PLCST: PROBABILISTIC LOCALIZATION AND CLASSIFICATION OF SOUNDS WITH TRANSFORMERS FOR SOUND EVENT LOCALIZATION AND DETECTION

Technical Report

Peipei Wu[1], *Jinzheng Zhao*[1], *Yaru Chen*[1], *Berghi Davide*[1], *Chenfei Zhu*[3], *Yin Cao*[2], *Yang Liu*[4], *Philip Jackson*[1], *Wenwu Wang*[1]

[1] University of Surrey, Centre for Vision, Speech and Signal Processing (CVSSP), Surrey, UK, {p.wu, j.zhao, yaru.chen, d.berghi,p.jackson, w.wang}@surrey.ac.uk
[2] Xi'an Jiaotong-Liverpool University, Department of Intelligent Science, Suzhou, China, {yin.k.cao}@gmail.com
[3] Daqian Information, Wuhan, China, {chenfei.zhu1994}@gmail.com
[4] Meta, Seattle, USA, {yangliuai}@meta.com

ABSTRACT

Sound Event Localization and Detection (SELD) is a task that involves detecting different types of sound events along with their temporal and spatial information, specifically, class-level events detection and their corresponding direction of arrivals at each frame. In DCASE 2023 Task 3, the recordings consist of real-world sound scenes with complex conditions, which contain simultaneous occurrences of up to 3 or even 5 events. Our submitted system for this task is based on the previously proposed method, PILOT (Probabilistic Localization of Sounds with Transformers). While PILOT combines transformers with CNN-based feature extraction modules and covers Sound Event Localization (SEL) tasks with sound activity detection, it requires modifications to address SELD tasks. In our architecture, we adapt PILOT's input features and output branches to SELD tasks and revise the loss function accordingly. We name our model Probabilistic Localization and Class of Sounds with Transformers (PLCST). Unlike other approaches, we do not generate additional samples from the development dataset or use other datasets for training, aiming to mitigate discrepancies. In addition, another benefit of our model is that the number of parameters is relatively small. Our experimental results demonstrate improvements in our system over the baseline methods.

Index Terms— Sound event localization and detection, transformer

1. INTRODUCTION

The purpose of the sound event localization and detection (SELD) task is to detect various types of activating sound events and localize them in the temporal and spatial domains. SELD technology can be used in numerous areas, such as improving speech quality for automatic speech recognition (ASR) and audio surveillance system for smart cities. Because of this task's prospect, SELD draws attention from both industry and academics.

Since this challenge started in the year of 2019, some relevant requirements of this task have been changed. For example, in the first 3 years of the challenges (the year 2019 to 2021), emulated multichannel recordings are provided. They spatialized event sample banks with spatial room impulse responses (SRIRs) in different rooms and mixed with spatial ambient noise recorded at the same locations to generate those recordings. The last year, in challenge 2022, the providing dataset changed to recordings of real sound scenes with manual annotations. Similar to previous iterations, this year, DCASE provides a dataset recorded from real scenes, which is named as Sony-TAu Realistic Spatial Soundscapes 2023 (STARSS23). Compared to the STARSS22 dataset used in DCASE2022, it adds an additional 4 hours of material captured in Tampere University distributed between the training and evaluation sets while all STARSS22 are maintained [1]. Different from datasets, the baseline for this task does not change too much. In the series, the baseline system selects a straightforward convolutional recurrent neural network (CRNN) from SELDNet with few modifications [2]. First, multi-head self-attention blocks are introduced [3]. In addition, Multi-ACCDOA is used to support detecting multiple instances of the same class when they are overlapping [4]. Finally, SALSA-lite [5] features are utilized for the microphone version of the dataset. This is to overcome the poor performance of GCC features in the presence of multiple overlapping sound events.

In this report, we use the part of the Probabilistic Localization of Sounds with Transformers (PILOT) as our network's architecture for this task [6]. Considering that the original PILOT can only detect sound events' activity rather than their types, we modified the input features and output branch to suit the SELD task. The whole network has four components, including the feature extraction module, the transformer encoder, the linear Gaussian system, and the classification network. Due to time constraints, we didn't generate data and used the development dataset for training only, and the complex features were not applied to the network. However, we will extend our work for the workshop and solve these problems.

The rest of this report is organized as follows. In Section 2, the proposed method and its training process are described. Section 3 shows the result of the proposed method on the development dataset. The last section concludes the experiments and provides some information linking to our extended work for the workshop.

2. PROPOSED METHOD

In this section, we provide an overview of the proposed network and its training processing. More details will be given in the workshop's paper.

2.1. Network Architecture

As mentioned in previous sections, the whole network has 4 different components, including the feature extraction module, the transformer encoder, the linear Gaussian system, and the classification layers.

The first component is the feature extraction module, designed to extract features from the input features map and generate embeddings for the rest network. Considering that both SELDNet and PILOT perform well in SELD and Sound event localization (SEL) tasks, we derive ideas from those networks. This module is constructed as three-layer CNNs followed by a fully-connection network. Each layer contains 64 kernels with the size of 3×3. Also, each conv layer is followed by batch normalization, rectified linear units (ReLU), and max-pooling. Similarly, the fully-connection network has three layers, with intermediate and final output dimensions 128. Finally, the output from the fully-connection network will be applied with two different linear layers to generate two embeddings, including positional embeddings and observation noise, for both the transformer encoder and the linear Gaussian system.

Next, the second component is the transformer encoder. We didn't amend the structure of the original transformer encoder. The parameters setting of the transformer encoder is listed as follows. The encoder has three layers, and each encoder layer contains 8 multi-heads. The dimensions of the input embedding and feedforward are 32 and 1024, respectively.

The output of the encoder will be applied by the linear Gaussian system and a classification network. The details of the linear gaussian system can be found in the original paper of PILOT [6]. The classification network is a three-layer fully-connection network with intermediate dimensions 256 and 128, respectively.

The overview of the proposed network is shown in Figure 1.

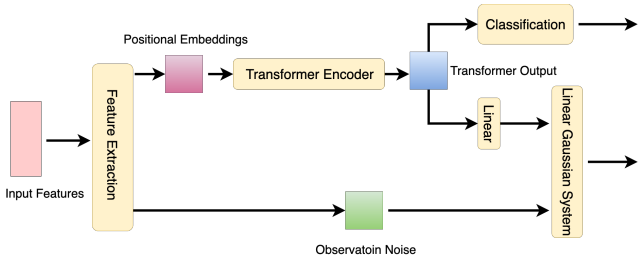

Figure 1: The overview of the proposed method's structure.

2.2. Network Training

Due to the time limitation, we only adopt the basic features processed on FOA types recordings in the dataset for the proposed networks. Therefore, we use the scripts offered by baseline to apply STFT with the default parameters settings on the recordings to obtain spectrograms for the phase map, which will be used for SEL. Considering that we need to classify different types of each sound event, we also extract mel spectrograms by provided scripts. Then, regarding the mel spectrogram and phase map having the same channel number 4, we concat them together before sending them into the proposed network. Due to the dimensions of the phase map and the mel spectrogram being 513×4 and 64×4, respectively, the dimension of the feature embeddings is 577×4.

The training data are cut into chunks with a fixed length of 1 ms, which is the same as the annotation. Adam optimizer is used to train the model, and a tri-stage rate scheduler is used to optimize the learning rate during training. To reach the deadline of the submission, the learning rate is set bigger than usual, starting at 0.05. The model is trained on the development dataset only with the specific splits ratio defined by the challenge. And the batch size is set as 128 for accelerating the training. The result of the current model will be influenced by these facts. However, we will improve this when we complete the paper for the workshop.

Table 1: Example of a figure with experimental results.

Methods	ER_{20}	F_{20}	LE_{CD}	LR_{CD}
mACCDOA baseline (p)	0.57	48.7	22	47.7
mACCDOA baseline (w)	0.74	15.4	97.3	25.1
PLCST	0.95	1.0	133.4	1.1

3. RESULTS ON THE DEVELOPMENT DATASET

We evaluate our proposed method on the development dataset of STARSS2023. Table 3 shows the experimental results of the proposed method for the development test dataset. We list out the experiment's result of the proposed method and baselines with different input features. Noted that we put both the published and we trained results of the baseline system in this table, which are distinguished by (p) and (w). The mACCDOA in the table means multi-ACCDOA features.

4. CONCLUSION, LIMITATION, AND FUTURE WORK

This report proposes a network named PLCST to solve the SELD task in the DCASE2023 challenge. However, we can see that our proposed method's current result is not competitive enough with the most distinct methods in this area. This is because we didn't provide a robust model by training with enough data and suitable parameter settings due to time being quite limited. Therefore, in the next step, we will generate more data for training, amend some network architecture, adjust the loss function, and find the best match parameters setting. Finally, the improved works will be submitted for the workshop.

5. REFERENCES

[1] A. Politis, K. Shimada, P. Sudarsanam, S. Adavanne, D. Krause, Y. Koyama, N. Takahashi, S. Takahashi, Y. Mitsufuji, and T. Virtanen, "STARSS22: A dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events," in *Proceedings of the 8th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022)*, Nancy, France, November 2022, pp. 125–129. [Online]. Available: https://dcase.community/workshop2022/proceedings

[2] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *CoRR*, vol. abs/1807.00129, 2018. [Online]. Available: http://arxiv.org/abs/1807.00129

[3] P. Sudarsanam, A. Politis, and K. Drossos, "Assessment of self-attention on learned features for sound event localization and detection," *CoRR*, vol. abs/2107.09388, 2021. [Online]. Available: https://arxiv.org/abs/2107.09388

[4] K. Shimada, Y. Koyama, S. Takahashi, N. Takahashi, E. Tsunoo, and Y. Mitsufuji, "Multi-accdoa: Localizing and detecting overlapping sounds from the same class with auxiliary duplicating permutation invariant training," 2022.

[5] T. N. T. Nguyen, D. L. Jones, K. N. Watcharasupat, H. Phan, and W.-S. Gan, "SALSA-lite: A fast and effective feature for polyphonic sound event localization and detection with microphone arrays," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, may 2022. [Online]. Available: https://doi.org/10.1109%2Ficassp43922.2022.9746132

[6] C. Schymura, B. T. Bönninghoff, T. Ochiai, M. Delcroix, K. Kinoshita, T. Nakatani, S. Araki, and D. Kolossa, "PILOT: introducing transformers for probabilistic sound event localization," *CoRR*, vol. abs/2106.03903, 2021. [Online]. Available: https://arxiv.org/abs/2106.03903