Investigating the Effect of Sound-Event Loudness on Crowdsourced Audio Annotations

Mark Cartwright, Justin Salamon, Ayanna Seals, Oded Nov and Juan Pablo Bello
New York University

Please address correspondence to: mark.cartwright@nyu.edu

1. Introduction

- Annotation is a time-consuming, but important step in machine listening
- Crowdsourcing is a popular way to quickly annotate, but understudied for audio
- To properly train machine listening models, we need to understand our data
- To further understand crowdsourced audio annotations and its limits, we investigate effect of sound-event loudness on:
  - Crowdsourced sound-event source annotations (i.e., start, end, class)
  - Crowdsourced sound-event proximity annotations (i.e. near or far).

2. Seeing Sound Dataset

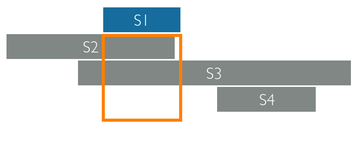
- Data from previous 3 x 3 x 2 full-factorial between subjects study:
  - 3 Audio visualizations
  - 3 Levels of max-polyphony (max number of overlapped sound events)
  - 3 Levels of gini-polyphony (concentration of sound events)
- 30 participants per condition; 10 soundscapes per condition
- Strong sound-event source annotations of 10 urban env. sound classes
- Sound-even proximity annotations (i.e. near / far)

Screenshot of Audio Annotator interface: https://github.com/CrowdCurio/audio-annotator

3. Relative Loudness Measures

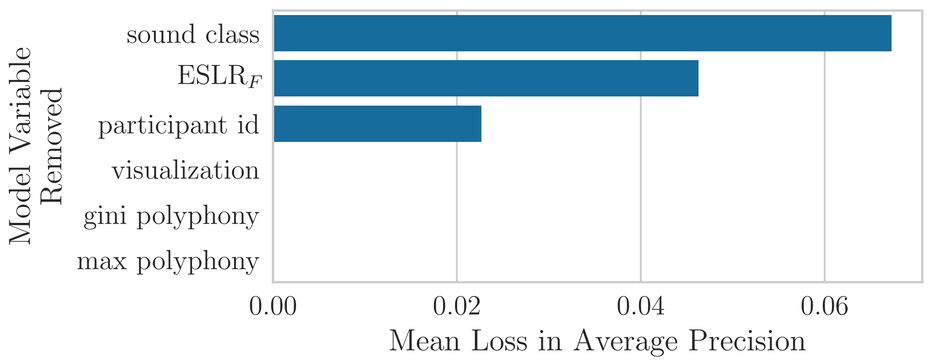
- Relative loudness based on LUFS
  - $ESLR_F$ - Sound-event loudness relative to all other sounds in scene:

  - $ESLR_L$ - Sound-event loudness relative to other co-occurring sounds:

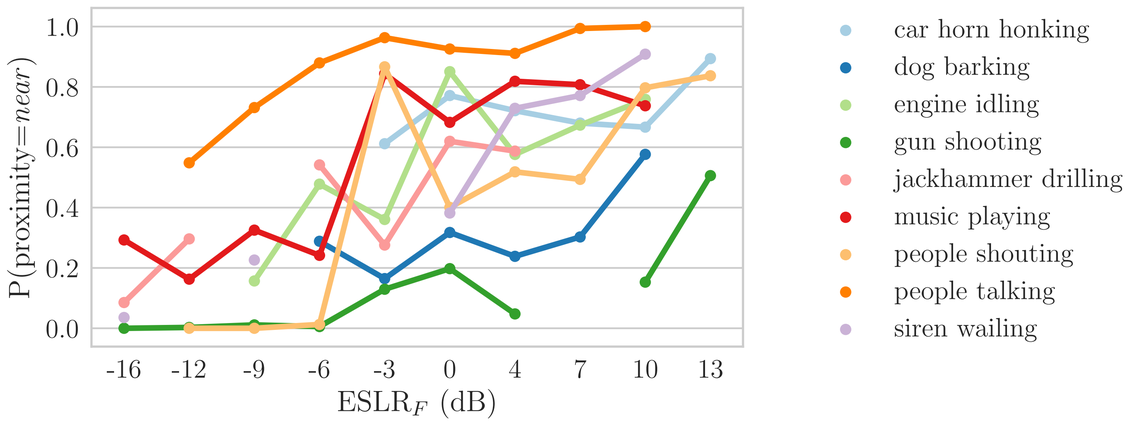


4. Effect of Loudness on Proximity Annotations

- Low annotator agreement (Krippendorff alpha = 0.31 [0.25, 0.41])
- **Logistic regression analysis, modeling *P(Proximity=near):***
  - Full model mean avg. precision = 0.80
  - Ablation analysis, removing variables from model:



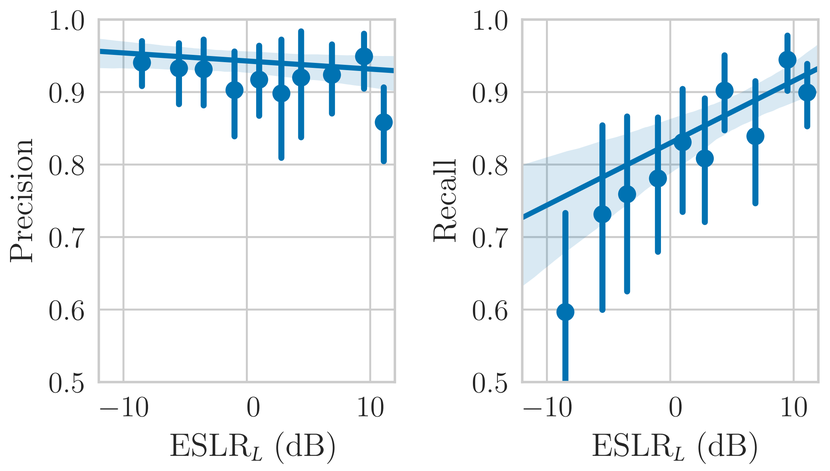
- **Relationship of sound-event loudness to proximity by class:**



- **Conclusions of sound-event proximity annotations analysis:**
  - While proximity labels are indicative of relative loudness, they are more affected by sound class and less so by annotator bias.
  - Listeners have class-dependent distance expectations. If a researcher wants to make use of proximity labels as a proxy for relative loudness, sound class must be accounted for when interpreting these labels.
  - Listeners have individualized thresholds for near and far. To account for this, annotators should complete a calibration task that estimates their thresholds.

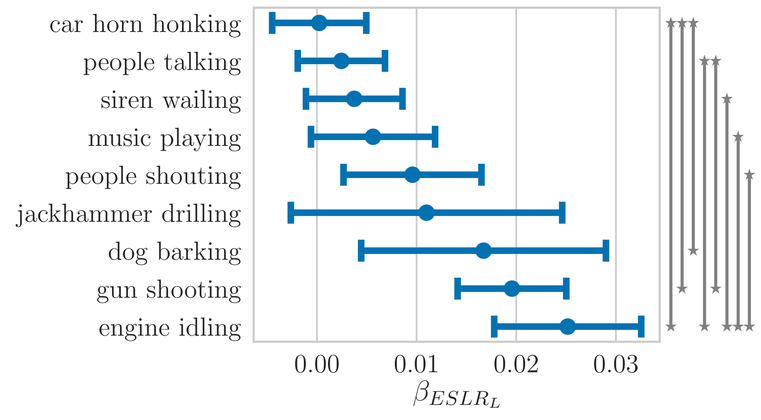
5. Effect of Loudness on Source Annotations

- Evaluation: segment-based metrics on 100 ms segments
- **Overall annotation quality**:



  - Diagonal line and shaded region in each plot is the robust linear regression fit and 95% C.I.
  - $ESLR_L$ is broken into decile bins and display only the means (dots) with 95% CIs (vertical lines).

- **Class-dependent annotation quality**:
  - Robust linear regression coefficients for recall regressed on $ESLR_L$ for each class:



  - Bars are 95% C.I.s
  - Vertical
  - lines on the right note significantly different pairs ($p < 0.05$).

- **Conclusions of sound-event source annotations analysis:**
  - Sound-event loudness affects overall event recall, but only minimally affects precision and onset/offset deviations
  - Results are largely driven by a small number of sound-event classes for which recall performance is more sensitive to relative loudness. The higher uncertainty for these classes could be accounted for in the training and evaluation of machine listening systems.

This work was partially supported by National Science Foundation award 1544753.