DOA Estimation based on Learnable TDOA Feature

Inkyu An[1] and Sung-eui Yoon[2]

***Abstract*— We propose the novel learnable time difference of arrival model to estimate direction-of-arrivals (DOAs). Different from existing TDOA features, our learnable TDOA model can consider semantic information of sound and distinguish various sound events. Our TDOA model is based on the self-attention mechanism. The self-attention mechanism is effective in learning semantic information of sound by analyzing the spatial relation of input audio.**

Our TDOA model are designed to learn TDOA information of sound events effectively. We compared our approach to the prior work, i.e., the speech-oriented sound source localization task [1]. We observed a significant improvement, i.e., a 32.7 % error reduction in the mean absolute error and a 2.1 % improvement in accuracy, compared to the prior work.

I. INTRODUCTION

Sound source localization (SSL) is a fundamental problem for robot auditions. There have been many efforts based on signal-processing-based techniques to deal with the SSL problems. Although meaningful progress exists thanks to signal-processing-based approaches, many issues remain in SSL.

Deep learning (DL)-based methods have recently been presented, and they give us significant improvements. He *et al.* [1] proposed the deep neural networks for multiple speaker localization. By adding noises, e.g., fan noises of the robot, to the training dataset, their method can be robust against the noises. Adavanne *et al.* [2] proposed the source localization method for multiple sound events, e.g., alarm, speech, and footstep. Their method can detect and localize multiple sound events simultaneously. These DL-based SSL methods were trained by multiple-channel audio datasets [1], [3].

It is popular for prior DL-based methods to utilize audio features based on signal processing techniques. To localize sound source positions, the time difference of arrival (TDOA) features, e.g., generalized cross correlation-phase transform (GCC-PHAT) [4], are widely used. The combination of TDOA features given a microphone array corresponds to the specific direction-of-arrival (DOA); thus, prior DL methods estimates DOAs by considering the combination of TDOA features. However, existing TDOA features can only encode the time difference between two coherent signals, but cannot consider semantic information of sound events.

This research was supported by the MSIT, Korea, under the ITRC support program(IITP-2022-2020-0-01460) supervised by the IITP

[1]Inkyu An is with School of Computing, Korea Advanced Institude of Science and Technology, Daejeon, South Korea, inkyu.an@kaist.ac.kr

[2]Sung-eui Yoon is with Faculty of School of Computing, Korea Advanced Institude of Science and Technology, Deajeon, South Korea sungeui@kaist.edu

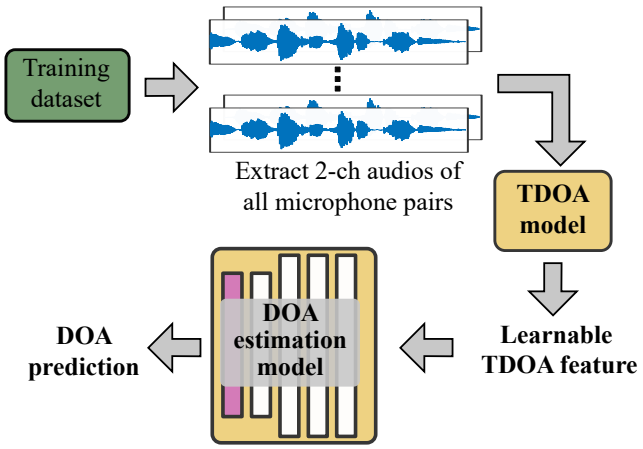

Fig. 1. The overview of our approach. 2-ch audios of microphone pairs in the dataset become an input of our approach. Out TDOA model learn to estimate TDOAs considering semantic information of sound. Our TDOA model computes multiple learnable TDOA features. Our DOA estimation model is designed to estimate DOAs from multiple learnable TDOA features. Our DOA estimation model learns the relation between the combination of multiple learnable TDOA features and the corresponding DOA. The white and purple boxes indicate the linear layer and the Sigmoid function, respectively.

We propose the novel learnable TDOA feature. Our TDOA feature, computed by our TDOA model, can learn semantic information of sound as well as time differences between two coherent audio signals. Our model can be trained by the training dataset and show better performance compared to the prior work [1]. Our model can distinguish various sound events by considering semantic information.

II. THE DOA ESTIMATION USING LEARNABLE TDOA FEATURE

The overview of our approach is shown in Fig. 1. We first extract every 2-ch audio of all microphone pairs from the training dataset. Extracted 2-ch audios enters to our TDOA model; our TDOA model can estimate TDOAs of two coherent sound signals considering semantic information. After our TDOA model computes the learnable TDOA features, we can estimate DOAs from the learnable TDOA features.

Learnable TDOA feature. Our TDOA model is effective to estimate TDOAs by consider semantic information. Our TDOA model is designed based on the self-attention mechanism [5]. The self-attention mechanism is a powerful approach to analyze spatial relations of input [6], [7]. For example, given an image containing a person, the self-attention mechanism can learn the spatial relation between each part of the person, e.g., head, body, and legs.

TABLE I

THE ACCURACY OF DOA ESTIMATIONS OF OURS AND THE PRIOR WORK.

	MAE (↓)	ACC (↑)
MLP-GCC [1]	4.61 degree	91.91 %
Ours	3.1 degree	93.9 %

The self-attention mechanism is also useful to learn semantic information of sound. The audio input contains consecutive sound signals capturing various sound events, e.g., speech and footstep. Each sound event has different spatial relations; for example, speech is a sequence of voices of a person, but footstep is a sequence of the sound of a person's foot tapping the floor. Our approach utilize the self-attention mechanism to learn spatial relations in the time domain and, thus, can distinguish different sound events.

The input of the TDOA model is 2-ch audios of all microphone pairs; thus, there should be multiple learnable TDOA features given all microphone pairs. The multiple learnable TDOA features go to the next step: estimating DOAs.

The DOA estimation model. The combination of TDOAs of all pairs given the microphone array corresponds to the specific DOAs; thus, we can estimate DOAs from the combination of learnable TDOA features. Our DOA estimation model is designed to learn those relations between the combination of TDOA features and the corresponding DOA.

Our DOA estimation model consists of four linear layers, and the Sigmoid function computes the DOA predictions. We utilize the binary cross-entropy loss between the DOA predictions and DOA labels to train our models.

III. RESULT AND DISCUSSION

We tested our approach in the speech-oriented SSL, i.e., estimating DOAs only of speech sounds. We compared our approach to the prior works [1] utilizing the existing TDOA feature, e.g., GCC-PHAT [4]. By comparing to prior works, we want to show the effectiveness of our learnable TDOA feature. We utilize the SSLR dataset [1] recorded from various conversations with additional noises. The SSLR dataset is recorded by the 4-ch circular microphone array; thus, there exist six microphone pairs in four microphones.

We utilize the evaluation metric proposed by [1], consisting of MAE and ACC; they are the mean absolute error and the accuracy of correct predictions, respectively.

We verify that our approach shows better accuracy for both metrics than the prior work in Table. I. We observe that our approach gives us a significant improvement, i.e., a 32.7 % error reduction in MAE and a 2.1 % improvement in ACC, compared to the prior work.

Those results show that our learnable TDOA feature is useful for localizing sound sources. Our learnable TDOA feature can be more helpful than existing TDOA features, e.g., GCC-PHAT. Moreover, we verify that our learnable TDOA feature efficiently considers semantic information. Considering semantic information is important to localize various speech in SSLR while ignoring additional noises.

REFERENCES

[1] Weipeng He, Petr Motlicek, and Jean-Marc Odobez, "Deep neural networks for multiple speaker detection and localization", in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 74–79.

[2] Sharath Adavanne, Archontis Politis, Joonas Nikunen, and Tuomas Virtanen, "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, 2018.

[3] Archontis Politis, Sharath Adavanne, Daniel Krause, Antoine Deleforge, Prerak Srivastava, and Tuomas Virtanen, "A dataset of dynamic reverberant sound scenes with directional interferers for sound event localization and detection", *arXiv preprint arXiv:2106.06999*, 2021.

[4] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay", *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need", *Advances in neural information processing systems*, vol. 30, 2017.

[6] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid, "Vivit: A video vision transformer", in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6836–6846.

[7] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang, "Transformer in transformer", *arXiv preprint arXiv:2103.00112*, 2021.