Reverberation-based Features for Sound Event Localization and Detection with Distance Estimation

Davide Berghi, *Member, IEEE* and Philip J. B. Jackson, *Member, IEEE*

***Abstract*—Sound event localization and detection (SELD) involves predicting active sound event classes over time while estimating their positions. The localization subtask in SELD is usually treated as a direction of arrival estimation problem, ignoring source distance. Only recently, SELD was extended to 3D by incorporating distance estimation, enabling the prediction of sound event positions in 3D space (3D SELD). However, existing methods lack input features designed for distance estimation. We argue that reverberation encodes valuable information for this task. This paper introduces two novel feature formats for 3D SELD based on reverberation: one using direct-to-reverberant ratio (DRR) and another leveraging signal autocorrelation to provide the model with insights into early reflections. Pre-training on synthetic data improves relative distance error (RDE) and overall SELD score, with autocorrelation-based features reducing RDE by over 3 percentage points on the STARSS23 dataset. The code to extract the features is available at `github.com/dberghi/SELD-distance-features`**

***Index Terms*—Distance Estimation, Sound Event Localization and Detection, Sound Source Localization, Reverberation.**

I. Introduction

SOUND event localization and detection (SELD) [1] integrates two subtasks: sound event detection (SED) and sound source localization (SSL). Thus, it involves identifying active sound events at any given time frame while estimating their spatial positions. SELD systems are important in many practical applications, e.g., human-robot interaction, security, augmented reality, accessibility, safety, and immersive production. SELD gained significant attention following its inclusion as a task in the Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge. Recent advancements in SELD research have tackled increasingly complex challenges, such as detecting moving sound events [2], ignoring external interfering sounds [3], distinguishing simultaneous same-class events originating from different directions of arrival (DOAs) [4], [5], [6], and leveraging the visual modality to tackle SELD as a multimodal task [7], [8], [9], [10].

The localization aspect of SELD is traditionally framed as a direction of arrival estimation (DOAE) problem, predicting the azimuth and elevation of sound events. However, this overlooks source distance, a crucial factor in many applications. The DCASE 2024 challenge addressed this by introducing distance estimation (3D SELD) [11]. Krause *et al*. [12] proposed two methods to support distance estimation in 3D SELD. The first method extends the multi-activity-coupled Cartesian DOA (multi-ACCDOA) vectors [6] to include distance estimation. Multi-ACCDOA vectors are commonly used in SELD as they simultaneously predict the DOAs of sound events and encode their activity level in the vector length (0 for inactive and 1 for active events). In the extended version presented in [12], the model predicts, for each event class *c*, track *n*, and time frame *t*, a 3-element DOA vector – the (*x, y, z*) coordinates on the unit sphere – along with a distance value $D_{nct} \in \langle 0, \infty)$. This representation, referred to as the multi-activity-coupled Cartesian Distance and DOA (multi-ACCDDOA) method, incorporates distance estimation into the original framework. The second method presented in [12] includes a separate output branch specifically for distance estimation. Alternatively, Hong *et al*. [13] proposed an approach where the model directly predicts the 3D positions of sound events in the form of (*x, y, z*) coordinates. This approach combines DOA and distance into a single representation, offering a unified localization framework. However, it requires an additional output branch to handle the SED subtask.

Selecting the right input features is crucial in designing a SELD system [14]. Commonly adopted features include log-mel spectrograms for the SED subtask, intensity vectors (IV) [4] for DOAE in first-order ambisonics (FOA) audio format, and generalized cross-correlation with phase transform (GCC-PHAT) [15] or SALSA-Lite [16] for microphone array (MIC) format. However, to the best of our knowledge, features specifically designed for distance estimation have not yet been explored within the context of 3D SELD. Useful information about the sound events distance is encoded in the acoustic reverberation [17], [18], [19], [20], [21], [22]. This paper proposes and evaluates two reverberation-based input feature extraction methods. The first uses the direct-to-reverberant ratio (DRR) as an indicator of the energy balance between direct and reverberant sound. The second leverages the autocorrelation function to extract information about early reflections and to estimate the initial time delay gap (ITDG), the interval between direct sound arrival and the first major reflection. Experiments on the STARSS23 dataset [7] demonstrate that incorporating these features alongside log-mel spectrograms and intensity vectors (IV) enhances distance estimation accuracy and overall SELD performance.

The main contributions of this paper are threefold: (1) we propose two methods for extracting reverberation-based input features to address distance estimation in 3D SELD; (2)

Research supported by EPSRC and BBC Prosperity Partnership AI4ME: Future Personalised Object-Based Media Experiences Delivered at Scale Anywhere EP/V038087. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising. Data supporting this study are available from https://zenodo.org/records/7880637 and from https://zenodo.org/records/10932241.

D. Berghi and P. J. B. Jackson are with the Centre for Vision, Speech, and Signal Processing, University of Surrey, Guildford, U.K. (e-mail: {davide.berghi, p.jackson}@surrey.ac.uk).

we conduct a preliminary study showing that autocorrelation-based features capture distance-related information by analyzing how an audio clip interacts with room impulse responses (RIRs) recorded at different distances; (3) we validate the effectiveness of the proposed features on real data.

II. Proposed Reverberation-based Features

Log-mel spectrograms and intensity vectors (IVs) work well for SED and DOAE but are not suited for distance estimation. To address this, we introduce two input features specifically designed to enhance distance estimation.

A. Direct-to-Reverberant Features

Distance cues can be extracted from the relationship between the direct and reverberant components of the captured audio signals [17]. To estimate the direct sound, $d(t)$, we employed the Weighted Prediction Error (WPE) dereverberation algorithm [23] applied to the omnidirectional channel W of the first-order ambisonic (FOA) audio format. Specifically, we adopted the Python implementation of the WPE algorithm released by Drude *et al*. [24] (taps=60; delay=5; iterations=5). The reverberant component, $r(t)$, is then estimated by subtracting the direct signal from the original signal in the temporal domain. To extract DRR features as 2D inputs to the model and to enable concatenation with the other SELD features (i.e., log-mel spectrograms and IVs), we calculate the DRR as a function of time and frequency, and then mapped it to log-mel space. The proposed DRR input features, $\mathbf{DRR}^{\mathrm{mel}}$, are defined as:

$$\mathbf{DRR}^{\mathrm{mel}}(t,k) = 10 \cdot \log_{10}\left(\mathbf{P}^{\mathrm{mel}}_{\mathrm{DRR}}(t,k)\right) \tag{1}$$

$$\mathbf{P}^{\mathrm{mel}}_{\mathrm{DRR}}(t,k) = \sum_{f=0}^{F} \mathbf{H}^{\mathrm{mel}}(k,f)\left(\frac{\mathbf{P}_{\mathrm{D}}(t,f)}{\mathbf{P}_{\mathrm{R}}(t,f)}\right) \tag{2}$$

where $\mathbf{H}^{\mathrm{mel}}$ denotes the mel filter bank, which maps the frequency spectrum to the mel scale, with k being the mel bin index. $\mathbf{P}_{\mathrm{D}}(t,f)$ and $\mathbf{P}_{\mathrm{R}}(t,f)$ are the power spectral densities (PSDs) of the direct and reverberant components, respectively. To prevent instability and avoid division by zero, the PSD values were clamped to a small positive constant, $\epsilon{=}1e{-}10$. Mathematically, $\mathbf{P}_{\mathrm{D}}(f,t)$ and $\mathbf{P}_{\mathrm{R}}(f,t)$ can be defined as $\mathbf{P}_{\mathrm{D}}(t,f){=}\max(|\mathbf{D}(t,f)|^2,\epsilon)$ and $\mathbf{P}_{\mathrm{R}}(t,f){=}\max(|\mathbf{R}(t,f)|^2,\epsilon)$, where $\mathbf{D}(t,f)$ and $\mathbf{R}(t,f)$ are the short-term Fourier transforms (STFTs) of the direct and reverberant components, $d(t)$ and $r(t)$, respectively.

In addition to the DRR features described, we explore a variant where $\mathbf{D}(t,f)$ and $\mathbf{R}(t,f)$ are separately converted into log-mel spectrograms and fed into the model. This approach, introduced in our DCASE2024 Task 3 submission [25], aims to give the network greater flexibility in learning task-relevant information. We refer to these features as D+R features.

B. Short-term Power of the Autocorrelation

For the second feature, we explore the role of early reflections in distance estimation, focusing on the ITDG, a key cue for perceiving distance [26], [27]. While early reflection delays also depend on room size, it is reasonable to assume that the earliest reflection originates from the floor [27], as shown in Fig. 1. From this assumption, Table I demonstrates that ITDG from floor reflections decreases as the source-microphone distance increases. These values assume a microphone height of $h_m{=}1.5$m, as in the STARSS23 dataset [7], and a source height of $h_s{=}0.9$m, reflecting the average event height in the training set, similar to a seated user. We also include sources at 1.5m, representing standing speakers. Although these conditions may not always hold, we argue that the model can learn prior knowledge about typical source heights based on class. For instance, speech is unlikely to originate from the ceiling or floor, whereas footsteps are naturally associated with the ground. Ideally, the model should determine when and how to incorporate such priors to refine distance estimation.

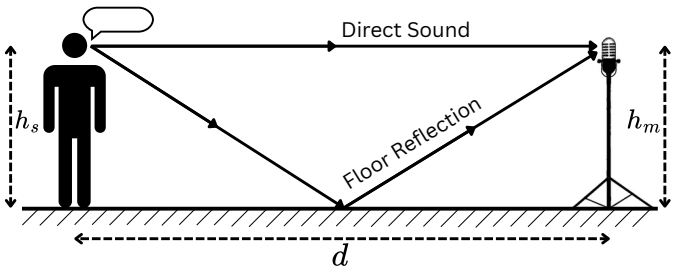

Fig. 1: Floor reflection path when source and receiver are at the same height ($h_s{=}h_m$) and separated by distance d.

TABLE I: Ideal direct sound and first reflection (1stRef) delays, with their corresponding initial time delay gaps (ITDGs), as the source distance increases, assuming the first reflection originates from the floor. We present cases for source heights of 1.5 m and 0.9 m, with microphone positioned at 1.5 m, sound speed of 343 m/s, and no additional interfering factors.

	Source Height: 1.5m			**Source Height: 0.9m**		
Dist	**Direct**	**1stRef**	**ITDG**	**Direct**	**1stRef**	**ITDG**
1.0 m	2.9 ms	9.2 ms	6.3 ms	3.4 ms	7.6 ms	4.2 ms
1.5 m	4.4 ms	9.8 ms	5.4 ms	4.7 ms	8.2 ms	3.5 ms
2.0 m	5.8 ms	10.5 ms	4.7 ms	6.1 ms	9.1 ms	3.0 ms
2.5 m	7.3 ms	11.4 ms	4.1 ms	7.5 ms	10.1 ms	2.6 ms
3.0 m	8.7 ms	12.4 ms	3.6 ms	8.9 ms	11.2 ms	2.3 ms

To design an input feature that captures early reflections, we conducted a preliminary study on ITDG variations across different source distances. We analyzed an 8s speech clip from the S3A Object-based Audio Drama dataset [28], [29] and convolved it with room impulse responses (RIRs) from SurrRoom 1.0 [30]. Specifically, we used the omnidirectional W channel of FOA RIRs recorded in the "Pop_Recording_Studio" at distances of [1m, 1.5m, 2m, 2.5m, 3m].

Fig. 2 shows the aligned RIRs, where the first reflection, i.e., the initial peak after the direct sound, shifts closer to the direct sound as distance increases, consistent with Table I. A later strong reflection, likely from the rear wall, also appears, with increasing delay at greater distances. While wall reflections depend on room size and geometry, making them unreliable for distance estimation, floor reflections offer a more robust and consistent cue for this task.

After spatializing the clip at different distances, we compute

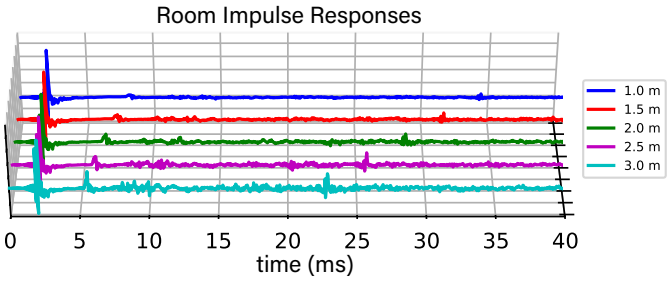

Fig. 2: RIRs from the omnidirectional FOA channel of the SurrRoom 1.0 dataset [30] ("Pop_Recording_Studio" room) used to spatialize speech at different distances. Direct sound peaks are temporally aligned for comparison.

the correlation coefficient and derive its energy envelope. The upper part of Fig. 3 shows the first 30 ms of the normalized autocorrelation coefficients (ACCs) at various distances. We observed that the second peak in the ACC aligns closely with the ideal ITDG values from Table I for h_s=1.5 m. To strengthen this representation capturing both the level and timing of the first reflection, we compute the short-term power of the ACC (stpACC). The stpACC is obtained by applying a Hann-windowed moving average (size: 8 samples) to the squared ACC coefficients. At 24 kHz sampling rate, this ∼0.3 ms window groups reflections from objects or surfaces within 10 cm of each other. The resulting stpACC features for spatialized speech signals are shown in the lower part of Fig. 3.

To leverage stpACC features for the 3D SELD task and allow concatenation with conventional SELD features (e.g., IVs and log-mel spectrograms), we represent them as 2D signals. This is achieved by computing the short-time autocorrelation function in the frequency domain as:

$$ACC(t,\tau) = \mathcal{F}^{-1}_{f\to\tau}(\mathbf{X}(t,f)\mathbf{X}^*(t,f)) \tag{3}$$

$$ACC^{\text{norm}}(t,\tau) = \frac{ACC(t,\tau)}{\max_\tau(|ACC(t,\tau)|)}, \quad \forall t \tag{4}$$

where $\mathbf{X}(t,f)$ is the STFT of the W channel, $(.)^*$ denotes complex conjugate, and $\mathcal{F}^{-1}_{f\to\tau}$ the inverse FFT from the frequency f to the time-lag domain τ. We then normalize each time bin t so that $ACC^{\text{norm}}(t,0)$=1, and convolve its square with an 8-sample Hann window to obtain $stpACC(t,\tau)$.

III. Experiments

A. Model Architecture

We evaluated the proposed distance features using a CNN-Conformer architecture, widely adopted for SELD [8], [31], [32]. It consists of a CNN encoder, a Conformer module [33], and feed-forward layers for 3D SELD predictions. The CNN encoder processes FOA-derived acoustic features, including IVs in log-mel domain (3 channels), log-mel spectrograms from FOA (4 channels), and the proposed distance features, DRR, D+R, or stpACC, forming an input of shape $C_{\text{in}}\times T_{\text{in}}\times F_{\text{in}}$. Here, T_{in} and F_{in} represent temporal and frequency (or time-lag) bins, respectively, with C_{in}=8 for DRR and sptACC or C_{in}=9 for D+R. The CNN encoder comprises four convolutional blocks with residual connections, each containing two 3×3 convolutional layers, BN, ReLU activation,

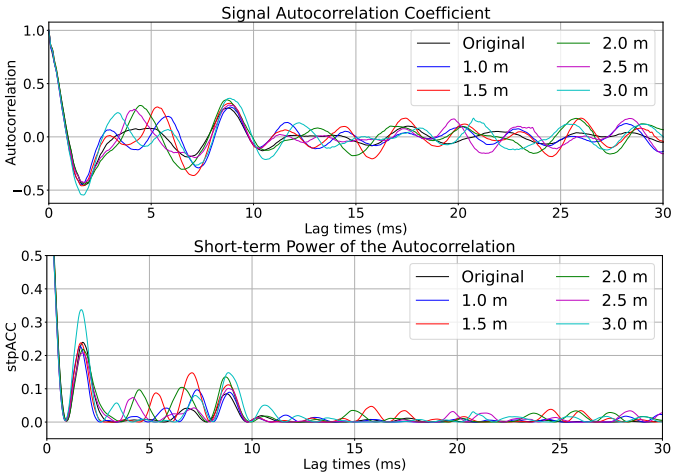

Fig. 3: Autocorrelation coefficient at varying distances (top). Short-term power of the autocorrelation (bottom).

and Avg pooling with a stride of 2, halving the temporal and frequency dimension at each block. The resulting tensor of shape $512\times T_{\text{in}}/16\times F_{\text{in}}/16$ is reshaped and frequency Avg pooling is applied to achieve a $T_{\text{in}}/16\times 512$ embedding. T_{in} is chosen so that $T_{\text{in}}/16$ matches the label frame rate (10 labels / sec). A Conformer module with four layers and eight attention heads processes this embedding, using depthwise convolutions with kernel size of 51. Finally, two feedforward layers predict multi-ACCDDOA vectors, modeling up to N=3 tracks [12]. As in previous 3D SELD works [11], [12], the model is trained using class-wise Auxiliary Duplicating Permutation Invariant Training (ADPIT) loss [6].

B. Dataset and Data Augmentation

We conducted our experiments using the STARSS23 dataset [7], which, to our knowledge, is the only public dataset for 3D SELD. Other well-known benchmarks for SELD, such as STARSS22 [34] or TAU-NIGENS Spatial Sound Events 2020 and 2021 [2], [3], do not include distance labels. STARSS23 [7] consists of ∼7.5h of real spatial recordings of acoustic scenes, temporally and spatially annotated, with 13 event classes. In our experiments, we employed the FOA audio format. The event class, DOA, and distance labels are provided at a resolution of 100ms. The dataset includes directional interferes, i.e., non-target sounds that should not be detected. Our experiments were conducted on the development set, which comes with a predefined train-test split. We evaluated our models on the test partition of the development set. To increase the size of the training data and mitigate overfitting, we augment the dataset by a factor of 8 using the audio channel swap (ACS) data augmentation [31]. We pre-trained our models using the synthetic 3D SELD data provided by the organizers of the DCASE2024 Task 3 Challenge, which consists of 20h of simulated data generated with RIRs, following a methodology similar to that described in [35]. We applied ACS data augmentation during pre-training too. We observed that pre-training our models is crucial for understanding the impact of different input features on distance estimation. This is likely because the model gains additional prior knowledge

TABLE II: Result with respective 95% confidence intervals achieved with the proposed distance input features. Each row represents a model trained using the concatenation of log-mel spectrograms, intensity vectors (IVs), and the different distance features. For the first row ("None"), only log-mel spectrograms and IVs are employed.

Distance Features	$F_{\leq 20°/1} \uparrow$	$DOAE \downarrow$	$RDE \downarrow$	$SELD \downarrow$
None	34.7% (29.7% - 39.3%)	**19.4°** (16.6° - 22.2°)	0.296 (0.273 - 0.355)	0.352 (0.332 - 0.385)
D+R	**36.4%** (31.1% - 41.6%)	22.0° (18.6° - 24.2°)	0.273 (0.234 - 0.298)	0.344 (0.314 - 0.367)
DRR	36.0% (30.8% - 41.4%)	20.1° (17.7° - 23.4°)	0.286 (0.240 - 0.315)	0.346 (0.319 - 0.368)
stpACC	35.9% (30.5% - 41.2%)	21.3° (11.9° - 30.6°)	**0.262** (0.225 - 0.296)	**0.341** (0.304 - 0.375)

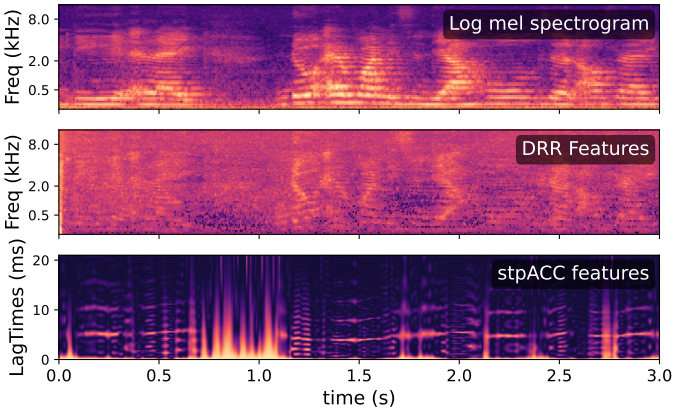

Fig. 4: Distance features with respective log mel spectrogram extracted from a sequence of STARSS23.

about sound events, enabling a better interpretation of the information encoded in the input features.

C. Metrics

To evaluate our models, we adopted the official metrics of the DCASE 2024 Task 3 Challenge [11] that are based on true positive (TP) and false positive (FP) predictions. A prediction is considered TP if the class prediction is correct and if its predicted DOA is within $\pm 20°$ from the target, and the relative distance error ($RDE{=}|L_p - L_r|/L_r$ with L_p and L_r being the predicted and reference distance, respectively) is smaller than 1. Metrics are computed at the frame level and for each class independently and then averaged across the number of classes. Based on these, the adopted metrics are the class- and location-dependent F1 score ($F_{\leq 20°/1}$), the class-dependent DOA error ($DOAE$), and the class-dependent relative distance error (RDE) [11]. We also include the $SELD$ score that encodes the overall 3D SELD performance and is achieved as: $SELD{=}\text{mean}((1 - F_{\leq 20°/1}), DOAE/180, RDE)$.

D. Hyper-parameters and Experimental Settings

We trained our models using 3-second audio chunks, extracted every 1 s for training and without overlap for testing. Spectrograms were generated via STFT with a 512-point Hann window and 150-sample hop size, yielding 480 temporal bins for 24kHz audio. Log-mel spectrograms (128 frequency bins) were computed for audio channels, DRR, D+R, and IV features. For stpACC features, we applied an STFT with a 1014-point Hann window. This ensures that, when considering only the positive time-lags $\tau > 0$, $stpACC(t, \tau)$ contains 512 time-lag bins, covering delays up to approximately 20 ms after the direct sound. We then downsample the time-lag dimension by a factor of 4 to achieve 128 bins and allow concatenation with the other features. Models were trained with batch size 32 using Adam optimizer for 50 epochs, selecting the best based on the lowest SELD score. The learning rate was 5e-5 for 30 epochs, then reduced by 5% per epoch.

E. Results

The results obtained using the proposed distance features are presented in Table II. Confidence intervals were estimated using the jackknife estimate of variance [36], applying the leave-one-out resampling technique to each of the 78 sequences in the test set. The table shows that all tested distance features contributed to a reduction in RDE, leading to an overall improvement in $SELD$ score. A small but not significant improvement was observed in $F_{\leq 20°/1}$. This is expected, as the distance features are specifically designed to enhance distance estimation, while the TP predictions used to compute $F_{\leq 20°/1}$ depend on an $RDE{<}1$ threshold, which is a relatively trivial condition to meet even without distance features.

The model without distance features achieved the lowest $DOAE$. However, considering the confidence intervals, this difference appears to fall within statistical noise. Nevertheless, the $DOAE$ confidence interval obtained with stpACC features is over three times larger than that of the other features, indicating greater variability and noisier estimates for this metric. Despite this, stpACC features also yielded the best RDE and the highest $SELD$ score. While DRR is a known indicator in distance perception [19], [22], [37], [38], D+R features yielded better distance estimates. We hypothesize that the model benefits from greater flexibility in learning task-relevant information from the direct and reverberant components separately.

IV. Conclusion

This paper introduces novel distance input features for 3D SELD, leveraging reverberation cues encoded in the audio signal. The first category of proposed features separates direct and reverberant components, which are either fed to the model independently or represented as a direct-to-reverberant ratio. The second category aims to enhance the model's understanding of early reflections, particularly the first reflection.

Experiments on STARSS23 indicate that the proposed features improve distance estimation, leading to enhanced overall SELD score. The most significant improvement was observed with short-term power of the autocorrelation features. Future research will explore the effectiveness of reverberation-based distance features across different 3D SELD data, including synthetic data, and various network architectures.

REFERENCES

[1] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, pp. 34–48, 2019.

[2] A. Politis, S. Adavanne, and T. Virtanen, "A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection," in *Detection and Classification of Acoustic Scenes and Events Workshop*, 2020.

[3] A. Politis, S. Adavanne, D. Krause, A. Deleforge, P. Srivastava, and T. Virtanen, "A dataset of dynamic reverberant sound scenes with directional interferers for sound event localization and detection," in *Detection and Classification of Acoustic Scenes and Events Workshop*, 2021.

[4] Y. Cao, T. Iqbal, Q. Kong, Y. Zhong, W. Wang, and M. D. Plumbley, "Event-independent network for polyphonic sound event localization and detection," in *Detection and Classification of Acoustic Scenes and Events Workshop*, 2020.

[5] Y. Cao, T. Iqbal, Q. Kong, F. An, W. Wang, and M. D. Plumbley, "An improved event-independent network for polyphonic sound event localization and detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 885–889.

[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 *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 316–320.

[7] 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 *International Conference on Neural Information Processing Systems*, 2023.

[8] D. Berghi, P. Wu, J. Zhao, W. Wang, and P. J. B. Jackson, "Fusion of audio and visual embeddings for sound event localization and detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2024.

[9] Y. Jiang, Q. Wang, J. Du, M. Hu, P. Hu, Z. Liu, S. Cheng, Z. Nian, Y. Dong, M. Cai, X. Fang, and C.-H. Lee, "Exploring audio-visual information fusion for sound event localization and detection in low-resource realistic scenarios," in *IEEE International Conference on Multimedia and Expo*, 2024, pp. 1–6.

[10] A. S. Roman, B. Balamurugan, and R. Pothuganti, "Enhanced sound event localization and detection in real 360-degree audio-visual soundscapes," *ArXiv*, vol. abs/2401.17129, 2024.

[11] D. Diaz-Guerra, A. Politis, P. Ariyakulam Sudarsanam, K. Shimada, D. Krause, K. Uchida, Y. Koyama, N. Takahashi, S. Takahashi, T. Shibuya, Y. Mitsufuji, and T. Virtanen, "Baseline models and evaluation of sound event localization and detection with distance estimation in DCASE 2024 Challenge," in *Detection and Classification of Acoustic Scenes and Events Workshop*, 2024, pp. 41–45.

[12] D. A. Krause, A. Politis, and A. Mesaros, "Sound event detection and localization with distance estimation," in *European Signal Processing Conference*, 2024, pp. 286–290.

[13] H. Hong, Q. Wang, J. Du, R. Wei, M. Cai, and X. Fang, "MVANet: Multi-stage video attention network for sound event localization and detection with source distance estimation," *ArXiv*, vol. abs/2411.14153, 2024.

[14] D. Berghi and P. J. B. Jackson, "Audio inputs for active speaker detection and localization via microphone array," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2023.

[15] C. Knapp and G. C. 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.

[16] T. N. Tho 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 *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 716–720.

[17] D. H. Mershon and L. E. King, "Intensity and reverberation as factors in the auditory perception of egocentric distance," *Perception & Psychophysics*, vol. 18, pp. 409–415, 1975.

[18] C. Sheeline, "An investigation of the effects of direct and reverberant signal interactions on auditory distance perception," Ph.D., Stanford University, 1982. [Online]. Available: https://ccrma.stanford.edu/files/papers/stanm13.pdf

[19] D. Griesinger, "The importance of the direct to reverberant ratio in the perception of distance, localization, clarity, and envelopment, part one." *The Journal of the Acoustical Society of America*, vol. 125, pp. 2483–2483, 2009.

[20] Y.-C. Lu and M. Cooke, "Binaural estimation of sound source distance via the direct-to-reverberant energy ratio for static and moving sources," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1793–1805, 2010.

[21] E. Georganti, T. May, S. van de Par, A. Harma, and J. Mourjopoulos, "Speaker distance detection using a single microphone," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 1949–1961, 2011.

[22] S. Chitreddy and P. Jackson, "Source Distance Perception with Reverberant Spatial Audio Object Reproduction of Real Rooms," in *Forum Acusticum*, 2020, pp. 2079–2086.

[23] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind mimo impulse response shortening," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 10, pp. 2707–2720, 2012.

[24] L. Drude, J. Heymann, C. Boeddeker, and R. Haeb-Umbach, "NARA-WPE: A Python package for weighted prediction error dereverberation in Numpy and Tensorflow for online and offline processing," in *Speech Communication; 13th ITG-Symposium*, 2018, pp. 1–5.

[25] D. Berghi and P. J. B. Jackson, "Leveraging reverberation and visual depth cues for sound event localization and detection with distance estimation," in *Techical Report of DCASE Challenge*, 2024.

[26] A. W. Bronkhorst and T. Houtgast, "Auditory distance perception in rooms," *Nature*, vol. 397, pp. 517–520, 1999.

[27] N. Kaplanis, S. Bech, S. J. Holdt, and T. van Waterschoot, "Perception of reverberation in small rooms: A literature study," in *Audio Engineering Society Conference*, 2014.

[28] C. Cieciura, E. Bargiacchi, and P. J. B. Jackson, "Authoring inter-compatible flexible audio for mass personalization," in *The 157th Audio Engineering Society Convention*, 2024.

[29] J. Woodcock, C. Pike, F. Melchior, P. Coleman, A. Franck, and A. Hilton, "Presenting the S3A object-based audio drama dataset," in *The 140th Audio Engineering Society Convention*, 2016.

[30] C. Cieciura, M. Volino, and P. J. B. Jackson, "SurrRoom 1.0 Dataset: Spatial room capture with controlled acoustic and optical measurements," in *The 154th Audio Engineering Society Convention*, 2023.

[31] 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.

[32] L. Xue, H. Liu, Y. Zhou, and L. Gan, "Resnet-conformer network using multi-scale channel attention for sound event localization and detection in real scenes," in *International Conference on Wireless Communications and Signal Processing*, 2023.

[33] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Interspeech*, 2020, pp. 5036–5040.

[34] 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," *Detection and Classification of Acoustic Scenes and Events Workshop*, 2022.

[35] 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," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2024.

[36] B. Efron and C. Stein, "The jackknife estimate of variance," *The Annals of Statistics*, vol. 9, no. 3, pp. 586–596, 1981.

[37] P. Zahorik, D. Brungart, and A. Bronkhorst, "Auditory distance perception in humans: A summary of past and present research," *Acta Acustica United With Acustica*, vol. 91, pp. 409–420, 2005.

[38] A. Kolarik, B. Moore, P. Zahorik, S. Cirstea, and S. Pardhan, "Auditory distance perception in humans: a review of cues, development, neuronal bases, and effects of sensory loss," *Atten Percept Psychophys*, vol. 78, p. 373–395, 2016.