LEARNING MULTI-TARGET TDOA FEATURES
FOR SOUND EVENT LOCALIZATION AND DETECTION

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

Sound event localization and detection (SELD) systems using audio recordings from a microphone array rely on spatial cues for determining the location of sound events. As a consequence, the localization performance of such systems is to a large extent determined by the quality of the audio features that are used as inputs to the system. We propose a new feature, based on neural generalized cross-correlations with phase-transform (NGCC-PHAT), that learns audio representations suitable for localization. Using permutation invariant training for the time-difference of arrival (TDOA) estimation problem enables NGCC-PHAT to learn TDOA features for multiple overlapping sound events. These features can be used as a drop-in replacement for GCC-PHAT inputs to a SELD-network. We test our method on the STARSS23 dataset and demonstrate improved localization performance compared to using standard GCC-PHAT or SALSA-Lite input features.

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. The DCASE SELD Challenge [1] provides first order ambisonics (FOA) recordings and signals captured from a 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 work, we therefore focus on how to better exploit information in the MIC recordings by learning to extract better features.

Generalized cross-correlations with phase transform (GCC-PHAT) [2] 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) for pairs of microphones. The TDOA measurements can then be mapped to direction-of-arrival (DOA) estimates, given the geometry of the array. However, GCC-PHAT
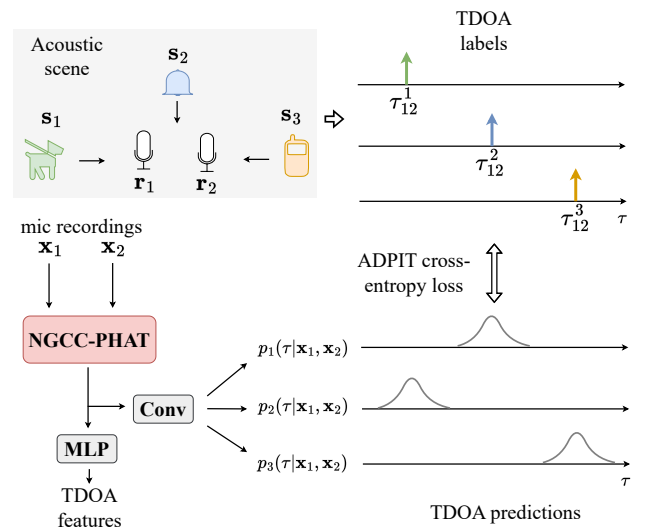
Figure 1: Overview of our pre-training strategy with $K = 3$ tracks. Given a set of sound events, we train a neural GCC-PHAT to predict the TDOA of each event. When the number of sound events is less than K, auxiliary duplication of the labels is used. In this illustration, only two microphones are shown for brevity.

is known to be sensitive to noise and reverberation [3]. GCC-PHAT can also fail to separate TDOAs for overlapping events, since two events at different locations can have the same TDOA for a given microphone pair, which yields only one correlation peak.

To improve separation of overlapping events, Xu et al. [4] proposed a beamforming approach, where phase differences from the cross-power spectrum are used as input features. Similarly, Cheng et al. [5] showed that localization performance can be improved by first filtering the audio signals using a sound source separation network before performing feature extraction. Several works [6, 7] have also proposed end-to-end localization from raw audio signals. The most widely adopted input feature is however the spatial cue-augmented log-spectrogram (SALSA) [8] and variants thereof (SALSA-Lite) [9], that combine directional cues with spectral cues in a single feature. This is done by calculating the principal eigenvector of the spatial covariance matrix for the different frequencies in the spectrogram.

Although some recent works [10, 11, 12] have approached TDOA estimation using learning-based methods, there is a lack

of research in how to combine this with the SELD task. Berg et al. [12] proposed using a shift-equivariant neural GCC-PHAT (NGCC-PHAT) network. However, this method, as it was originally proposed, only supports single-source TDOA estimation and was not evaluated in a real-world localization scenario.

In this work, we describe how NGCC-PHAT can be trained to extract TDOA features for multiple sound sources. We show that such features can be learnt by employing permutation invariant training, which allows for prediction of TDOAs for multiple overlapping sound events. Furthermore, we show that these features can be used with an existing SELD-pipeline on a real-world dataset, for better performance compared to using traditional input features. The material presented in this work is an extension of our DCASE 2024 challenge submission [13].

2. METHOD

2.1. Background

Consider an acoustic scene, as shown in Figure 1, with M microphones located at positions $\mathbf{r}_m \in \mathbb{R}^3$ for $m = 1, \dots, M$. Furthermore, let $\mathbf{s}_p \in \mathbb{R}^3$, $p = 1, \dots, P$ denote the locations of the active sound events. For a given time frame, each microphone records a signal x_i, which is composed of the sum of active events as

$$x_i[n] = \sum_{p=1}^{P} (h_{p,i} * u_p)[n] + w_i[n], \quad n = 1, \dots, N, \quad (1)$$

where u_p is the p:th active event, $h_{p,i}$ is the room impulse response from the p:th event to the i:th microphone, w_i is additive noise and N is the number of samples. Furthermore, we define the TDOA for microphone pair (i, j) and the p:th event as

$$\tau_{ij}^p = \lfloor \frac{F_s}{c} \left(||\mathbf{s}_p - \mathbf{r}_i||_2 - ||\mathbf{s}_p - \mathbf{r}_j||_2 \right) \rceil, \quad (2)$$

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

The GCC-PHAT is defined as

$$R_{ij}[\tau] = \frac{1}{N} \sum_{k=0}^{N-1} \frac{X_i[k] X_j^*[k]}{|X_i[k] X_j^*[k]|} e^{\frac{i 2\pi k \tau}{N}}, \quad (3)$$

where (X_i, X_j) are the discrete Fourier transforms of (x_i, x_j). The feature is calculated for time delays $\tau = -\tau_{\max}, \dots, \tau_{\max}$, where $\tau_{\max} = \max_{i,j} \lfloor ||\mathbf{r}_i - \mathbf{r}_j||_2 F_s / c \rceil$ is the largest possible TDOA for any pair of microphones. In an anechoic and noise-free environment with a single sound event u_p, this results in $R_{ij}[\tau] = \delta_{\tau_{ij}^p}[\tau]$, where

$$\delta_{\tau_{ij}^p}[\tau] = \begin{cases} 1, & \tau = \tau_{ij}^p, \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

In practice, GCC-PHAT will often yield incorrect TDOA estimates due to noise and reverberation. In the case of multiple overlapping sound events, the different events may interfere and result in difficulties resolving peaks in their signal correlations.

NGCC-PHAT attempts to alleviate this problem by filtering the input signals using a learnable filter bank with L convolutional filters, before computing GCC-PHAT features R_{ij}^l, $l = 1, \dots, L$ for each channel in the signals independently. In theory, such a filter bank can perform source separation so that different channels

in the NGCC-PHAT correspond to TDOAs for different sound events. Note that for an ideal filter bank that perfectly separates the p:th sound event to the l:th channel, we would have $R_{ij}^l[\tau] = \delta_{\tau_{ij}^p}[\tau]$ in an anechoic and noise-free environment, due to the shift-equivariance of the convolutional filters.

2.2. Permutation Invariant Training for TDOA Estimation

We extend NGCC-PHAT to predict time delays for multiple events in a single time frame using auxiliary duplicating permutation invariant training (ADPIT) [14], by creating separate target labels for each active sound event. This is done by training a classifier network to predict the TDOA of all active events for all pairs of microphones by treating it as a multinomial classification problem. The L correlation features are first processed using another series of convolutional layers with C output channels. These are then projected to K different output tracks which are assigned to the different events. The last layer of the NGCC-PHAT network therefore outputs probability distributions $p_k(\tau|\mathbf{x}_i, \mathbf{x}_j)$ for $k = 1, \dots, K$ over the set of integer delays $\tau \in \{-\tau_{\max}, \dots, \tau_{\max}\}$, as illustrated in Figure 1.

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

$$l_\alpha(\mathbf{x}_i, \mathbf{x}_j) = -\frac{1}{K} \sum_{k=1}^{K} \sum_{\tau=-\tau_{\max}}^{\tau_{\max}} \delta_{\tau_{ij}^{\alpha(k)}}[\tau] \log p_k(\tau|\mathbf{x}_i, \mathbf{x}_j). \quad (5)$$

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, giving the total loss

$$\mathcal{L} = \frac{2}{M(M-1)} \sum_{\substack{i,j=1 \\ i<j}}^{M} \min_{\alpha \in \text{Perm}([K])} l_\alpha(\mathbf{x}_i, \mathbf{x}_j). \quad (6)$$

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 when combined with spectral features that are suitable for classification.

When the assumption $P = K$ does not hold, the formal implication is that α needs to cover another set of event arrangements. Our approach is equivalently to transform each such case into subcases where the assumption holds. Time frames with no active events ($P = 0$) are discarded in the loss calculation, since no TDOA label can be assigned. When $1 \le P \le K - 1$, we perform auxiliary duplication of events following the method in [14], which makes the loss invariant to both permutations and which events that are duplicated. Furthermore, in the case of $K < P$, it is possible to compute the loss for all subsets of K events from P and use the minimum.

3. EXPERIMENTAL SETUP

3.1. Using TDOA Features for SELD

In order to show the benefits of better TDOA features for SELD, we demonstrate how they can be used in conjunction with a SELD-system. This involves two training phases: 1) pre-training of the

NGCC-PHAT network for TDOA prediction and 2) training the SELD-network using the TDOA features as input. The NGCC-PHAT network operates on raw audio signals and consists of four convolutional layers, the first being a SincNet [15] layer, and the remaining three use filters of length 11, 9, and 7 respectively. Here, each convolutional layer has $L = 32$ channels and together form the filter bank mentioned in Section 2.1, which is applied independently to audio from the different microphones. GCC-PHAT features are computed channel-wise for each microphone pair, and the features are then processed by another four convolutional layers, where the final layer has $C = 16$ output channels.

The maximum delay used is chosen for compatibility with the setup in the STARSS23 dataset [16], which uses a tetrahedral array with $M = 4$ microphones. The diameter of the array is 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 have 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 convolutional layer to $K = 3$ output tracks. Although the maximum polyphony in a single time frame in the dataset is five, we use $K = 3$ tracks since the computational complexity of permutation invariant training scales as $\mathcal{O}(K!)$ and more than three simultaneous events are rare. When more than three events are active, for pre-training we randomly select labels for three events and discard the rest.

When training the SELD-network, we extract the TDOA input features for longer audio signals by windowing the NGCC-PHAT computation without overlap. We use an input duration 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 for the same time-frame in order to better distinguish between different types of event. For this we use log mel-spectrograms (MS) with $F = 64$ spectral features for each recording.

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 multi-layer perceptron to map the 13 time-delays to 64 dimensions. The TDOA features are then reshaped and concatenated with the M spectral features channel-wise, as shown in Figure 2, 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 pooling variants that determine the size of the input features to the SELD-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 resulting network variants *Small* and *Large* for this reason.

For SELD-training, we use a CST-Former [17] network that 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 [17] for more details about this architecture.

3.2. Dataset and Model Training

We train all our models on a mixture of real spatial audio recordings and simulated recordings. The real recordings are from the STARSS23 [16] audio-only dev-train dataset, which consists of
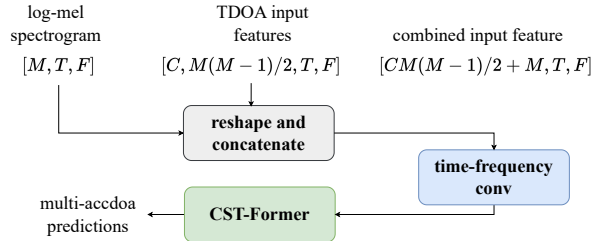

Figure 2: Illustration of how TDOA features are used together with log mel-spectrograms as input to the CST-Former network.

about 3.5 hours of multi-channel audio recordings. The dataset has up to 5 simultaneous events from 13 different classes. For data augmentation, we use channel-swapping [18], 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 [19] and consists of 20 hours of synthesized recordings, where the audio is taken from the FSD50K [20] dataset. In addition, we generate an additional 2 hours of synthesized recordings using Spatial Scaper [21] with impulse responses from the TAU [22] and METU [23] databases. This additional data contains sounds from classes that occur rarely in STARSS23, namely "bell", "clapping", "doorCupboard", "footsteps", "knock" and "telephone". The total amount of training data is about 50 hours.

The NGCC-PHAT network was trained for one epoch with a constant learning rate of 0.001, after which the weights were frozen. The CST-Former network was then trained for 300 epochs using the AdamW optimizer [24] 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 [14] format, with distances included as proposed in [25]. 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.

Evaluations were done using the DCASE 2024 SELD challenge metrics [1, 26]. This includes the location dependent F-score F_{LD}, the DOA error $DOAE$ and the relative distance error RDE, which is the distance error divided by the ground truth distance to the event. Each metric is calculated class-wise and then macro-averaged across all classes. Furthermore, the location dependent F-score only counts predicted events as true positives if they are correctly classified and localized, such that predictions with $DOAE$ larger than $T_{DOA} = 20°$ or RDE larger than $T_{RD} = 1$ are counted as false positives. We focus on evaluating the performance of our method compared to that of other commonly used input features with the same SELD-network, and do not compare to other (e.g. FOA-based) state-of-the-art methods.

4. RESULTS

Our main results are presented in Table 1, where we compare our method to GCC with MS and to SALSA-Lite. Our method performs better in terms of F_{LD} and $DOAE$, for both the Small and Large variant of the network, although SALSA-Lite has the lowest RDE for the Large variant. When increasing the model size, the results improve for both SALSA-Lite and NGCC, but not for GCC. Since GCC features are less informative, the increase in model size results
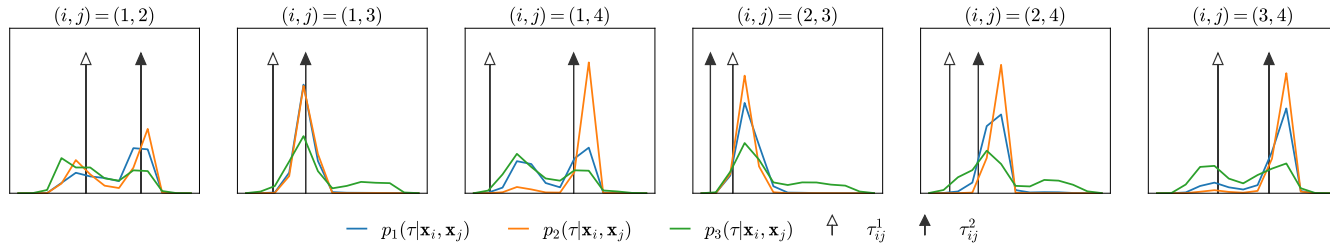
Figure 3: An example of the TDOA predictions $p_k(\tau|\mathbf{x}_i, \mathbf{x}_j)$ from the pre-trained NGCC-PHAT network using $K = 3$ output tracks. Predictions are shown for all six microphone combinations (i, j) at a single time frame with two events and ground truth TDOAs τ_{ij}^1 and τ_{ij}^2.

Table 1: Macro-averaged test results on STARSS23 [16] dev-test.

Input feature	$F_{LD} \uparrow$	$DOAE \downarrow$	$RDE \downarrow$	#params
CST-Former Small				
GCC + MS	15.7 ± 1.0	27.7 ± 2.1	0.78 ± 0.02	550K
SALSA-Lite	24.6 ± 2.0	27.0 ± 1.2	$\mathbf{0.41 \pm 0.02}$	530K
NGCC + MS	$\mathbf{26.0 \pm 2.0}$	$\mathbf{25.8 \pm 2.3}$	0.42 ± 0.01	663K
CST-Former Large				
GCC + MS	14.2 ± 1.1	28.4 ± 1.9	0.84 ± 0.03	1.37M
SALSA-Lite	26.1 ± 1.0	26.4 ± 3.6	$\mathbf{0.42 \pm 0.02}$	1.35M
NGCC + MS	$\mathbf{28.2 \pm 2.8}$	$\mathbf{23.2 \pm 1.8}$	0.50 ± 0.02	1.49M

Table 2: Ablations of the number input channels used in the TDOA input features for CST-Former Small.

C	$F_{LD} \uparrow$	$DOAE \downarrow$	$RDE \downarrow$	#params
1	24.4 ± 2.3	29.7 ± 3.3	0.44 ± 0.08	608K
4	24.2 ± 0.8	$\mathbf{23.2 \pm 2.5}$	0.46 ± 0.01	619K
16	$\mathbf{26.0 \pm 2.0}$	25.8 ± 2.3	$\mathbf{0.42 \pm 0.01}$	663K

in overfitting. The same can be said for the increase in RDE when using NGCC + MS, since the TDOA features from both GCC and NGCC mostly contain angular cues, but less information about spatial distance. Note that GCC + MS and NGCC + MS use exactly the same CST-Former architecture, so the extra parameter count when using NGCC comes from the pre-trained feature extractor. When using SALSA-Lite, the pooling operations in the convolutional layers were adjusted in order to achieve a similar model size.

In order to verify the importance of using more than one input channel for TDOA features, we ablate the number of channels C in the NGCC-PHAT network. The results are shown in Table 2, where it can be seen that increasing the number of channels from 1 to 16 increases performance in terms of all metrics. This agrees with the intuition that using more than one input channels enables the pre-training to better separate spatial cues from different events. Furthermore, the cost for increasing the number channels in terms of the increase in model parameters is relatively small.

We also ablate the number of tracks K used for TDOA-prediction during pre-training, and present the location dependent F-score for values of T_{DOA} in Figure 4. Due to the sensitivity of the macro-averaged F-score to incorrect predictions for rare classes in the test data, we instead use the micro-averaged statistic. At the default $20°$ threshold, the effect of increasing the number of tracks is small, but asymptotically it is clear that using $K = 3$ tracks increases the F-score regardless of how many events are active. Note that the number of tracks only affects the complexity in the pre-training stage of NGCC-PHAT, and not the overall parameter
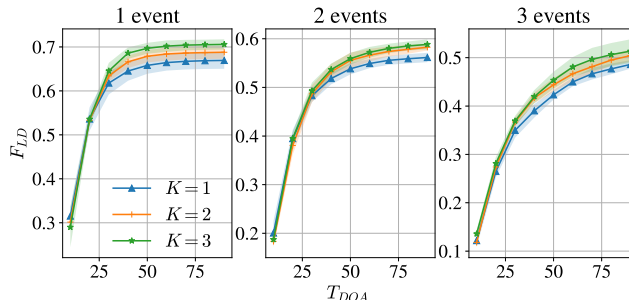

Figure 4: Micro-averaged F-score as a function of the angular threshold T_{DOA} using different number of output tracks K during TDOA pre-training. Evaluation was done using CST-Former Small.

count of the final model, since all C channels are used as input to the network, and the mapping to K tracks can be discarded.

Finally, we show examples of TDOA predictions in Figure 3. When the TDOAs of the events are well-separated, the different tracks yield different peaks at approximately the correct time delays. However, for the microphone pairs where events are tightly spaced, the predictions fail to separate the different TDOAs.

5. CONCLUSIONS

In this work we proposed an input feature based on NGCC-PHAT and showed its usefulness as input to a SELD-network. Permutation invariant training for the TDOA estimation problem enabled NGCC-PHAT to learn TDOA features for multiple overlapping sound events, and improved SELD performance compared to using GCC-PHAT or SALSA-Lite input features.

These results indicate that our NGCC-PHAT pre-training for TDOA classification provides a good feature extractor for the SELD task. Intuitively, better TDOA prediction in the feature extractor ought to yield better SELD results, but further studies are needed to validate this. Evaluating TDOA prediction performance would however involve new methodology, such as heuristics for peak selection from the output tracks, as well as selecting useful evaluation metrics. The downstream network could be resilient to some type of information our current loss function aims to suppress. In addition, a source-wise or class-wise TDOA format could be beneficial. We therefore anticipate future work to explore other pre-training options and end-to-end training.

Focusing on the feature extractor, we made minimal effort to address the other challenges of the dataset. We leave for future work to incorporate known techniques, such as class balancing, additional data augmentation, temporal filtering and ensemble voting.

6. REFERENCES

[1] "Audio and Audiovisual Sound Event Localization and Detection with Source Distance Estimation," https://dcase.community/challenge2024/task-audio-and-audiovisual-sound-event-localization-and-detection-with-source-distance-estimation, 2024, [Accessed 2024-07-03].

[2] 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.

[3] 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.

[4] 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 *Proc. Interspeech 2020*, 2020, pp. 5091–5095.

[5] S. Cheng, J. Du, Q. Wang, Y. Jiang, Z. Nian, S. Niu, C.-H. Lee, Y. Gao, and W. Zhang, "Improving sound event localization and detection with class-dependent sound separation for real-world scenarios," in *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2023, pp. 2068–2073.

[6] H. Sundar, W. Wang, M. Sun, and C. Wang, "Raw waveform based end-to-end deep convolutional network for spatial localization of multiple acoustic sources," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 4642–4646.

[7] Y. He and A. Markham, "SoundDoA: Learn Sound Source Direction of Arrival and Semantics from Sound Raw Waveforms," in *Proc. Interspeech 2022*, 2022, pp. 2408–2412.

[8] 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.

[9] 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.

[10] X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 2814–2818.

[11] D. Salvati, C. Drioli, and G. L. Foresti, "Time Delay Estimation for Speaker Localization Using CNN-Based Parametrized GCC-PHAT Features," in *Interspeech*, 2021, pp. 1479–1483.

[12] 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.

[13] A. Berg, J. Engman, J. Gulin, K. Åström, and M. Oskarsson, "The LU System for DCASE 2024 Sound Event Localization and Detection Challenge," DCASE2024 Challenge, Tech. Rep., June 2024.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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

[20] 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.

[21] 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.

[22] 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

[23] 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

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

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

[26] A. Politis, A. Mesaros, S. Adavanne, T. Heittola, and T. Virtanen, "Overview and evaluation of sound event localization and detection in dcase 2019," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 684–698, 2020.