THE LU SYSTEM FOR DCASE 2024 SOUND EVENT LOCALIZATION AND DETECTION CHALLENGE

Technical Report

Axel Berg[1,2]*, Johanna Engman*[1]*, Jens Gulin*[1,3]*, Karl Åström*[1]*, Magnus Oskarsson*[1]*,*

[1]Computer Vision and Machine Learning, Centre for Mathematical Sciences, Lund University, Sweden {firstname.lastname@math.lth.se}
[2] Arm, Lund, Sweden {axel.berg@arm.com}
[3] Sony Europe B.V., Lund, Sweden {jens.gulin@sony.com}

ABSTRACT

This technical report gives an overview of our submission to task 3 of the DCASE 2024 challenge. We present a sound event localization and detection (SELD) system using input features based on trainable neural generalized cross-correlations with phase transform (NGCC-PHAT). With these features together with spectrograms as input to a Transformer-based network, we achieve significant improvements over the baseline method. In addition, we also present an audio-visual version of our system, where distance predictions are updated using depth maps from the panorama video frames.

Index Terms— sound event localization and detection, time difference of arrival, generalized cross-correlation

1. INTRODUCTION

The sound event localization and detection (SELD) task consists of classifying different types of acoustic events, while simultaneously localizing them in 3D space. In previous editions of the challenge, the localization amounted to predicting the direction of arrival (DOA), whereas this year's challenge also involves estimating the distance relative to the microphone array. The audio recordings can be used in two formats: first order ambisonics (FOA), which combines recordings from 32 microphones, or 4-channel recordings from a tetrahedral microphone array (MIC). In recent years, most systems submitted to the challenge have utilized the former format, whereas the latter has been less explored. In this report, we therefore focus on how to better exploit information in the MIC recordings.

Generalized cross-correlations with phase transform (GCC-PHAT) [1] combined with spectral audio features is the basis for most SELD methods for microphone arrays. The spectral features contain important cues on what type of sound event is active, whereas the purpose of GCC-PHAT is to extract the time-differences of arrival (TDOA) between pairs of microphones. The TDOA measurements can then be mapped to DOAs given the geometry of the array. However, GCC-PHAT is known to be sensitive to noise and reverberation [2]. GCC-PHAT may also fail to resolve TDOAs for multiple events simultaneously, because of limited sampling resolution. To this end, spatial cue-augmented log-spectrogram (SALSA) [3] and variants thereof (SALSA-Lite) [4] have been proposed, which combine directional cues with spectrograms in a single feature.

In this report, we explore the neural GCC-PHAT (NGCC-PHAT) [5], that filters audio signals and outputs multiple correlation features per microphone pair. We show that such features can be learnt by employing permutation invariant training, which allows for prediction of TDOAs for multiple overlapping sound events.

This work was partially supported by the strategic research project ELLIIT and the Wallenberg AI, Autonomous Systems and Software Program (WASP), funded by the Knut and Alice Wallenberg Foundation. Model training was enabled by the Berzelius resource provided by the Knut and Alice Wallenberg Foundation at the National Supercomputer Centre in Sweden.

2. TDOA FEATURE EXTRACTION

For TDOA audio features, we use NGCC-PHAT [5], which is a form of generalized cross-correlation that can be trained to predict time delays between pairs of microphones. Given two signals $\mathbf{x}_i, \mathbf{x}_j$ received in a time frame from microphones i and j, NGCC-PHAT infers the corresponding TDOA for a sound event.

We extend NGCC-PHAT to predict time delays for multiple events in a single time frame using auxiliary duplicating permutation invariant training (ADPIT) [6] by creating separate target labels for each active sound event. For a given time frame, let $\mathbf{r}_i, \mathbf{r}_j \in \mathbb{R}^3$ denote the locations of microphone i and j, and $\mathbf{s}_k \in \mathbb{R}^3$ be the location of the k:th active event. We then define the corresponding TDOA as

$$\tau_{ij}^k = \lfloor \frac{F_s}{c} \left(||\mathbf{s}_k - \mathbf{r}_i||_2 - ||\mathbf{s}_k - \mathbf{r}_j||_2 \right) \rceil, \tag{1}$$

where F_s is the sampling rate, c is the speed of sound and $\lfloor \cdot \rceil$ denotes rounding to the nearest integer.

We then train a classifier to predict the TDOA of all active events for all pairs of microphones by treating it as a multinomial classification problem, where predictions from K separate output tracks are assigned to the different events. The last layer of the NGCC-PHAT network therefore outputs K probability distributions $p_k(t|\mathbf{x}_i, \mathbf{x}_j)$ over the set of integer delays $t \in \{-\tau_{\text{max}}, ..., \tau_{\text{max}}\}$, where $\tau_{\text{max}} = \max_{i,j} \lfloor ||\mathbf{r}_i - \mathbf{r}_j||_2 F_s / c \rceil$ is the largest possible TDOA for any pair of microphones. The corresponding label is a unit impulse response defined as

$$\delta_{\tau_{ij}^k}[t] = \begin{cases} 1, & t = \tau_{ij}^k, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

With K as the number of tracks, assume for now that there are also K active events. Furthermore, let $\text{Perm}([K])$ denote the set of

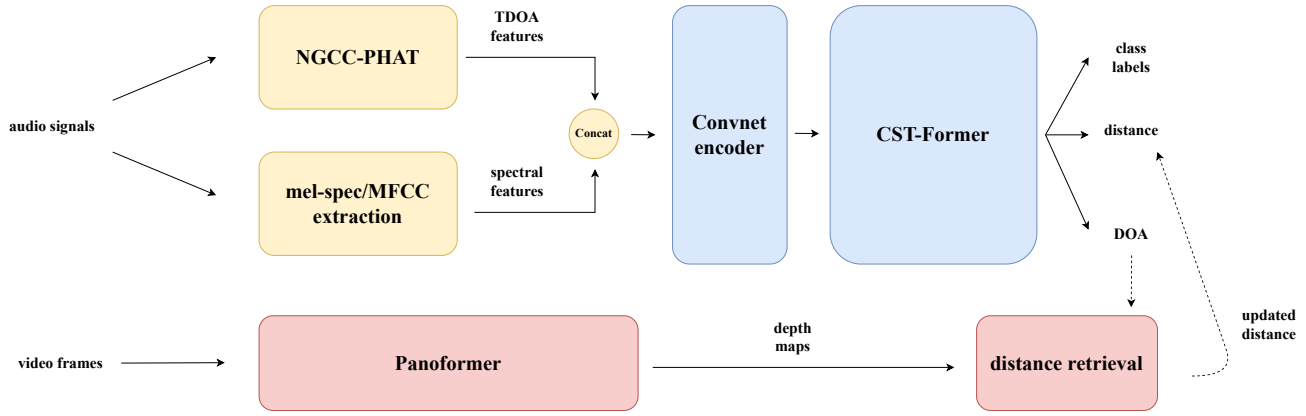

Figure 1: Overview of our SELD system. Audio features from a pre-trained NGCC-PHAT network are fed together with spectral features to a CST-former network that outputs SELD predictions. In the audio-visual system, we update the distance predictions from frame-wise depth-maps extracted from a pre-trained Panoformer-network.

permutations of the events $\{1, \dots, K\}$. For a single microphone pair (i, j) and an event arrangement $\alpha \in \text{Perm}([K])$, the loss is calculated using the average cross-entropy over all output tracks as

$$l_\alpha(\mathbf{x}_i, \mathbf{x}_j) = -\frac{1}{K} \sum_{k=1}^{K} \sum_{t=-\tau_{\max}}^{\tau_{\max}} \delta_{\tau_{ij}^{\alpha(k)}}[t] \log p_k(t|\mathbf{x}_i, \mathbf{x}_j). \quad (3)$$

Due to the ambiguity in assigning different output tracks to different events, we calculate the loss for all possible permutations of the events and use the minimum. The loss is then averaged over all $M(M-1)/2$ microphone pairs, where M is the total number of microphones, and the total loss then becomes

$$L = \frac{2}{M(M-1)} \sum_{\substack{i,j=1 \\ i<j}}^{M} \min_{\alpha \in \text{Perm}([K])} l_\alpha(\mathbf{x}_i, \mathbf{x}_j). \quad (4)$$

Note that this loss function is class-agnostic, since the output tracks are not assigned class-wise. The main purpose of the TDOA features are therefore to provide better features for localization in combination with some other audio representation that is suitable for event detection and classification. Time frames with no active events are discarded during NGCC training, since no TDOA label can be assigned. When at least one, but fewer than K events are active, we duplicate the label for the last active event, such that predictions for each track can be assigned to a TDOA label.

3. SYSTEM DESIGN

3.1. Audio System

Our audio system consists of an NGCC-PHAT network together with CST-former [7], which is a Transformer-based SELD network that utilizes self-attention across the temporal, spectral and channel dimensions independently. We consider only the tetrahedral microphone array recordings and do not use first order ambisonics in our system. Training of our system consists of two phases: 1) pre-training of the NGCC-PHAT network for TDOA prediction as described in Section 2 and 2) training the CST-former for the SELD task. A high-level overview of our system is shown in Figure 1.

The NGCC-PHAT network operates on raw audio signals and consists of four convolutional layers with 32 output channels, the first being a SincNet [8] layer, and the remaining three use filter of length 11, 9, and 7 respectively. Here, each convolutional layer is applied independently to audio from the $M = 4$ different microphones. GCC-PHAT features are then computed channel-wise for all pair-wise combinations of microphones, and the features are then processed by another four convolutional layers, where the final layer has $C = 16$ output channels. The tetrahedral microphone array used for the recordings has a diameter of 8.4 cm, which corresponds to a maximum TDOA of $\tau_{\max} = 6$ delays at a sampling rate of $F_s = 24$ kHz. In total, the TDOA features therefore has shape, $[C, M(M-1)/2, 2(\tau_{\max}+1)] = [16, 6, 13]$.

During pre-training for TDOA-prediction, the 16 channels are then mapped by a final convolutional layers to $K = 3$ output tracks. Although the maximum polyphony in a single time frame in this year's challenge is five, we use $K = 3$ tracks since the computational complexity of PIT-training scales as $\mathcal{O}(K!)$ and more than three simultaneous events are rare anyway. When more than three events are active, we randomly select labels for three events and discard the rest.

When training the SELD-network, we extract TDOA features for longer audio signals by windowing the NGCC-PHAT computation without overlap. We use the default challenge setup of 5 second audio inputs, which corresponds to $T = 250$ TDOA features when using a window length of 20 ms. Since the TDOA features are designed to be class-agnostic, we combine them with spectral features in order to better distinguish between different types of event. For this we use log mel-spectograms (MS) or mel-frequency cepstral coeficcients (MFCC) with $F = 64$ spectral features for each of the $M = 4$ input channels.

When merging the spectral features with the TDOA features, we first concatenate the 16 channels for the 6 microphone pairs of the TDOA features, and use a linear projection to map the 13 time-delays to 64 dimensions. The TDOA features are then concatenated with the M spectral features channel-wise, resulting in a combined feature size of $[CM(M-1)/2 + M, T, F] = [100, 250, 64]$.

The combined feature is passed through a small convolutional network with 64 output channels with pooling over the time and spectral dimensions. Here we use two different variants that deter-

mine the size of the input features to the CST-former network: 1) pooling over 5 time windows and 4 frequencies, which produces features of size [64, 50, 16], or 2) pooling over 5 time windows and no pooling over frequencies, which results in features of size [64, 50, 64]. We call the second variant *large* for that reason.

The CST-former network consists of Transformer blocks, where each block contains three self-attention modules: temporal attention, spectral attention and channel attention with unfolded local embedding. We use the default configuration with two blocks, each with eight attention heads, and refer to [7] for more details about this architecture.

3.2. Audio-Visual System

For better depth predictions, we have tested the use of a pre-trained model for monocular depth estimation. We use the model Panoformer [9], that is trained on panoramic indoor images (the Stanford 2D-3D Semantic dataset [10]). We sample every third frame in the video sequence. The image data is then downsampled to a resolution of 512×1024, and normalized with mean $(0.485, 0.456, 0.406)$ and standard deviation $(0.229, 0.224, 0.225)$ in RGB respectively. The normalized images are passed to the Panoformer model which outputs 512×1024 metric depth maps. The depth maps are scaled with a factor of 25 and truncated to unsigned 8-bit for storage purposes. When we run our audio model we can get a depth estimate for each detection by looking up the corresponding depth value in the depth map position corresponding to our predicted angle.

We update the depth maps for all classes, except for "water tap" and "knock", since these events are often occluded and not visible to the camera. Furthermore, for human-related classes (speech and laughter), we subtract 30 degrees from the predicted elevation angle when retrieving the depth, in order to avoid retrieving depth from blurred faces.

4. EXPERIMENTS

4.1. Dataset and Training Setup

We train all our models on a mixture of real spatial audio recordings and simulated recordings. The real recordings are from the STARSS23 [11] audio-only development dataset, which consists of about 7 hours of multi-channel audio recordings. For data augmentation, we use channel-swapping [12], which expands the dataset by a factor of 8 by swapping the input channels and corresponding DOA labels in different combinations.

The simulated data is provided as a part of the DCASE 2024 challenge [13] and consists of 20 hours of synthesized recordings, where the audio clips are taken from the FSD50K [14] dataset. In addition, we generate an additional 2 hours of synthesized recordings using Spatial Scaper [15] with impulse responses from the TAU [16] and METU [17] databases. This additional data only contains sounds from classes that occur rarely in the real recordings, namely "bell", "clapping", "doorCupboard", "footsteps", "knock" and "telephone". The total amount of training data amounts to about 78 hours.

The NGCC-PHAT network was trained for 1 epoch with a constant learning rate of 0.001, after which the weights were frozen. For the challenge submission, we trained the rest of the system for 175 epochs using the AdamW optimizer [18] with a batch size of 64, a cosine learning rate schedule starting at 0.001 and weight decay of 0.05. The mean squared error was used as loss function with labels in the Multi-ACCDOA [6] format, with distances included as proposed in [19]. In order to penalize errors in predicted distance relative to the proximity of the sound events, we scale loss-terms for the distance error with the reciprocal of the ground truth distance.

For reporting results on the STARSS23 development dataset, we train the models on the same data as described above, but omitting the dev-test split and use the non-augment version of this split for evaluation instead. The number of training epochs is also increased from 175 to 300[1].

Table 1: Macro-averaged test results on STARSS23 [11] audio-only dev-test. Our results are obtained using CST-former [7]. Large models discards pooling over frequencies in the input features.

Model	Input feature	$F_{LD} \uparrow$	$DOAE \downarrow$	$RDE \downarrow$	#params
Baseline [20]	GCC + MS	9.9	38.1	0.30	744k
CST-Former	GCC + MS	16.3	28.7	0.79	550k
	SALSA-Lite	26.4	28.2	0.38	530k
	NGCC + MS	29.0	23.9	0.38	663k
	NGCC + MFCC	28.7	20.8	0.38	663k
CST-Former (large)	NGCC + MS	32.0	21.8	0.44	1.49M
	NGCC + MFCC	26.8	26.5	0.57	1.49M

Table 2: Macro-averaged test results on STARSS23 [11] audio-visual dev-test using different types of input features. Our results are obtained using CST-former [7]. Updated distance estimates are performed using Panoformer [9]. Large models discards pooling over frequencies in the input features.

Model	Input feature	$F_{LD} \uparrow$	$DOAE \downarrow$	$RDE \downarrow$	#params
Baseline [20]	GCC + MS	11.8	38.5	0.29	2.7M
CST-Former + Panoformer	GCC + MS	21.3	28.7	0.32	20.9M
	SALSA-Lite	27.0	28.2	0.28	20.9M
	NGCC + MS	29.8	23.9	0.28	21.0M
	NGCC + MFCC	29.4	20.8	0.28	21.0M
CST-F. (large) + Panoformer	NGCC + MS	33.4	21.8	0.28	21.9M
	NGCC + MFCC	29.0	26.5	0.28	21.9M

4.2. Experimental Results on Development Dataset

We report our results on the STARSS23 dev-test split for the audio-only system in Table 1, where we compare our method with the challenge baseline using the metrics defined in the challenge. The metrics used are the macro-averaged location-dependent F-score (F_{LD}), which is thresholded such that true positives must have an angular error less than 20° and relative distance errors must be less than 1, as well as the macro-averaged DOA error (DOAE) and relative distance error (RDE).

We also report results of training our system with standard GCC-PHAT and SALSA-lite input features. Notably, NGCC features yields large improvements in DOAE compared to these.

In Table 2, we present results where the distance predictions are updated using depth maps. This results in significantly lower relative distance errors compared to the audio-only system, as well as small improvements in terms of F-score.

[1]Code will be made available at `https://github.com/axeber01/ngcc-seld/`

5. REFERENCES

[1] C. Knapp and G. Carter, "The Generalized Correlation Method for Estimation of Time Delay," *IEEE transactions on acoustics, speech, and signal processing*, vol. 24, no. 4, pp. 320–327, 1976.

[2] B. Champagne, S. Bédard, and A. Stéphenne, "Performance of time-delay estimation in the presence of room reverberation," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 2, pp. 148–152, 1996.

[3] T. N. T. Nguyen, K. N. Watcharasupat, N. K. Nguyen, D. L. Jones, and W.-S. Gan, "Salsa: Spatial cue-augmented log-spectrogram features for polyphonic sound event localization and detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1749–1762, 2022.

[4] 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, 2022, pp. 716–720.

[5] A. Berg, M. O'Connor, K. Åström, and M. Oskarsson, "Extending GCC-PHAT using Shift Equivariant Neural Networks," in *Proc. Interspeech 2022*, 2022, pp. 1791–1795.

[6] 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," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 316–320.

[7] Y. Shul and J.-W. Choi, "Cst-former: Transformer with channel-spectro-temporal attention for sound event localization and detection," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 8686–8690.

[8] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," in *2018 IEEE spoken language technology workshop (SLT)*. IEEE, 2018, pp. 1021–1028.

[9] Z. Shen, C. Lin, K. Liao, L. Nie, Z. Zheng, and Y. Zhao, "Panoformer: Panorama transformer for indoor 360 depth estimation," in *European Conference on Computer Vision*. Springer, 2022, pp. 195–211.

[10] I. Armeni, A. Sax, A. R. Zamir, and S. Savarese, "Joint 2D-3D-Semantic Data for Indoor Scene Understanding," *ArXiv e-prints*, Feb. 2017.

[11] K. Shimada, A. Politis, P. Sudarsanam, D. A. Krause, K. Uchida, S. Adavanne, A. Hakala, Y. Koyama, N. Takahashi, S. Takahashi, T. Virtanen, and Y. Mitsufuji, "STARSS23: An audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events," in *Advances in Neural Information Processing Systems*, vol. 36. Curran Associates, Inc., 2023, pp. 72 931–72 957.

[12] Q. Wang, J. Du, H.-X. Wu, J. Pan, F. Ma, and C.-H. Lee, "A four-stage data augmentation approach to resnet-conformer based acoustic modeling for sound event localization and detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1251–1264, 2023.

[13] D. A. Krause and A. Politis, "[DCASE2024 Task 3] Synthetic SELD mixtures for baseline training," Apr. 2024. [Online]. Available: https://doi.org/10.5281/zenodo.10932241

[14] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50K: An open dataset of human-labeled sound events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2021.

[15] I. R. Roman, C. Ick, S. Ding, A. S. Roman, B. McFee, and J. P. Bello, "Spatial Scaper: a library to simulate and augment soundscapes for sound event localization and detection in realistic rooms," *arXiv preprint arXiv:2401.12238*, 2024.

[16] A. Politis, S. Adavanne, and T. Virtanen, "TAU Spatial Room Impulse Response Database (TAU-SRIR DB)," Apr. 2022. [Online]. Available: https://doi.org/10.5281/zenodo.6408611

[17] O. Olgun and H. Hacihabiboglu, "METU SPARG Eigenmike em32 Acoustic Impulse Response Dataset v0.1.0," Apr. 2019. [Online]. Available: https://doi.org/10.5281/zenodo.2635758

[18] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[19] D. A. Krause, A. Politis, and A. Mesaros, "Sound event detection and localization with distance estimation," *arXiv preprint arXiv:2403.11827*, 2024.

[20] "DCASE 2024 Task 3 baseline code repository," https://github.com/partha2409/DCASE2024_seld_baseline, [Accessed 12-06-2024].