## Technical Summary: Sound Event Localization and Detection (SELD) Enhancements and Datasets

This document summarizes advancements in sound event localization and detection (SELD), focusing on two key areas: enhancement of 3D SELD performance using reverberation-based features and the characteristics of datasets designed for SELD research.

### Reverberation-based Features for 3D SELD with Distance Estimation

**Objective:** To enhance 3D SELD performance by incorporating reverberation-based features for improved distance estimation, leveraging the distance-related information encoded in reverberation.

**Methodology:**

1. **Feature Extraction:** Two novel feature formats were developed:
   * **Direct-to-Reverberant Ratio (DRR):** This feature utilizes the ratio between direct and reverberant sound components to estimate distance. The direct sound component, (d(t)), is estimated using Weighted Prediction Error (WPE) dereverberation. The reverberant component, (r(t)), is derived by subtracting the direct signal from the original signal. The DRR feature is calculated in the log-mel space: [ \mathbf{DRR}^{\mathrm{mel}}(t,k)=10\cdot\log\_{10}\left(\mathbf{P}*{\mathrm{ DRR}}^{\mathrm{mel}}(t,k)\right) \tag{1} ] [ \mathbf{P}*{\mathrm{DRR}}^{\mathrm{mel}}(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}}) is the mel filter bank, and (\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. A variant (D+R) separately feeds log-mel spectrograms of (\mathbf{D}(t,f)) and (\mathbf{R}(t,f)) into the model.
   * **Short-Term Power of the Autocorrelation (stpACC):** This feature extracts information about early reflections, including the initial time delay gap (ITDG), by leveraging signal autocorrelation. The autocorrelation function is computed in the frequency domain and normalized as: [ ACC(t,\tau)=\mathcal{F}*{f\rightarrow\tau}^{-1}(\mathbf{X}(t,f)\mathbf{X}^{\*} (t,f)) \tag{3} ] [ ACC^{\mathrm{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. The short-term power is derived by applying a Hann-windowed moving average to the squared ACC coefficients.
2. **Model Architecture:** A CNN-Conformer architecture processes FOA-derived acoustic features (intensity vectors (IVs), log-mel spectrograms) and the proposed distance features. The CNN encoder contains four convolutional blocks with residual connections and Avg pooling. Frequency Avg pooling is applied after reshaping the resulting tensor. A Conformer module (four layers, eight attention heads) processes the embedding. Two feedforward layers predict multi-ACCDOA vectors. The model is trained using class-wise Auxiliary Duplicating Permutation Invariant Training (ADPIT) loss.
3. **Dataset and Augmentation:** Experiments used the STARSS23 dataset (FOA audio format), augmented by a factor of 8 using audio channel swap (ACS). Models were pre-trained on synthetic 3D SELD data from the DCASE2024 Task 3 Challenge, also with ACS augmentation.
4. **Metrics:** Evaluation used the DCASE 2024 Task 3 Challenge metrics: class- and location-dependent F1 score ((F\_{\leq 20^{\circ}/1})), class-dependent DOA error ((DOAE)), class-dependent relative distance error ((RDE)), and the overall SELD score: (SELD)=(\mathrm{mean}((1-F\_{\leq 20^{\circ}/1}),DOAE/180,RDE)).

**Technical Findings:**

* Incorporating distance features reduced the Relative Distance Error (RDE) and improved the overall SELD score.
* The stpACC features yielded the best RDE (0.262) and the highest SELD score (0.341).
* D+R features outperformed DRR features, indicating separate learning of direct and reverberant components is beneficial.
* The model without distance features achieved the lowest DOAE (19.4�), although the statistical significance is questionable.

**Technical Conclusions:**

Reverberation-based features, especially stpACC, are effective in improving distance estimation and overall SELD performance. These features provide complementary information to conventional SELD features.

### Sound Event Localization and Detection Datasets

**Key Objectives:**

* To facilitate research in sound event localization and detection.
* To provide realistic, reverberant spatial sound scenes for algorithm training and testing.
* To incorporate moving sources and directional interferers to increase complexity.
* To enable the development of robust SELD systems capable of handling dynamic acoustic environments.

**Main Technical Specifications/Characteristics:**

* **Reverberant Spatial Sound Scenes:** Datasets simulate realistic acoustic environments with varying degrees of reverberation.
* **Moving Sources:** Include sound events originating from moving sources, creating dynamic spatial audio scenes.
* **Directional Interferers:** Incorporate directional noise sources.
* **Dynamic Reverberant Sound Scenes:** Feature dynamic reverberation, where the reverberation characteristics change over time.

**Technical Conclusions:**

The SELD datasets are a valuable resource for developing and testing advanced algorithms, addressing critical challenges in real-world SELD applications by focusing on realistic acoustic conditions, moving sources, and directional interference.