LEAD Dataset: How Can Labels for Sound Event Detection Vary Depending on Annotators?

APSIPA ASC 2024

Naoki Koga*, Yoshiaki Bando†, Keisuke Imoto*

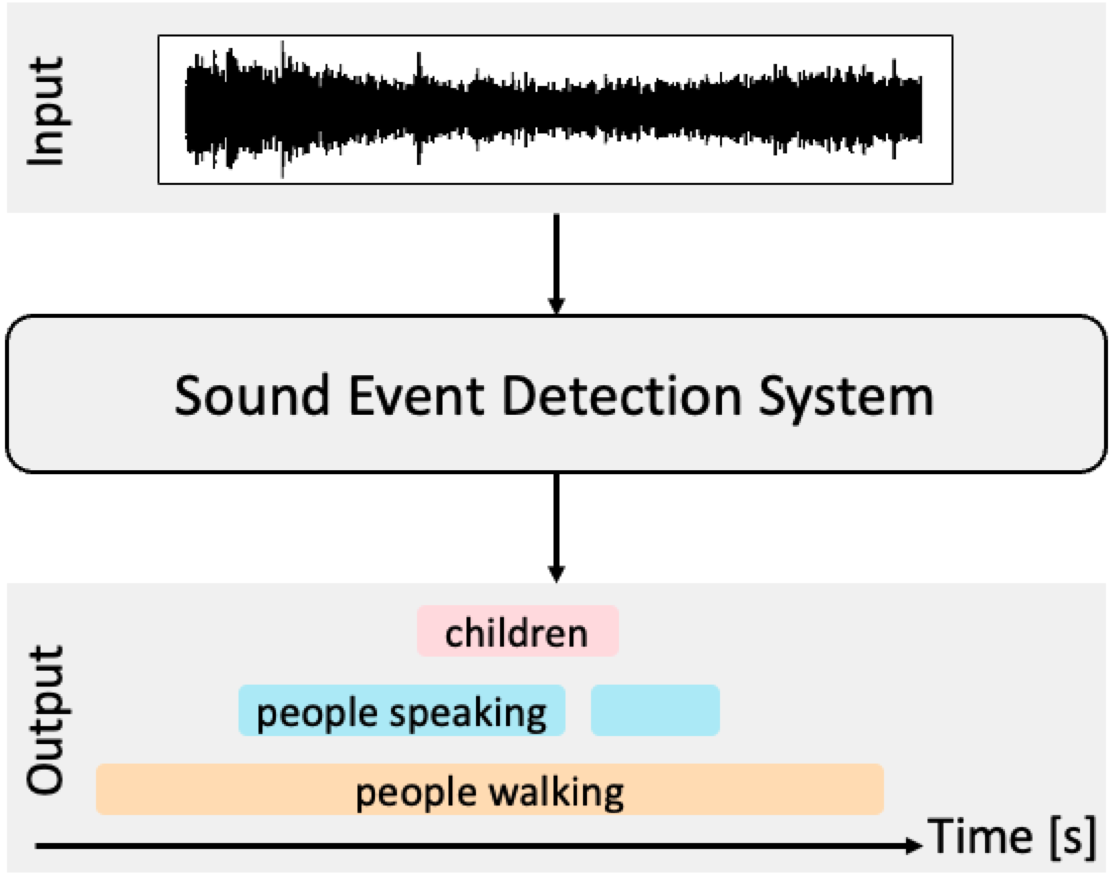
*Doshisha University, †National Institute of Advanced Industrial Science and Technology (AIST)

Imoto Lab

Dec. 5, 2024

Background: Sound event detection (SED)

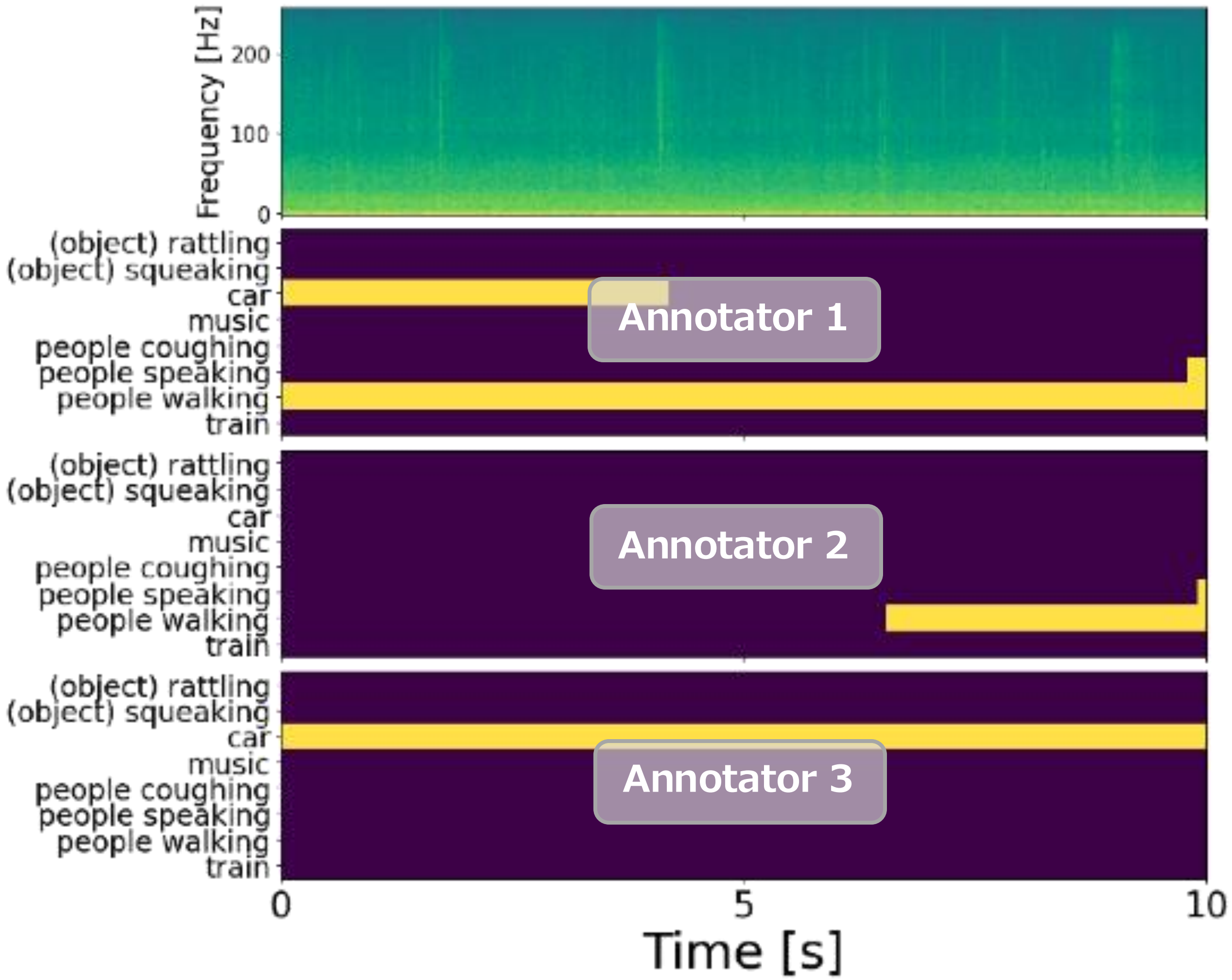
- **Task of estimating the types, onsets, and offsets of sound events[1, 2]**
  - e.g., “children,” “people walking”



[1] A. Mesaros, et al., “Sound event detection in the DCASE 2017 challenge,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 6, pp. 992–1006, 2019.
[2] A. Mesaros, et al., “Sound event detection: A tutorial,” IEEE Signal Processing Magazine, vol. 38, no. 5, pp. 67–83, 2021.

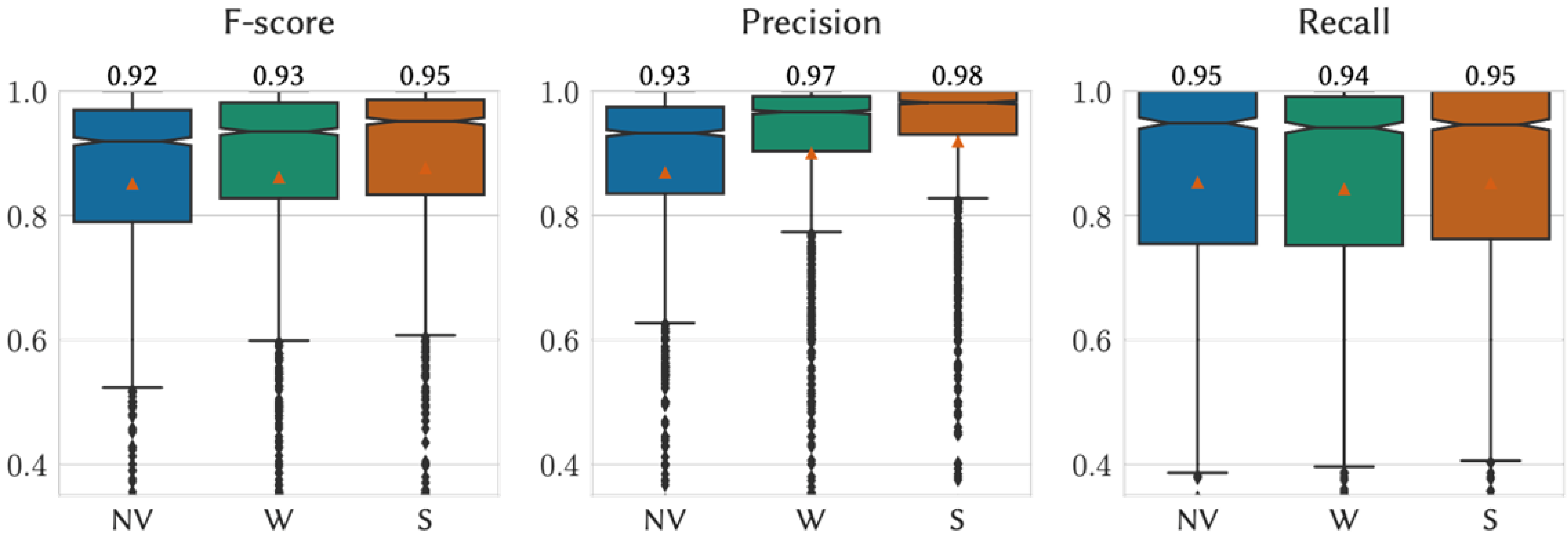
Problem: Annotating strong labels in SED

- **Variations in the types, onsets, and offsets of sound events**
  - different types, onsets and offsets depending on three annotators



Related work 1: Effects of visualizations on the annotations [1]

- **Visualization improves the quality of strong labels assigned to sound signals.**

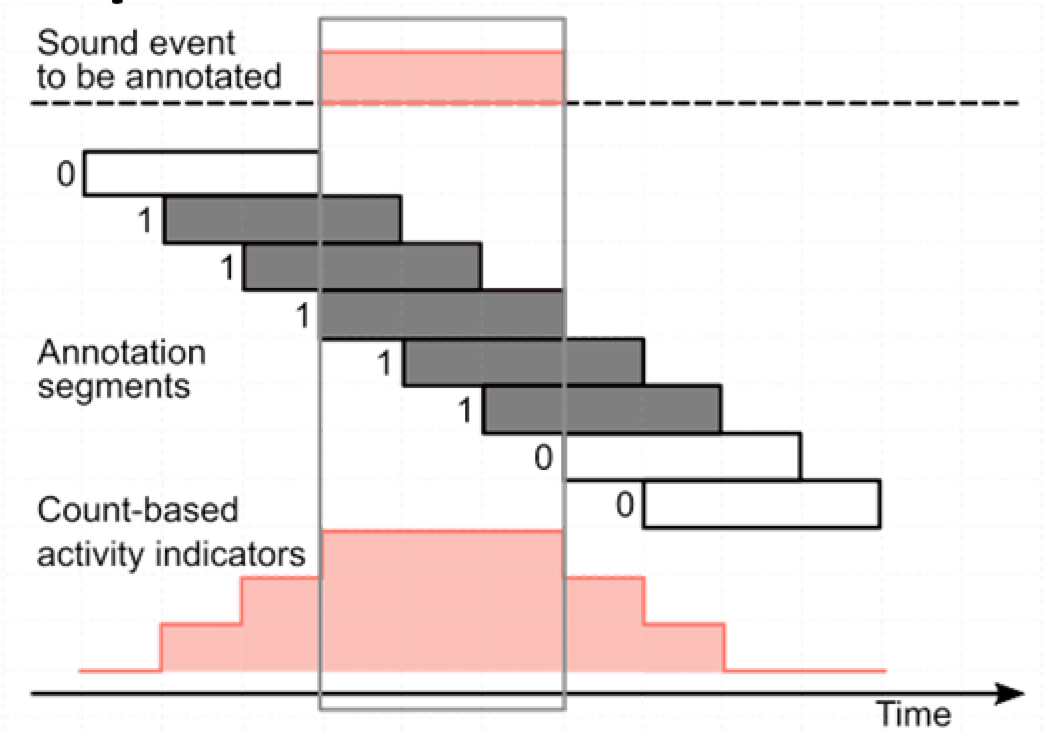


NV: no visualization, W: waveform, S: spectrogram

[1] M. Cartwright, et al., "Seeing sound: Investigating the effects of visualizations and complexity on crowdsourced audio annotations," ACM Transactions on Computer-Human Interaction, vol. 1, no. 29, pp. 1–21, 2017.

Related work 2: Reliable strong labels in SED [1]

- **Substitution of the annotation of strong labels with the annotation of multiple weak labels**



What are the characteristics of the variations in strong labels?

[1] I. Martín-Morató et al., “Strong labeling of sound events using crowdsourced weak labels and annotator competence estimation,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 902–914, 2023.

Overview of our contributions

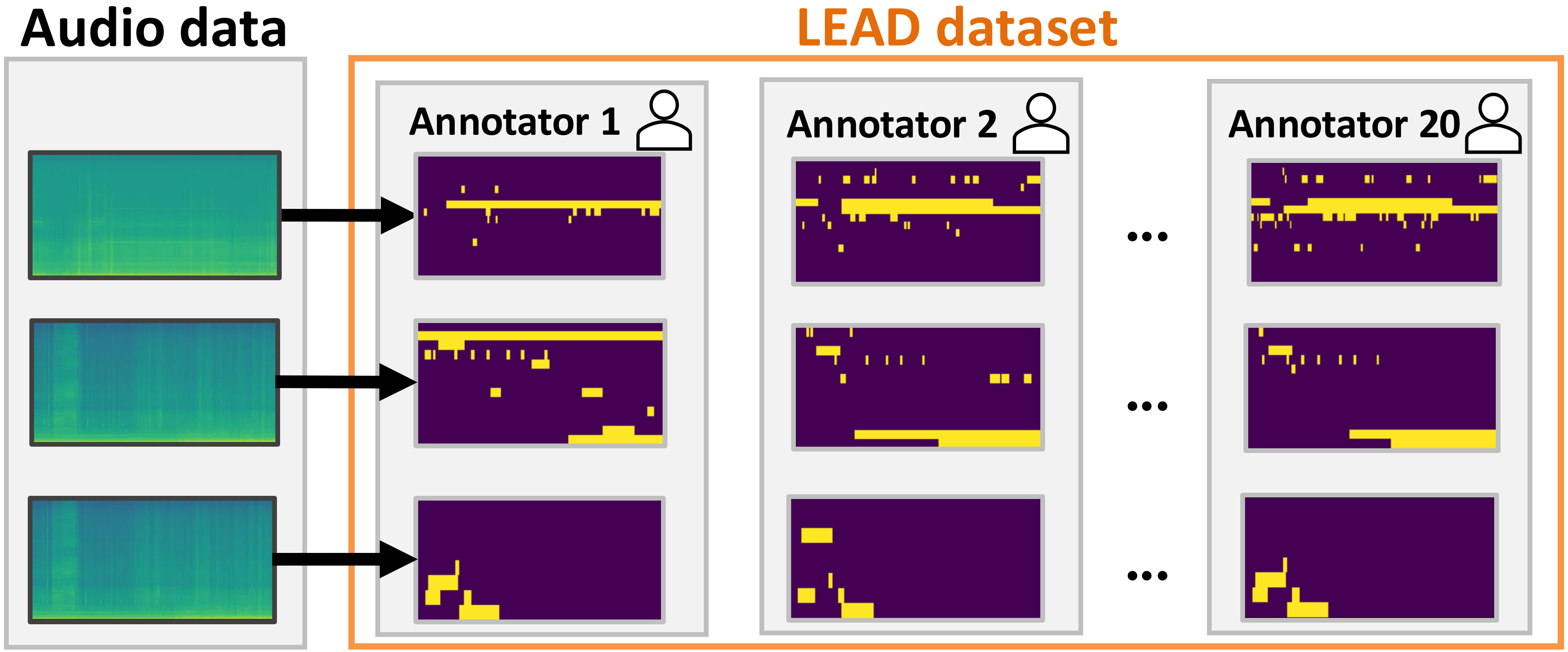
- Building the LEAD dataset
- Analyses with the LEAD dataset
- Experiment with the LEAD dataset

Overview of our contributions

- **Building the LEAD dataset**
- Analyses with the LEAD dataset
- Experiment with the LEAD dataset

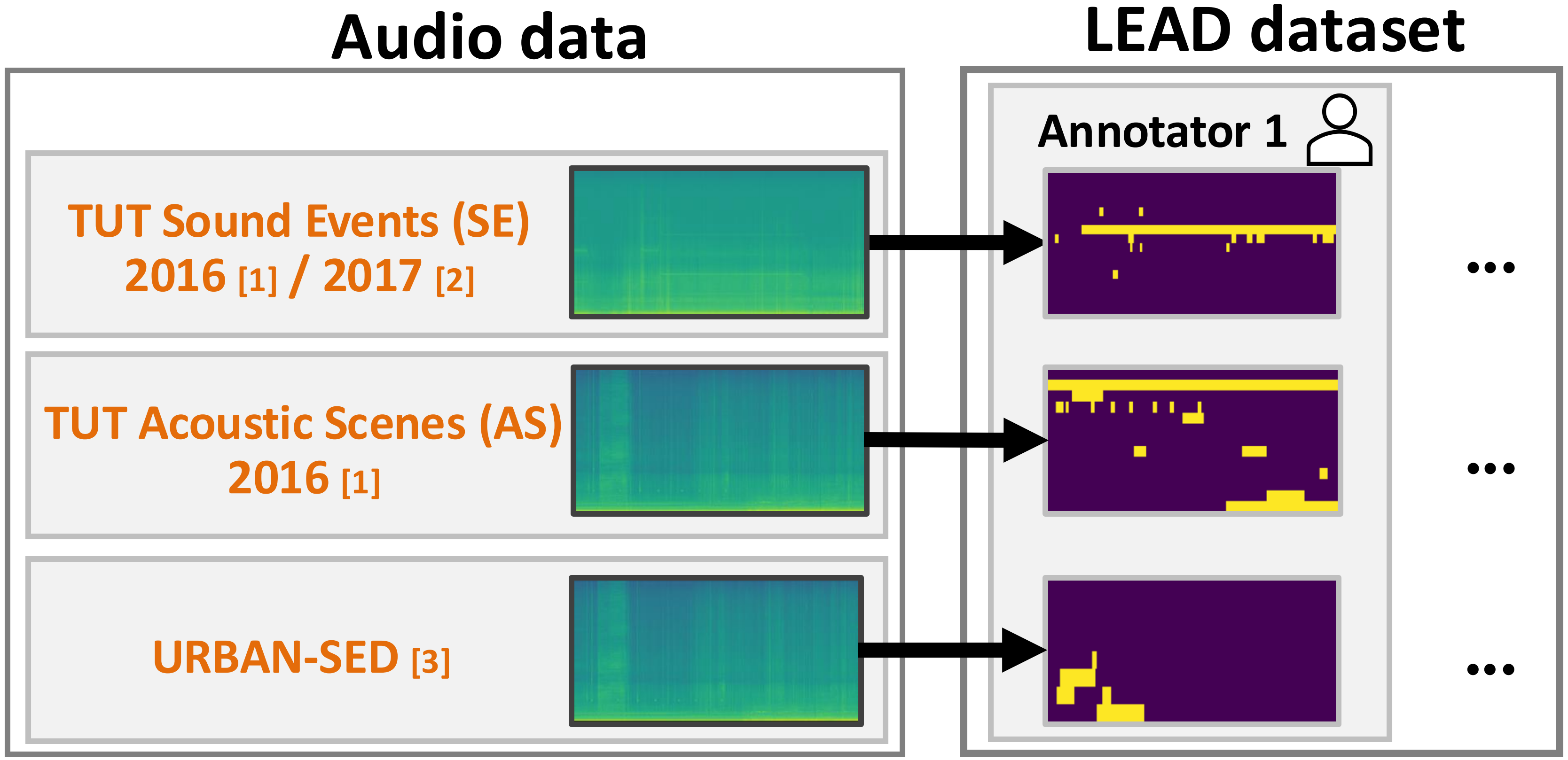
Building the LEAD dataset

- **Dataset for a better understanding of the variations in strong labels**
  - The LEAD dataset provides distinct strong labels for each clip annotated by 20 different annotators.



Data source

- **LEAD dataset has 20 strong labels for 5.67 hours of sound.**
  - Sound from four previous datasets

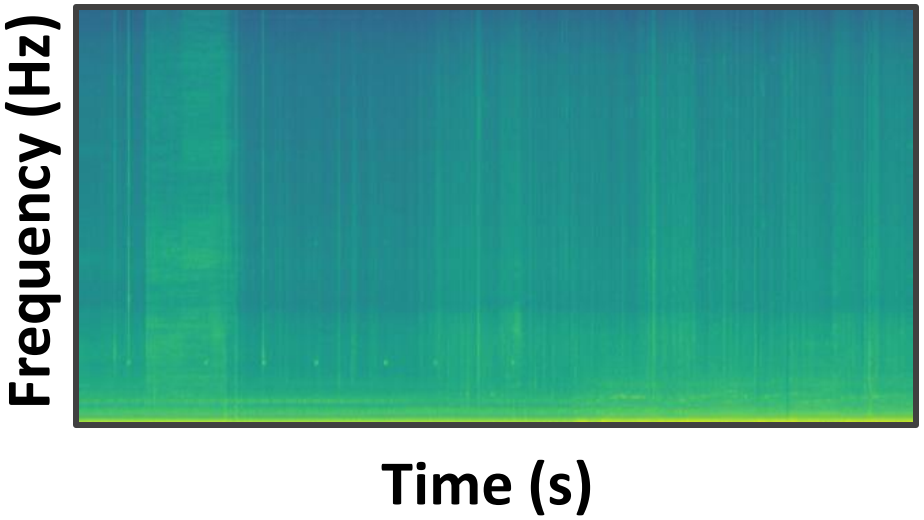


[1] A. Mesaros, et al., “TUT database for acoustic scene classification and sound event detection,” Proc. EUSIPCO, pp. 1128–1132, 2016.
[2] A. Mesaros, et al., “Challenge setup: Tasks, datasets and baseline system,” Proc. Workshop on DCASE, pp. 85–92, 2017.
[3] J. Salamon, et al., “Scaper: A library for soundscape synthesis and augmentation,” Proc. WASPAA, pp. 344–348, 2017.

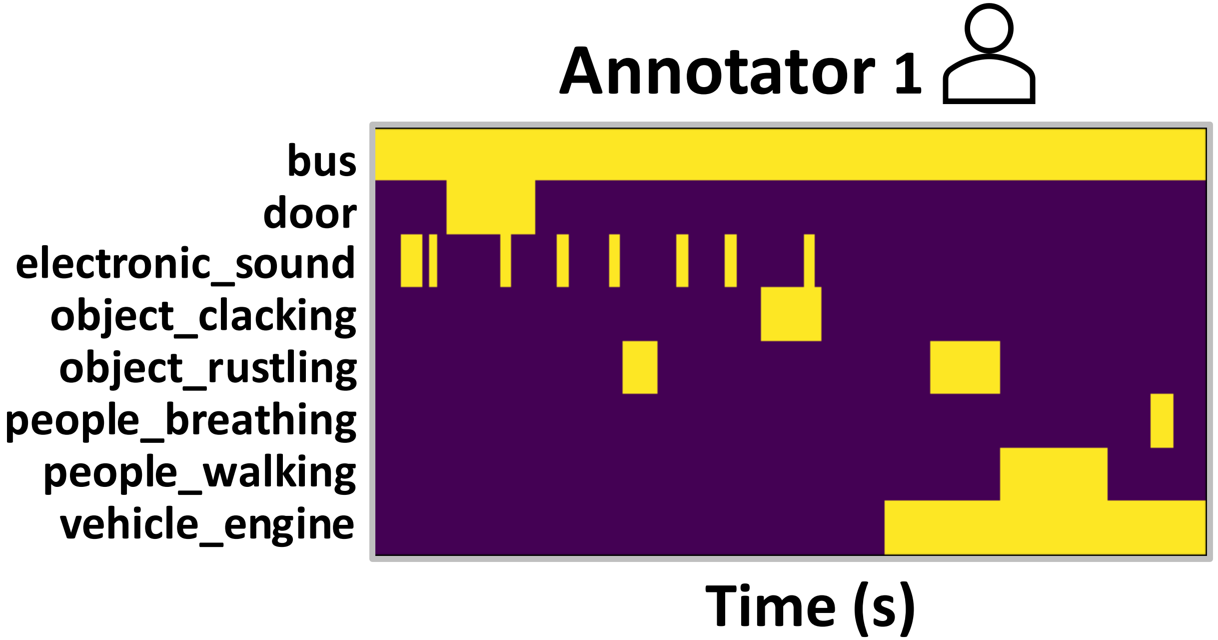
Annotation procedure

1. **Assigning strong labels to sound signals**
2. **Assigning a confidence score (CS) to the selected class, onset, and offset for each sound event instance**

The sound of bus

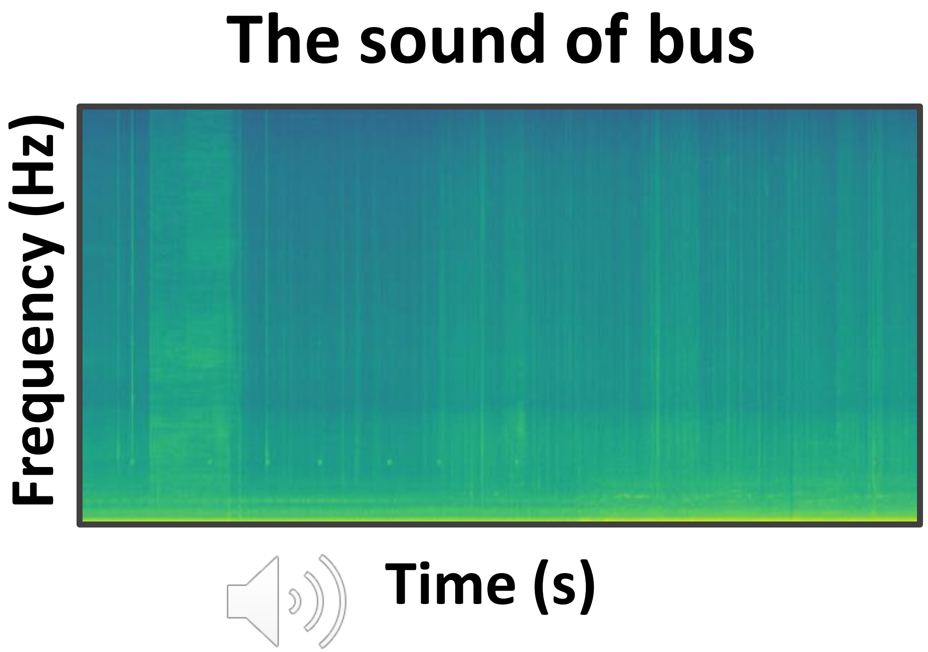


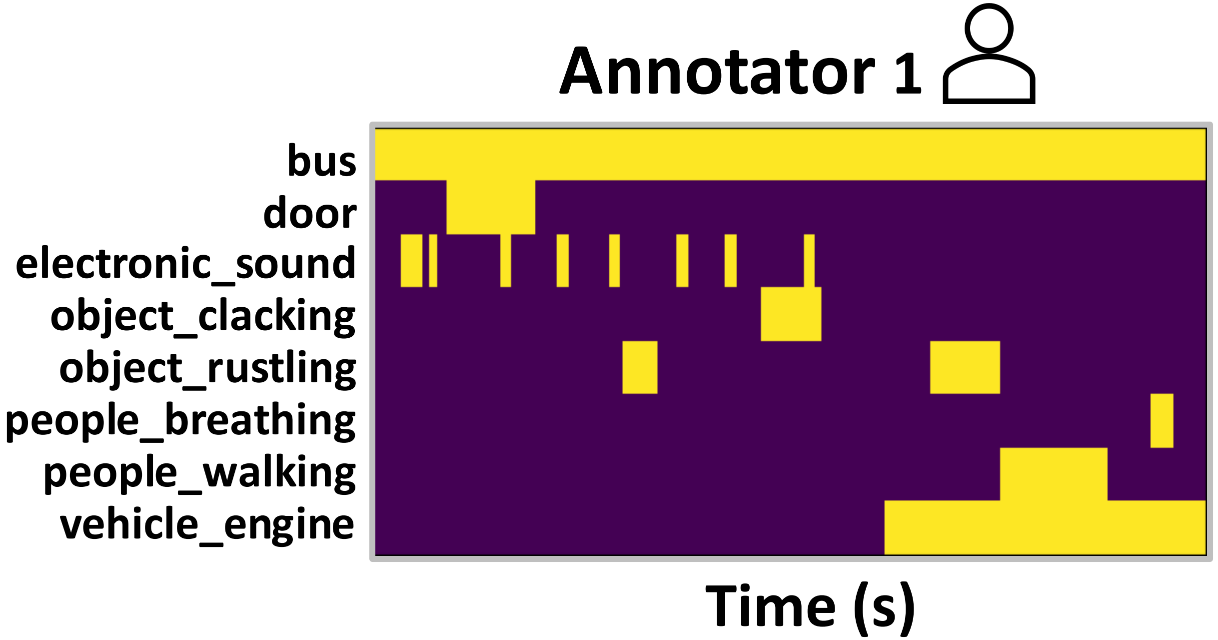
Annotator 1



Assigning strong labels to sound signals

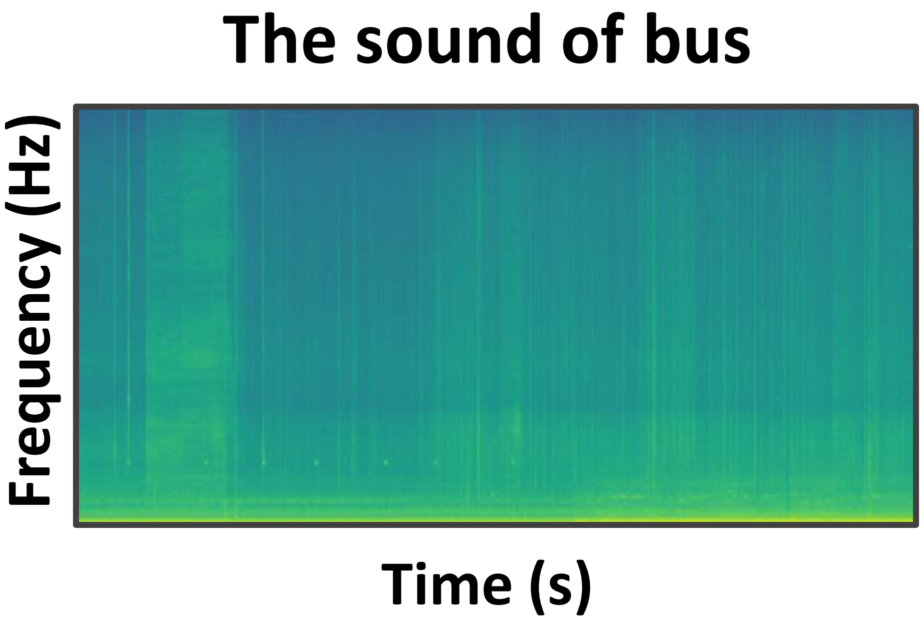
- All annotators selected classes, onsets, and offsets from the candidates.

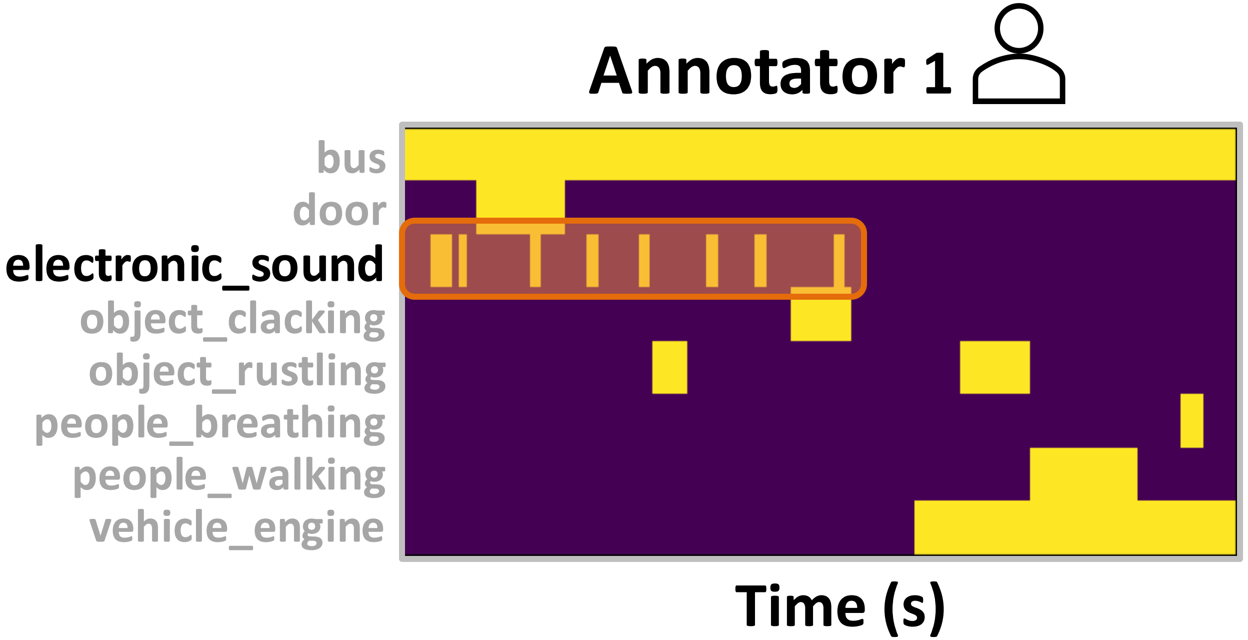




Assigning strong labels to sound signals

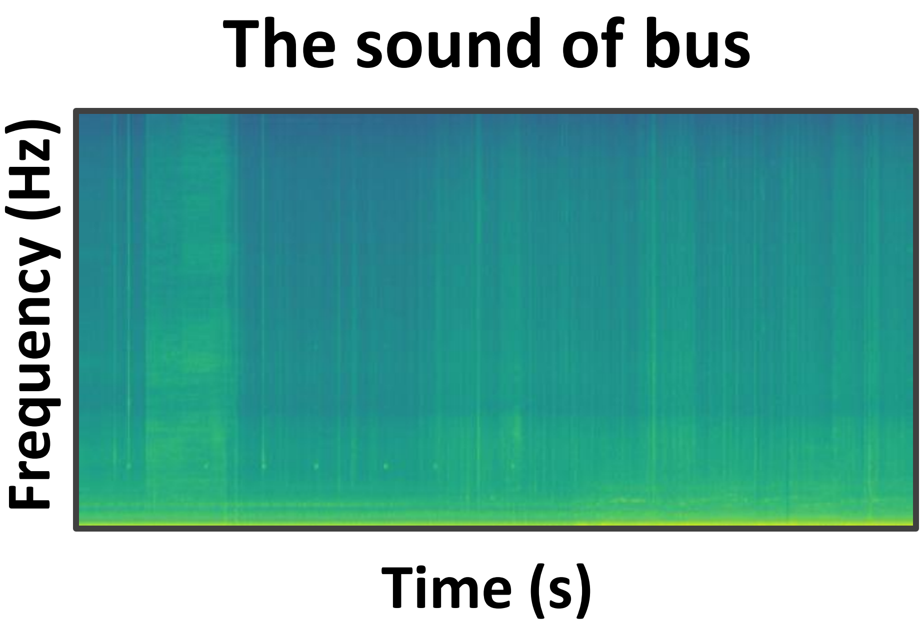
- **All annotators selected classes, onsets, and offsets from the candidates.**

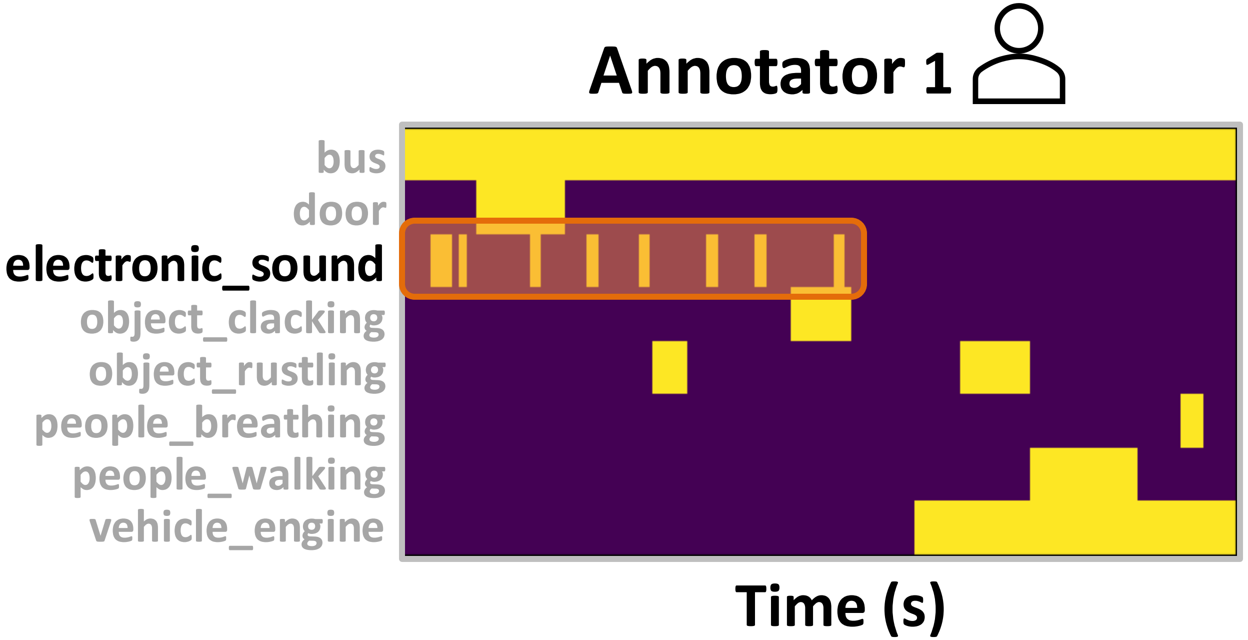




Assigning confidence scores (CSs) to sound events

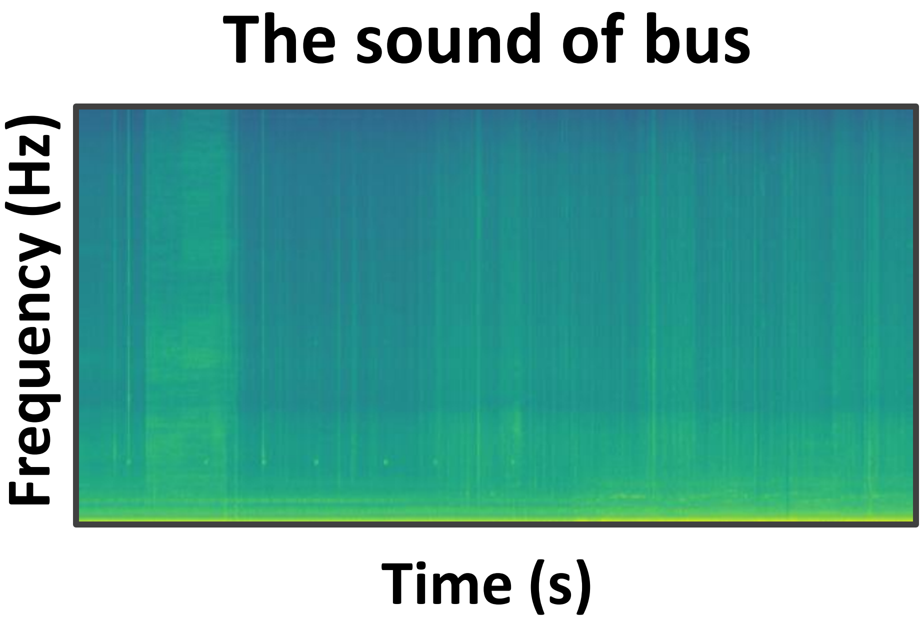
- **Assigning a CS to the selected class, onset, and offset for each sound event instance**
  - A five-point scale ranging from “very unconfident” to “very confident”

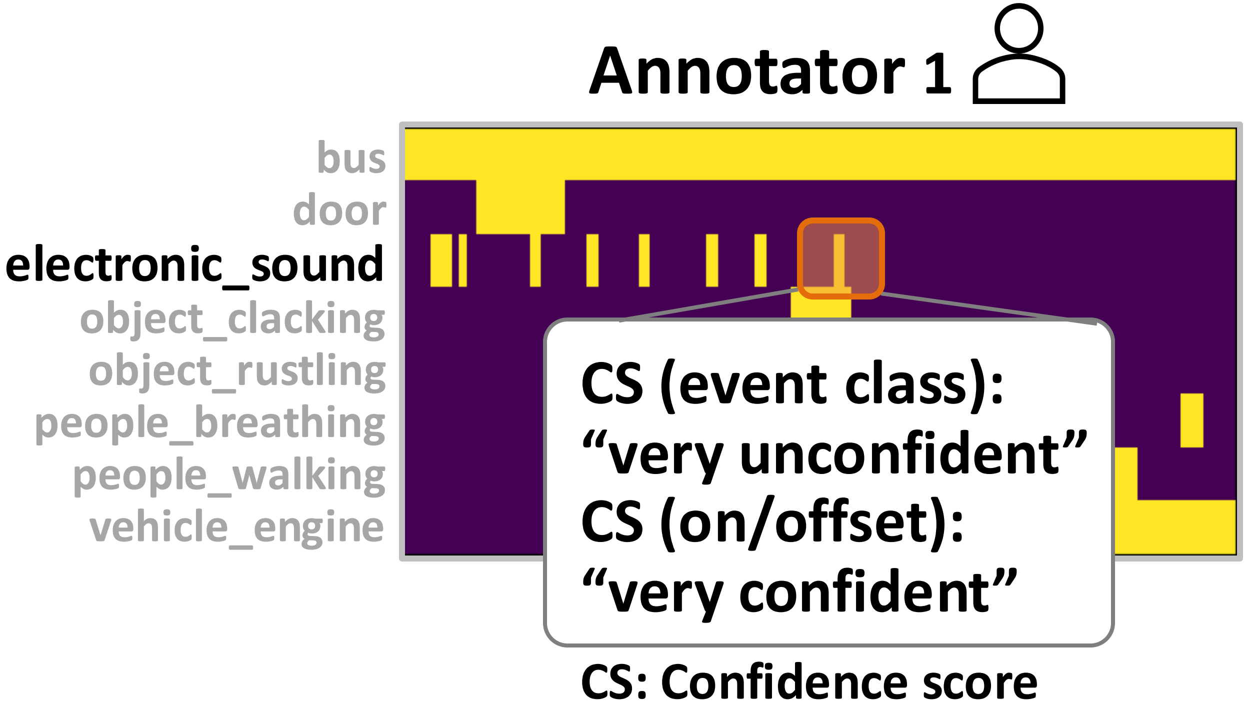




Assigning confidence scores (CSs) to sound events

- **Assigning a CS to the selected class, onset, and offset for each sound event instance**
  - A five-point scale ranging from “very unconfident” to “very confident”





Overview of our contributions

- Building the LEAD dataset
- **Analyses with the LEAD dataset**
- Experiment with the LEAD dataset

Overview of the analyses with the LEAD dataset

- **Analysis 1: Categorizing the variations in strong labels manually**
- **Analysis 2: Confirming a relationship between strong labels and CSs**
  - Analysis 2-a: Objective analysis of CSs
  - Analysis 2-b: Subjective analysis of CSs

Analysis 1: Categorizing the variations in strong labels

- **Purpose: To make the variations in strong labels distinguishable**

Variations in sound event classes

- Label deletion
- Label insertion
- Label integration
- Label substitution
- Hierarchical label substitution

Temporal variations in strong labels

Analysis 1: Categorizing the variations in strong labels

- **Purpose: To make the variations in strong labels distinguishable**

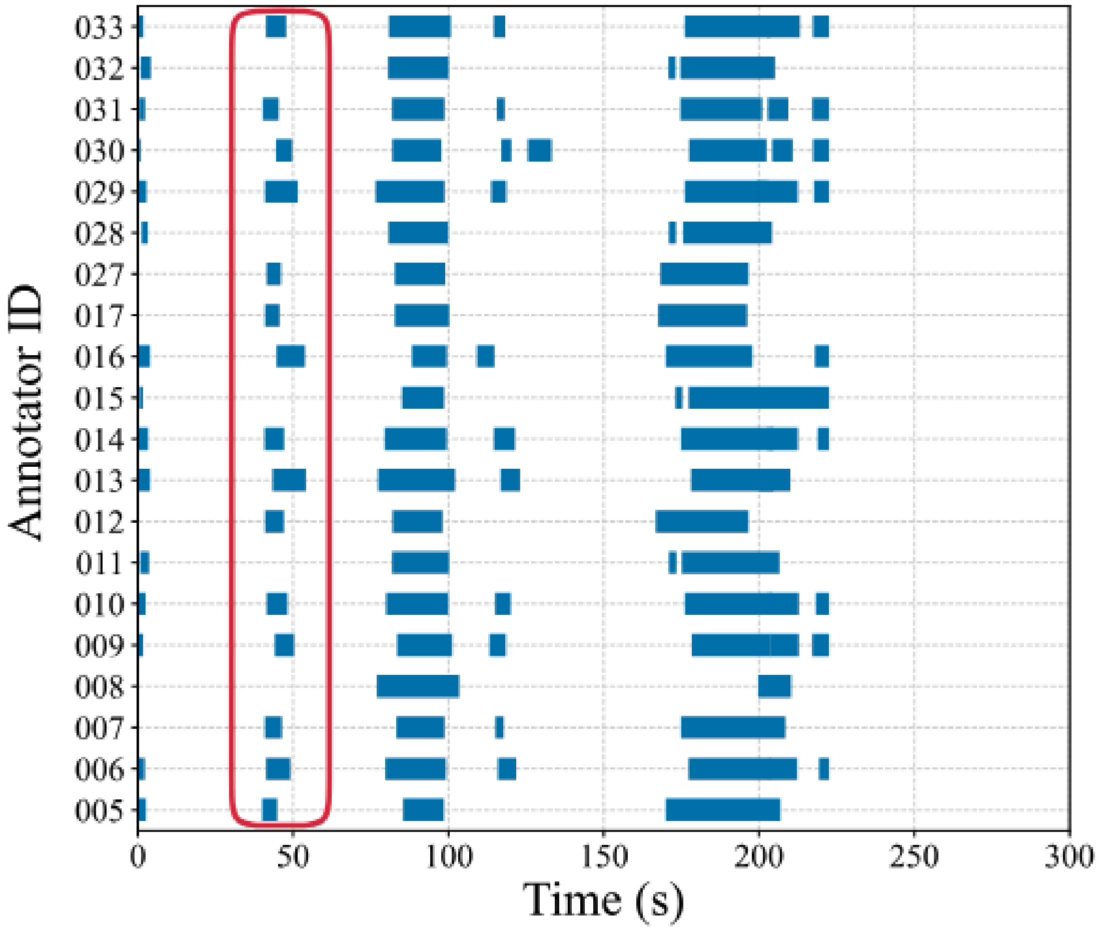
Variations in sound event classes

- Label deletion
- Label insertion
- Label integration
- Label substitution
- Hierarchical label substitution

Temporal variations in strong labels

Label deletion

- **An event label which is not assigned to an expected sound event**
  - e.g., “bird_singing” in b006.wav



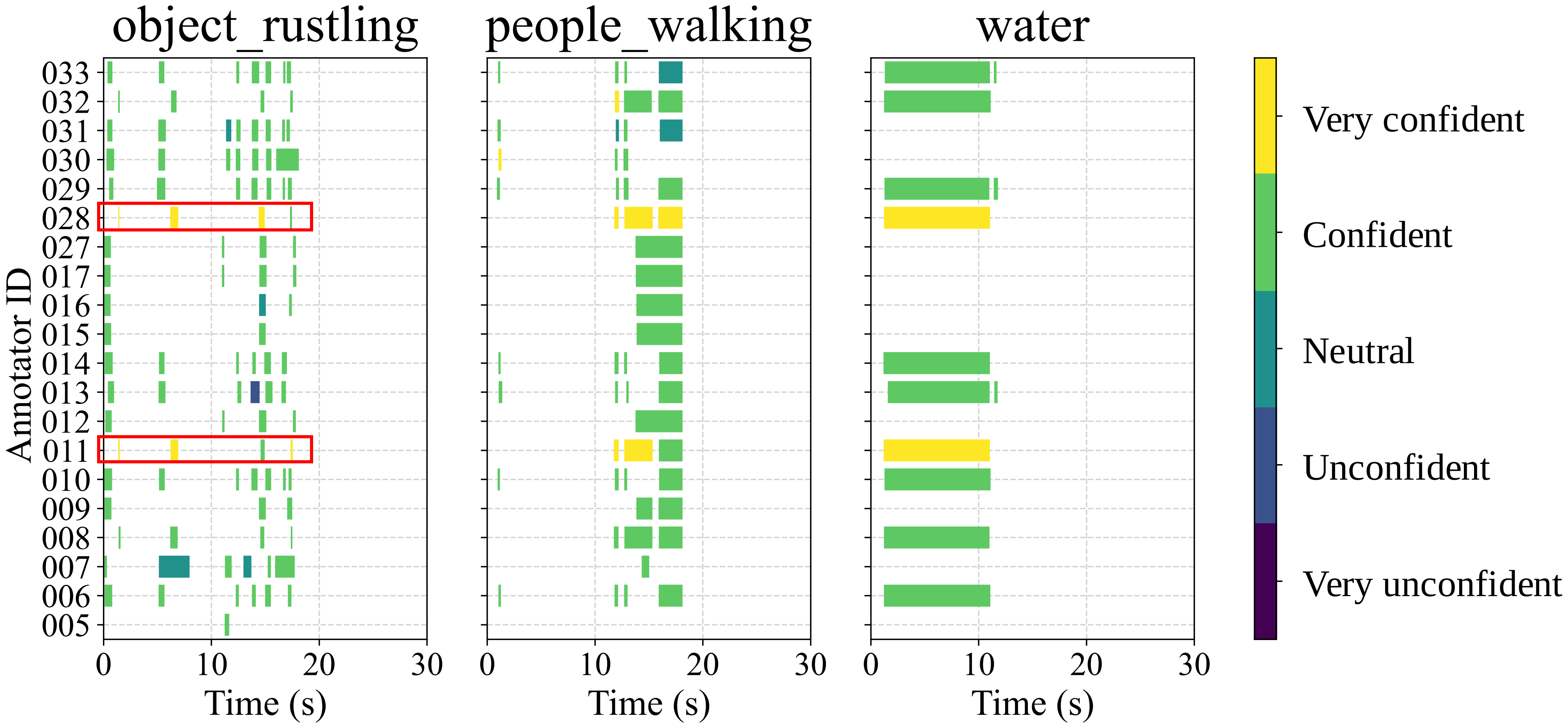
Analysis 2-a: Objective analysis of CSs

- **Investigating the differences of average CSs between real-world and synthetic sounds**
  - No significant difference can be seen between the average CSs of URBAN-SED and other datasets.

Strong label Sound length	TUT SE 2016	TUT SE 2017 120 s – 360 s	TUT AS 2016 30 s	URBAN-SED 10 s
Confidence score on sound event class	3.94 ± 0.16	4.00 ± 0.17	3.94 ± 0.12	4.03 ± 0.21
Confidence score on onset/offset	4.04 ± 0.12	4.00 ± 0.19	4.11 ± 0.11	4.13 ± 0.13

Analysis 2-b: Subjective analysis of temporal CSs

- **Visualizing the relationship between CSs and 20 strong labels**
  - Temporally shorter sound events with high CS were more likely to vary over time.



Overview of our contributions

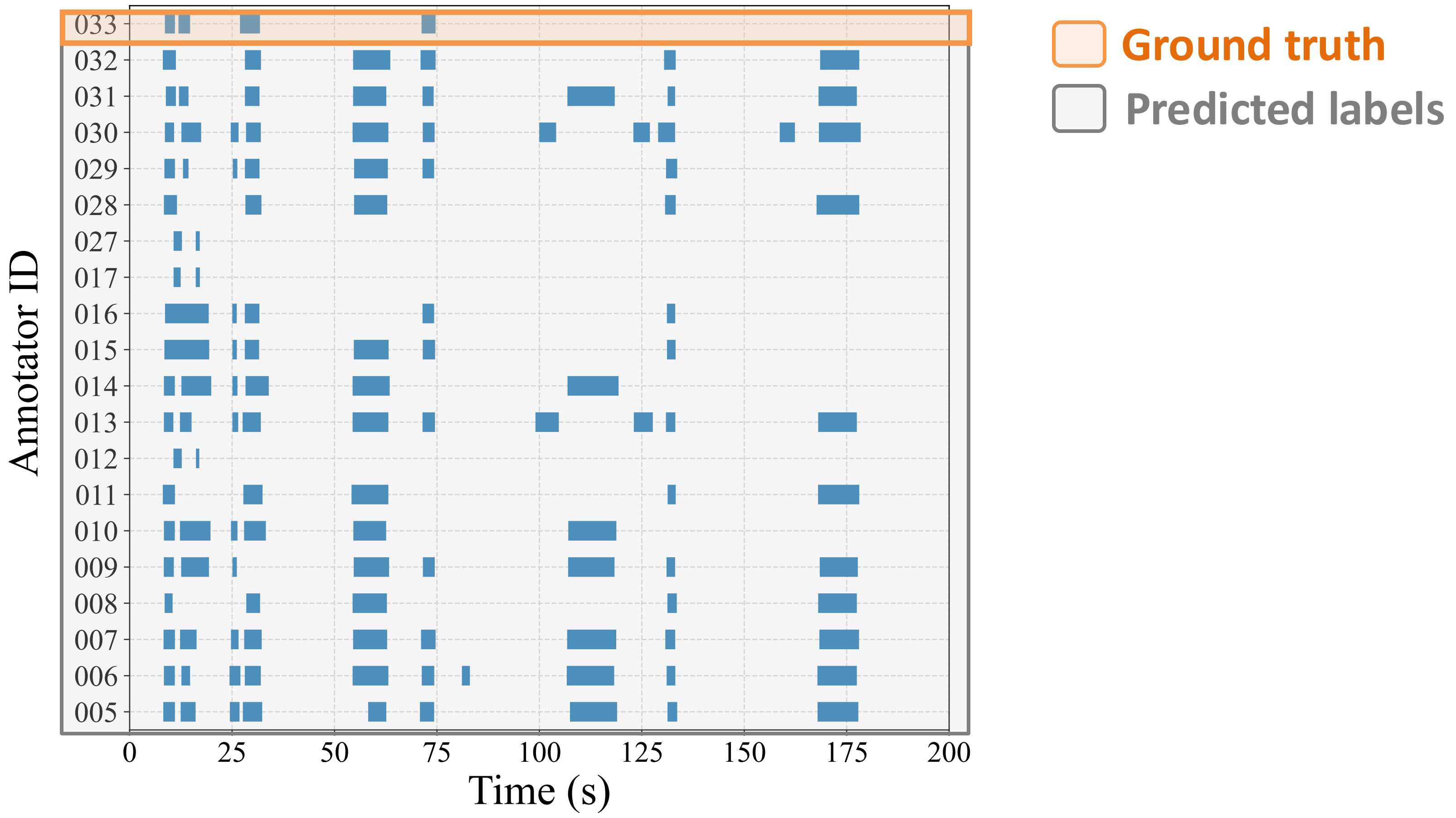
- Building the LEAD dataset
- Analyses with the LEAD dataset
- **Experiment with the LEAD dataset**

Overview of the experiment

- **Pseudo-detection performance with strong labels**
  - Purpose: To Investigate the influence of the variations in strong labels on the detection performance in SED
  - We calculated a pseudo-detection performance for each data source of the LEAD dataset.

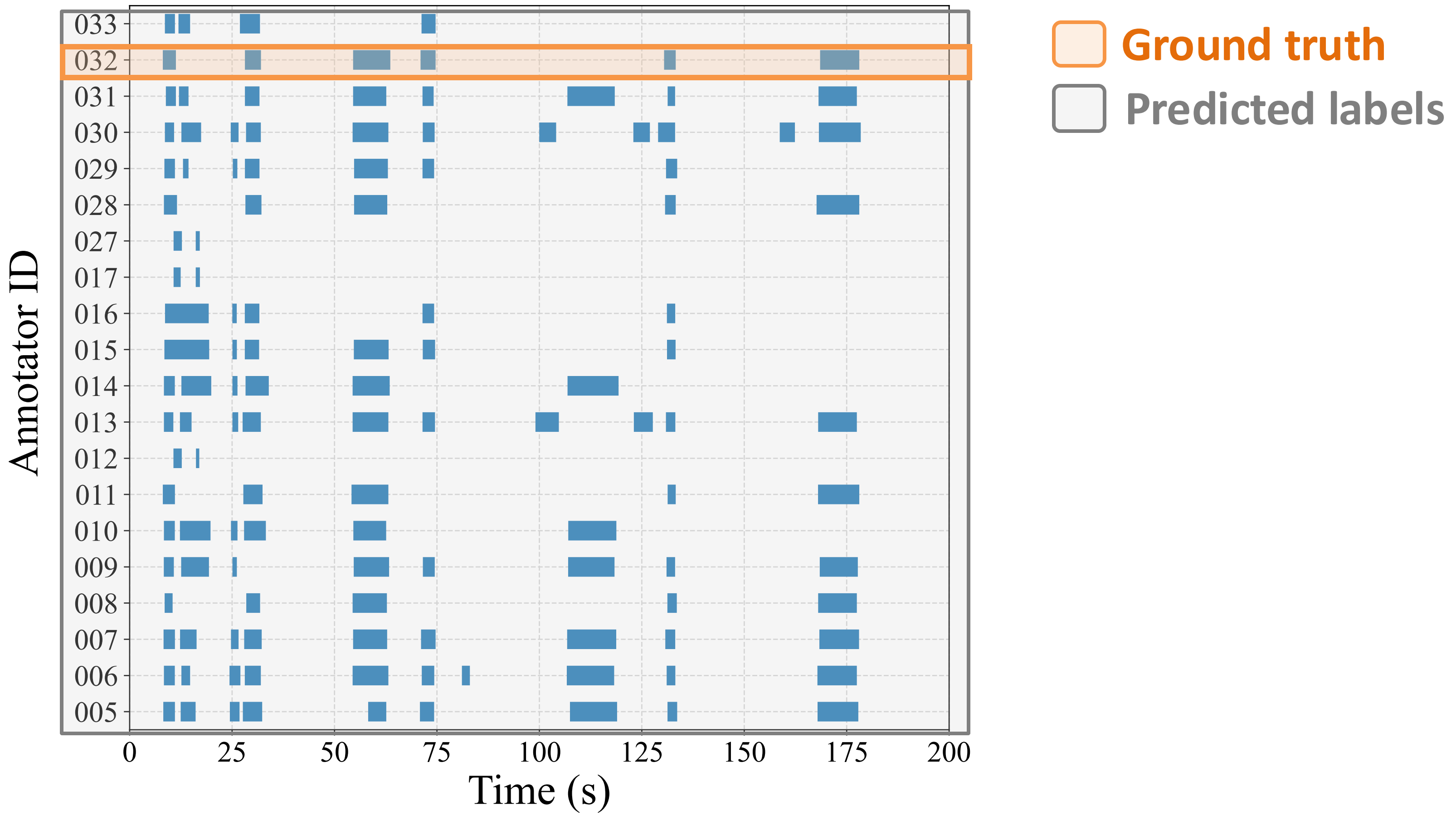
How to calculate the pseudo-metrics

- **Picking up one annotator’s strong label as the ground truth**
  - Calculating the metrics of the sound signals in each data source of the LEAD dataset



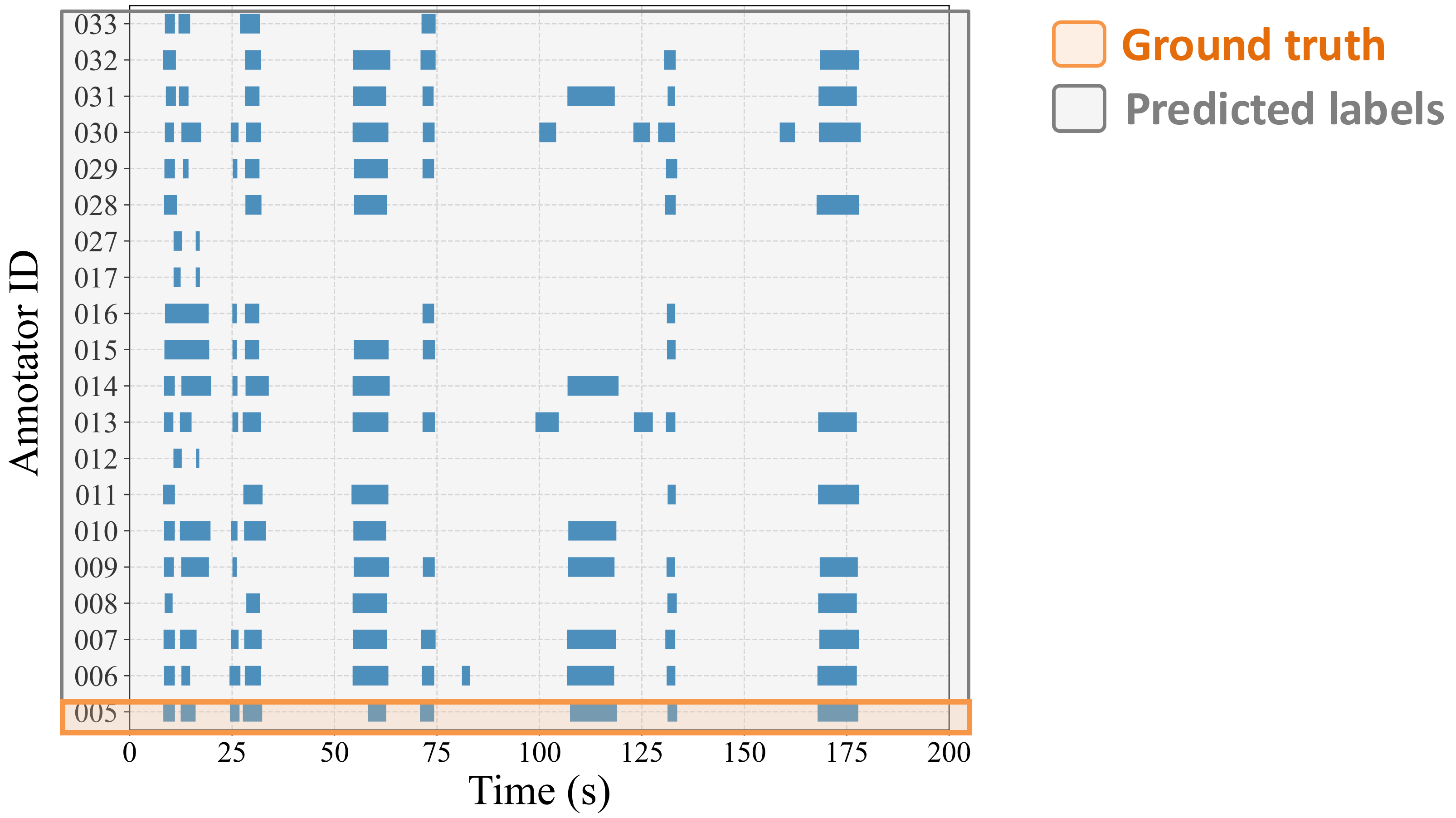
How to calculate the pseudo-metrics

- **Picking up one annotator's strong label as the ground truth**
  - Calculating the metrics of the sound signals in each data source of the LEAD dataset



How to calculate the pseudo-metrics

- **Picking up one annotator’s strong label as the ground truth**
  - Calculating the metrics of the sound signals in each data source of the LEAD dataset



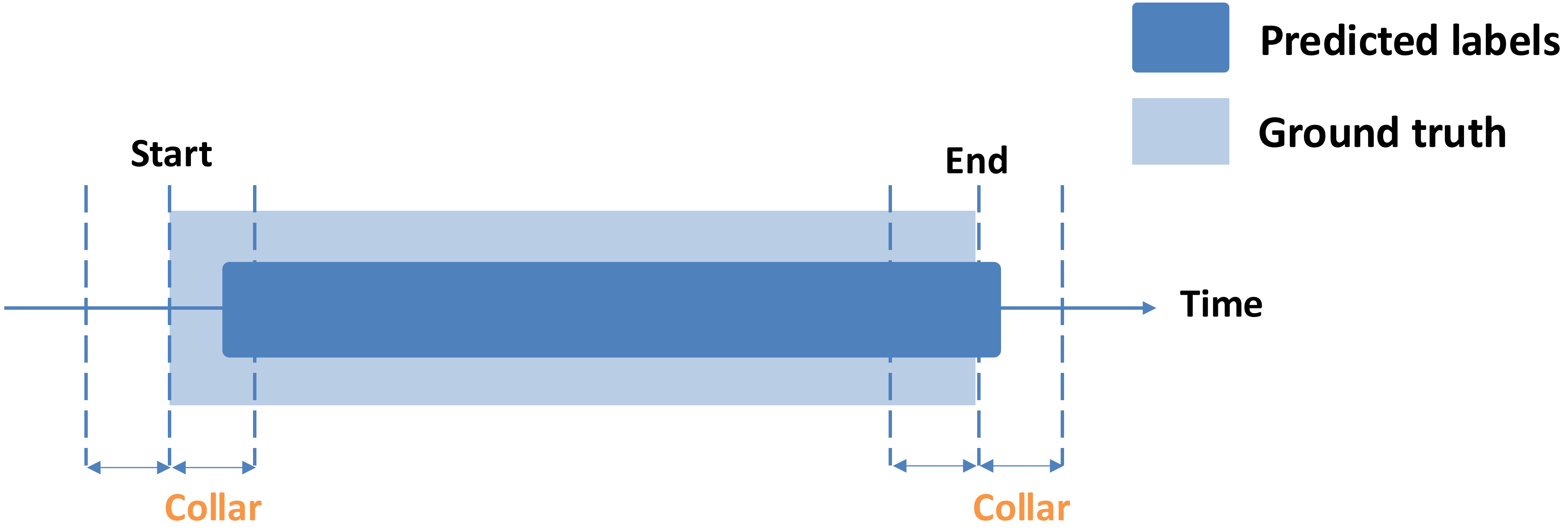
Conditions of metrics

- **Segment-based micro-F-score [1]**
- **Event-based micro-F-score [1]**
- **Intersection-based micro-F-score [2]**

[1] A. Mesaros, et al., "Metrics for polyphonic sound event detection," Applied Sciences, vol. 6, no. 6, p. 1–17, 2016.
[2] C. Bilen, et al., "A framework for the robust evaluation of sound event detection," Proc. ICASSP, pp. 61–65, 2020.

Conditions of metrics

- **Event-based micro-F-score** [1]
  - Collar: 0.20 seconds



[1] A. Mesaros, et al., “Metrics for polyphonic sound event detection,” Applied Sciences, vol. 6, no. 6, p. 1–17, 2016.
[2] C. Bilen, et al., “A framework for the robust evaluation of sound event detection,” Proc. ICASSP, pp. 61–65, 2020.

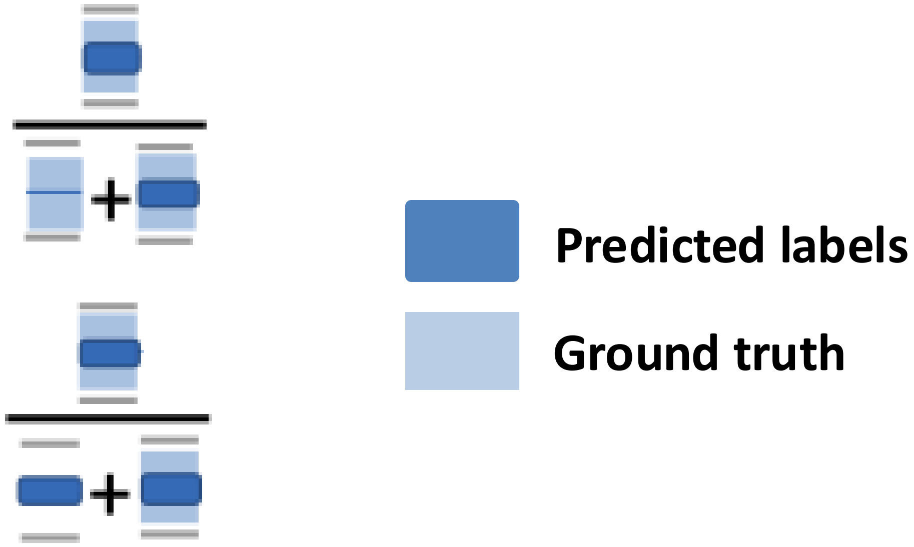
Conditions of metrics

Intersection-based micro-F-score [2]

- $\rho_{GTC} = 0.1$
  - GTC: ground truth intersection criterion
- $\rho_{DTC} = 0.1$
  - DTC: detection tolerance criterion

Ground truth intersection criteria

Detection tolerance criteria



[1] A. Mesaros, et al., “Metrics for polyphonic sound event detection,” Applied Sciences, vol. 6, no. 6, p. 1–17, 2016.
[2] C. Bilen, et al., “A framework for the robust evaluation of sound event detection,” Proc. ICASSP, pp. 61–65, 2020.

Pseudo-detection performance

- **Event-based micro-F-score got lower performance than the intersection-based micro-F-score.**

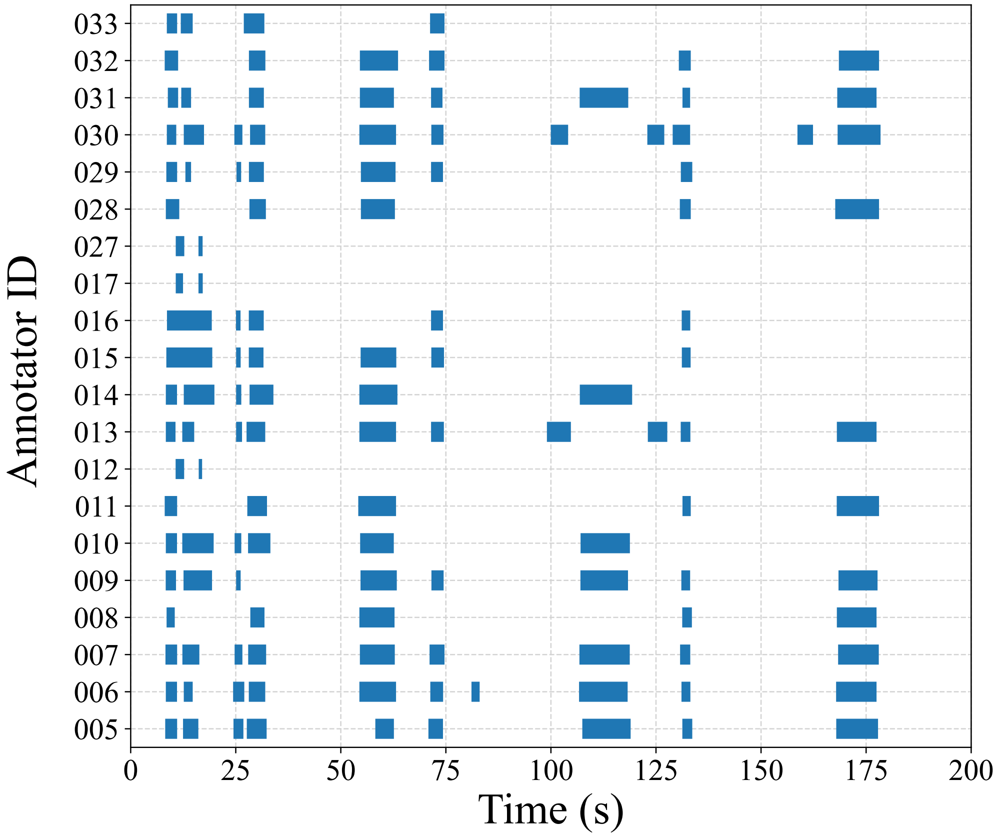
Strong label	Event-based micro-F-score	Intersection-based micro-F-score
TUT SE 2016	8.17%	53.87%
TUT SE 2017	5.33%	32.59%
TUT AS 2016	32.59%	54.25%
URBAN-SED	50.51%	40.52%

Lower

Additional experiment

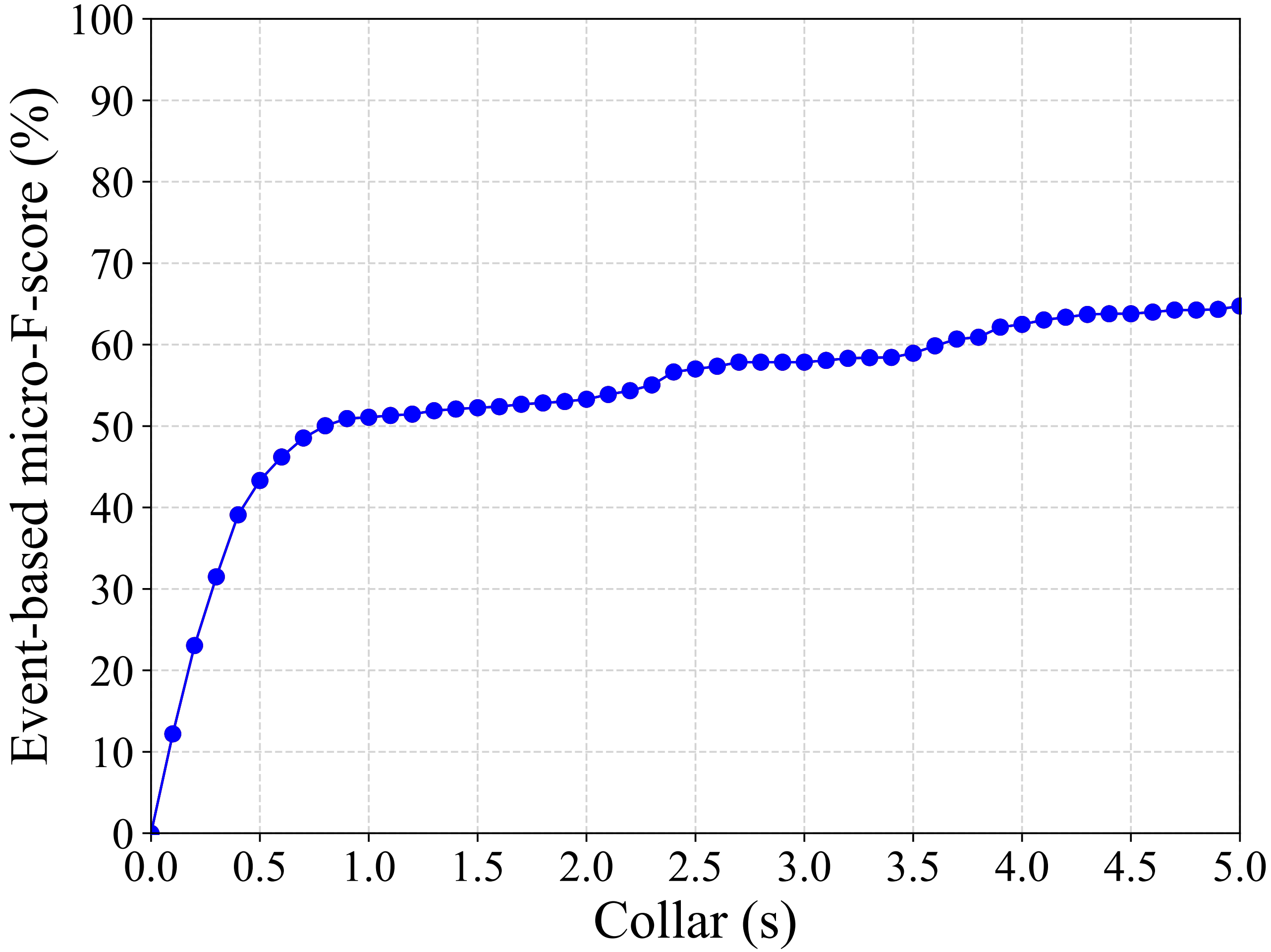
Influence of the collar setting on the detection performance

- We checked the detection performance with various collar settings of the event-based micro-F-score using “water” in one sound signal.
- The collars setting: intervals of 0.1 seconds from 0.0 to 5.0 seconds



Event-based micro-F-score

- **Event-based micro-F-score rapidly increased up to 1.0 seconds.**
  - Collar setting should be adjusted depending on the training data.



Conclusion

- **Purpose of our study**
  - To gain a better understanding of the variations in strong labels
- **Contributions**
  - Building a large-scale dataset including the variations
  - Analyses: Classification of the variations in strong labels
  - Experiment: The temporal variations in sound events affect the detection performance.
- **Future work**
  - Development a robust model against the variations in strong labels with the LEAD dataset