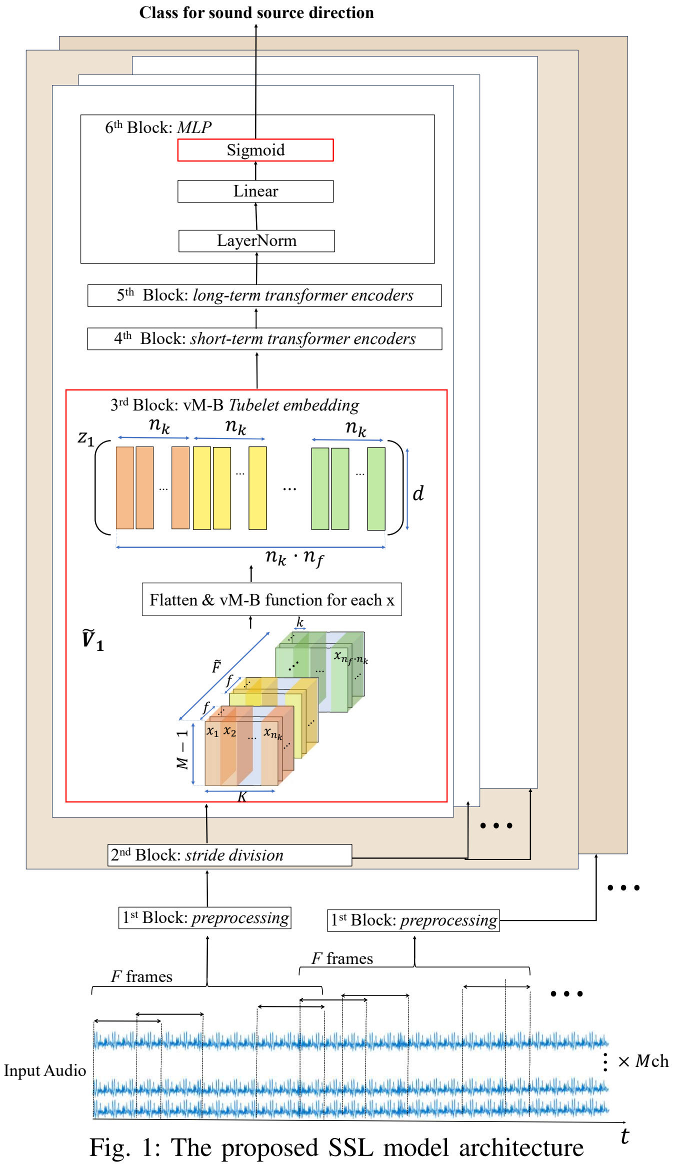

Fig. 1: The proposed SSL model architecture

architectures, demonstrating performance gains in SSL. The von Mises – Bernoulli (vM-B) distribution has been applied in DNNs (vM-B-DNN [7], vM-B ResNet [14]) to explicitly model this periodicity, improving SSL robustness. However, these approaches usually span only short temporal contexts and lack attention mechanisms, limiting their effectiveness in real-world, dynamic, and multi-source scenarios.

Our previous work [5] applied ViViT, a time-series-specialized ViT, to SSL, demonstrating its effectiveness in real-world settings. However, it was limited to single-source localization and could not handle multiple sources, a crucial requirement for practical applications.

III. Proposed Method

Fig.1 illustrates the architecture of the proposed model. Similar to the base model [5], it consists of six main blocks:

1) Preprocessing – Constructs a relative phase matrix from the input.
2) Stride Division – Divides the input matrix into submatrices with as long a temporal context as possible while keeping the number of submatrices unchanged.
3) vM-B Tubelet Embedding – Converts each submatrix into tokens.
4) Short-Term Transformer Encoders – Processes tokens with local features (short temporal context).
5) Long-Term Transformer Encoders – Processes tokens with global features (long temporal context).
6) MLP (Multi-Layer Perceptron) – Estimates the sound direction by integrating extracted features.

To enable multiple SSL and handle the periodicity of the input phase, we improved the 3rd and 6th blocks, respectively. These modified components are highlighted with red rectangles in Fig. 1 and are described in the following sections.

A. MLP (Multi-Layer Perceptron)

To extend the base model for multiple sound source scenarios, the classification task must be converted from multi-class classification to multi-label classification. To achieve this, the categorical cross-entropy (CCE) loss function is replaced with the sum of BCE losses for all output elements, following the approach in [15]. Additionally, to apply BCE loss, a sigmoid activation function is introduced at the final layer of the MLP in the 6th block. This ensures that all output elements fall within the range $[0, 1]$, making BCE loss computation feasible.

The BCE loss with logits is defined as follows:

$$\mathcal{L}_{BCE}(y,z) = \frac{1}{N}\sum_{i=1}^{N}\left[-y_i \log(\sigma(z_i)) - (1-y_i)\log(1-\sigma(z_i))\right], \quad (1)$$

$$\sigma(z_i) = \frac{1}{1+e^{-z_i}}, \quad (2)$$

where z_i denotes the model's logit output for the i-th element, $y_i \in 0, 1$ is the corresponding binary ground truth label, and $\sigma(\cdot)$ is the sigmoid function.

B. vM-B Tubelet Embedding

The vM-B Tubelet Embedding in the 3rd block addresses the periodicity of phase information. Tubelet Embedding, originally proposed in ViViT [4], extracts non-overlapping tensors from a 3D input, applies a linear transformation, and constructs tokens.

In the base model, tokenization was performed while preserving the correlation between microphones by avoiding division along the microphone axis during tensor extraction. The vM-B Tubelet Embedding extends this by incorporating the vM-B function into the linear transformation process. This enables phase-aware conversion of non-overlapping tensors $\boldsymbol{x}_{ij} (j = 1, 2, \ldots, n_k \cdot n_f)$ extracted from the input $\tilde{\boldsymbol{V}}_i$ into tokens $\boldsymbol{z}_{ij}$.

The token vector $\boldsymbol{z}_{ij}$ is computed as follows:

$$\boldsymbol{X}_{ij} = \text{Flatten}(\boldsymbol{x}_{ij}), \quad (3)$$

$$\boldsymbol{z}_{ij} = \boldsymbol{A}\cos(\boldsymbol{X}_{ij}) + \boldsymbol{B}\sin(\boldsymbol{X}_{ij}), \quad (4)$$

where Flatten() is a function that converts a three-dimensional tensor $\boldsymbol{x}_{ij} \in R^{f\times(M-1)\times k}$ into a vector $\boldsymbol{X}_{ij} \in$

$R^{(f\times(M-1)\times k)}$, and $\boldsymbol{A}, \boldsymbol{B} \in R^{d\times(f\times(M-1)\times k)}$ are weight matrices randomly initialized during training.

IV. Evaluation

For evaluation, we created four datasets and conducted three experiments using three different metrics.

A. Dataset

A $5 \times 8 \times 3\,\mathrm{m}$ room ($RT_{60} = 0.2\,\mathrm{s}$) was used. An 8-ch TAMAGO microphone array was placed at the center, and a speaker was positioned 1m away. Speech from the Centre for Speech Technology Research VKTS (CSTR-VKTS) corpus [16] was recorded in 72 directions at 5-degree intervals. Diffuse noise (BUS, CAF, PED, STR from CHiME4 [17]) was played from four corners. Recordings were in 16-bit, 16 kHz (see details in [5]).

We calculated RMS every second, removed segments with RMS 5% of the max, and chunked the rest into 30 - second segments as the recorded sound set.

- Dataset1: one speech chunk + one noise type at SNR = $-10, -5, 0, 20\,\mathrm{dB}$.
- Dataset2: mix of two speech chunks ($\geq$ 10-degree apart, different content) + one noise type at the same SNR levels.
- Dataset3 and Dataset4: same as Dataset1/2 but recorded in a more reverberant room ($RT_{60} = 0.6$) for testing in reverberant conditions.

For both Dataset1 and Dataset2, the SNR = -5 dB data was reserved exclusively for testing. The remaining data was split 8:1:1 into training, validation, and test sets.

B. Metrics

Three metrics were used. Mean Absolute Error (MAE) represents the average absolute error between the estimated and ground truth directions. Accuracy (Acc) was defined as the ratio of correct results to the total number of test samples, while Accuracy±5 (Acc±5) was defined as the proportion of results with an absolute error of 5 degrees or less relative to the total number of test samples.

Since SELDNet is based on a regression model, different definitions were applied to Acc and Acc±5. Specifically, Acc was calculated as the ratio of results with an absolute error of 2.5 degrees or less, while Acc±5 was calculated as the ratio of results with an absolute error of 7.5 degrees or less.

C. Experiments

To evaluate the performance of the proposed model, we conducted the following three experiments at four different SNR levels (-10, -5, 0, and 20 dB):

- Experiment 1: Effectiveness of each proposed technique (ablation test)
- Experiment 2: Performance comparison with other methods
- Experiment 3: Robustness to reverberation

Experiment 1 examined the effectiveness of each enhancement technique we applied to the original ViViT. Specifically, we conducted an ablation test on vM-B tubelet embedding and BCE loss, proposed in this paper, as well as stride division, introduced in [5]. For training and evaluation, we used a single-speaker dataset (Dataset1). Since BCE loss was introduced for SSL of multiple sound sources, in addition to the model trained on Dataset1, we also used a model trained on Dataset2, which contains a mix of single-source and two-source data. We then conducted comparative experiments, using Acc as the evaluation metric, to assess the impact of BCE loss versus the CCE loss for both single-source and two-source test data.

Experiment 2 compared the proposed method with MUSIC, vM-B ResNet, and SELDNet. MUSIC [3] is a signal processing-based localization method known for its robustness to noise. vM-B ResNet [14] is a ResNet-based localization method that incorporates phase information. SELDNet [8] is designed for simultaneous SSL and event detection. Since vM-B ResNet is originally designed for single-source localization, it was evaluated only in single-source scenarios. Although SELDNet is capable of handling multiple sources, it assumes that at most one source of the same type is present. As this study focuses solely on speech as the source type for evaluation, SELDNet was used only in the single-source localization scenario. The evaluation was conducted under two scenarios: one in which only a single source was present and another where both single and two-source cases were mixed. In the single-source scenario, the model was trained and evaluated using Dataset1, while in the mixed-source scenario, evaluation was performed using Dataset2. In all evaluations, it was assumed that the system knew the number of sources in advance. When only one source was present, the direction corresponding to the highest output value was considered the sound source direction. When two sources were present, the method of determining the source directions depended on the loss function. For BCE loss, the top two directions with the highest output values were selected as the source directions, while for CCE loss, the directions corresponding to the two highest peaks in the output vector were selected.

Experiment 3 assessed the robustness of the proposed method in reverberant environments by comparing it with MUSIC under two conditions: single-source only and mixed single+two-source cases were mixed. For the single-source scenario, we used a model trained on Dataset1, which has an RT_{60} of 0.2 s, and evaluated its performance on both the test data from Dataset1 and Dataset3, which has a longer reverberation time of $RT_{60} = 0.6\,\mathrm{s}$. In the mixed-source scenario, we used a model trained on Dataset2 and compared its performance on both the test data from Dataset2 and Dataset4.

The proposed model was trained using the same preprocessing and parameters as in [5], while the comparison models were trained following the preprocessing methods and parameters specified in their respective papers. MUSIC was implemented using the open-source software HARK [18], with all parameters set to HARK's recommended default values. Training was conducted for a maximum of 50 epochs, with

TABLE I: Effectiveness of each proposed technique (training:dataset1/test:dataset1)

vM-B	Stride	Binary	-10			-5			0			20		
Tubelet	Division	CE	MAE↓	Acc↑	Acc±5↑	MAE↓	Acc↑	Acc±5↑	MAE↓	Acc↑	Acc±5↑	MAE↓	Acc↑	Acc±5↑
		✓	19.68	67.88	77.76	14.04	78.63	84.49	9.763	85.44	89.21	1.811	96.59	97.63
	✓	✓	9.275	76.79	89.04	3.7	89.28	95.79	1.44	95.41	98.49	0.131	99.79	99.87
✓	✓	✓	**5.173**	**87.5**	**94.05**	**1.826**	**95.42**	**97.94**	**0.778**	**98.33**	**99.12**	**0.08**	**99.89**	**99.93**

TABLE II: BCE loss vs. CCE loss (Acc)

Training Data	Test Dataset	#Sources	BCE				CCE			
			-10	-5	0	20	-10	-5	0	20
dataset1	dataset1	1	87.50	95.42	**98.33**	99.89	**87.75**	**95.43**	98.29	**99.94**
dataset2	dataset2	1 & 2	88.64	95.54	**98.01**	99.36	**89.43**	**95.80**	97.93	**99.38**
	part of dataset2	1	**86.51**	**96.60**	**97.29**	**99.36**	71.37	80.91	85.33	95.06
	part of dataset2	2	88.7	95.51	98.03	99.36	**89.95**	**96.23**	**98.30**	**99.50**

TABLE III: Performance Comparison

Method	#Sources	-10			-5			0			20		
		MAE↓	Acc↑	Acc±5↑	MAE↓	Acc↑	Acc±5↑	MAE↓	Acc↑	Acc±5↑	MAE↓	Acc↑	Acc±5↑
Proposed	1	5.17	**87.50**	94.05	1.83	**95.42**	97.94	**0.78**	**98.33**	99.12	**0.08**	**99.89**	99.93
vM-B ResNet	1	38.13	35.39	55.30	29.07	46.48	66.60	22.59	54.21	74.34	11.08	77.60	87.79
SELDNet	1	**3.14**	67.75	**95.98**	**1.49**	83.25	**99.61**	1.41	86.64	**99.41**	1.15	93.24	99.52
MUSIC	1	68.41	15.08	32.63	48.31	27.16	53.64	28.38	29.31	74.11	3.41	58.84	98.80
Proposed	1 & 2	**4.71**	**88.69**	**94.45**	**1.74**	**95.51**	**98.03**	**0.89**	**98.03**	**99.05**	**0.38**	**99.37**	**99.60**
MUSIC	1 & 2	63.78	11.45	26.07	49.61	19.22	41.13	36.31	27.80	56.68	16.59	41.63	81.85

TABLE IV: Single Source Localization for different reverberation conditions (RT_{60}=0.2s and RT_{60}=0.6s)

RT60	Method	-10			-5			0			20		
		MAE↓	Acc↑	Acc±5↑	MAE↓	Acc↑	Acc±5↑	MAE↓	Acc↑	Acc±5↑	MAE↓	Acc↑	Acc±5↑
0.2s	Proposed	**5.17**	**87.50**	**94.05**	**1.83**	**95.42**	**97.94**	**0.78**	**98.33**	**99.12**	**0.08**	**99.89**	**99.93**
	MUSIC	68.41	15.08	32.63	48.31	27.16	53.64	28.38	39.21	74.11	3.409	58.84	98.80
0.6s	Proposed	**2.71**	**83.90**	**96.93**	**0.99**	**89.94**	**99.29**	**0.50**	**92.15**	**99.85**	**0.29**	**94.41**	**99.99**
	MUSIC	69.21	14.21	32.19	48.49	27.11	53.82	26.95	41.65	75.55	1.82	66.69	99.86

TABLE V: Multiple Source Localization for different reverberation conditions (RT_{60}=0.2s and RT_{60}=0.6s)

RT60	Method	-10			-5			0			20		
		MAE↓	Acc↑	Acc±5↑	MAE↓	Acc↑	Acc±5↑	MAE↓↑	Acc↑	Acc±5↑	MAE↓	Acc↑	Acc±5↑
0.2s	Proposed	**4.71**	**88.69**	**94.45**	**1.74**	**95.51**	**98.03**	**0.89**	**98.03**	**99.05**	**0.38**	**99.37**	**99.60**
	MUSIC	63.78	11.45	26.07	49.61	19.22	41.13	36.31	27.80	56.68	16.59	41.63	81.85
0.6s	Proposed	**5.38**	**79.95**	**94.35**	**2.62**	**86.45**	**97.66**	**2.01**	**89.06**	**98.44**	**1.42**	**91.53**	**99.01**
	MUSIC	63.42	11.27	31.16	49.4	18.9	47.58	37.41	25.92	61.95	23.34	40.09	81.25

early stopping applied if the validation loss did not decrease for 5 consecutive epochs.

D. Results and discussion

Tables I and II summarize the findings from Experiment 1. Table I highlights the impact of each enhancement to ViViT under the single-source scenario, while Table II contrasts BCE and CCE losses using models trained on Dataset1 for single-source evaluation and on Dataset2 for both single- and two-source settings.

From Table I, it is clear that combining stride division with BCE loss leads to improved results, and the addition of vM-B tubelet embedding further enhances all performance metrics (MAE, ACC, and Acc±5), resulting in the strongest overall performance. These outcomes indicate that earlier methods were unable to fully exploit the periodicity inherent in phase inputs, whereas the vM-B tubelet embedding enables a more effective use of this periodicity, significantly improving localization accuracy.

Table II shows the evaluation of BCE loss in multi-source scenarios. When trained on Dataset1 and tested on single-source data, all methods performed similarly; although designed for multiple sources, BCE achieved comparable performance to CCE. In contrast, CCE appears better on Dataset2, which mainly consists of two-source data and thus favors models specialized for that case. Splitting the Dataset2 test set reveals this trend: BCE outperforms on single-source data due to CCE's fixed two-source assumption, which leads to source count errors. On two-source data, CCE performs slightly better since its fixed assumption aligns with the data. In summary, CCE performs well when the number of sources matches between training and testing, but lacks flexibility. BCE, formulated as a multi-label classification task, generalizes better across source counts. Given this, CCE loss achieves high performance when the number of sources in the training and evaluation data matches but fails to maintain good performance when they do not. On the other hand, BCE loss, formulated as a multi-label classification problem, enables the model to generalize even when the number of sources in the training and evaluation datasets differs.

Table III shows the results of Experiment 2. First, when evaluating the model trained on Dataset1 with single-source data, the proposed method and SELDNet demonstrated good performance. Comparing these two methods, the proposed method achieved the highest accuracy (Acc), while SELDNet performed slightly better in the other two metrics. This difference can be attributed to the fundamental approach of treating SSL as either a classification or a regression task. Since the proposed method formulates localization as a classification problem, it excels at accurately estimating the sound source direction, leading to superior Acc performance. On the other hand, SELDNet formulates localization as a regression problem, which introduces some estimation errors but prevents the predicted results from deviating significantly from the correct direction. As a result, it achieves better performance in MAE and Acc±5. Therefore, improving MAE and Acc±5 can be achieved by making the classification problem more similar to a regression problem, specifically by increasing the angular resolution of the proposed method. When evaluating the model trained on Dataset2, the proposed method outperformed the MUSIC method in all metrics. In particular, when the SNR is below 0, MUSIC suffers from theoretical limitations, leading to a significant performance drop. However, the proposed method maintains good performance even in scenarios where single-source and two-source cases are mixed.

Tables IV and V show the result of Experiment 3. Table IV presents the results for the single-source scenario, while Table V shows the results for the mixed single- and two-source scenario. Across all SNR conditions, the proposed method consistently outperforms MUSIC. In general, MUSIC is known to be vulnerable to reverberation as it does not handle multipath effects well. In this experiment, however, MUSIC demonstrated stable results despite changes in reverberation, regardless of the number of sound sources. Nevertheless, its absolute performance remains low under SNR conditions of 0 dB or lower due to its theoretical limitations. On the other hand, the proposed method maintains high performance under both noise and reverberation. It achieves an Acc of over 90%, and in almost all cases, MAE remains below 5 degrees, which is the angular resolution, demonstrating its high accuracy.

From the results of these experiments, it was demonstrated that the proposed method outperforms conventional methods in terms of both performance and robustness under noise and reverberation, even in the presence of multiple sound sources.

V. CONCLUSION

This paper presented improvements to deep learning-based SSL and extensions to support multiple sound sources for real-world applications such as robot audition. Specifically, we extended the ViViT-based model and introduced binary cross entropy loss to reformulate the problem from multi-class classification to multi-label classification, enabling localization of multiple sound sources. Additionally, we incorporated von Mises-Bernoulli (vM-B) tubelet embedding to account for phase periodicity. Experimental results demonstrated that the proposed method achieves high accuracy and robustness under various acoustic conditions, including low SNR, reverberant. and multi-source scenarios, validating its effectiveness. Future work includes extending the model for online inference using streaming audio, and integrating the localization model into realtime robot audition systems.

REFERENCES

[1] Hiroshi G. Okuno and Kazuhiro Nakadai, "Robot audition: Its rise and perspectives," in *ICASSP*, 2015, pp. 5610–5614.

[2] David F. Rosenthal and Hiroshi G. Okuno, Eds., *Computational auditory scene analysis*, Lawrence Erlbaum Associates Publishers, 1998.

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

[4] A. Arnab et al., "Vivit: A video vision transformer," *CoRR*, vol. abs/2103.15691, 2021.

[5] H. Yokota et al., "A video vision transformer for sound source localization," in *Eropean Signal Processing Conference (EUSIPCO)*, 2024.

[6] K. Nakadai et al., "Sound source localization based on von-Mises-Bernoulli deep neural network," in *SII*, 2020, pp. 658–663.

[7] K. Itoyama et al., "Assessment of von Mises-Bernoulli deep neural network in sound source localization," in *Interspeech*, 2021, pp. 2152–2156.

[8] S. Adavanne et al., "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 34–48, 2019.

[9] C. Schymura et al., "Exploiting attention-based sequence-to-sequence architectures for sound event localization," in *EUSIPCO*, 2021, pp. 231–235.

[10] S. Park et al., "Many-to-many audio spectrogram transformer: Transformer for sound event localization and detection.," in *DCASE*, 2021, pp. 105–109.

[11] P.-A. Grumiaux et al., "Improved feature extraction for crnn-based multiple sound source localization," in *Proc. of the 29th European Signal Processing Conference (EUSIPCO)*, 2021.

[12] S. Kapka and M. Lewandowski, "Sound source detection, localization and classification using consecutive ensemble of crnn models," Tech. Rep., DCASE 2019 Challenge, 2019.

[13] N. Yalta et al., "Sound source localization using deep learning models," *J. Robotics Mechatronics*, vol. 29, pp. 37–48, 2017.

[14] M. Bozkurtlar et al., "Real time sound source localization using von-Mises ResNet," in *SII*, 2024, pp. 466–471.

[15] L. Perotin et al., "Crnn-based multiple doa estimation using acoustic intensity features for ambisonics recordings," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 22–33, 2019.

[16] J. Yamagishi et al., "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," 2019.

[17] E. Vincent et al., "The CHiME-4 challenge: Advances in robust speech recognition for multiple speaker environments," in *Proceedings of the 2016 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2016, pp. 335–340.

[18] K. Nakadai et al., "Development, deployment and applications of robot audition open source software HARK," *Journal of Robotics and Mechatronics*, vol. 29, no. 1, pp. 16–25, 2017.