INVESTIGATING THE EFFECT OF SOUND-EVENT LOUDNESS ON CROWDSOURCED AUDIO ANNOTATIONS

Mark Cartwright, Justin Salamon, Ayanna Seals, Oded Nov, Juan Pablo Bello

New York University

ABSTRACT

Audio annotation is an important step in developing machine-listening systems. It is also a time consuming process, which has motivated investigators to crowdsource audio annotations. However, there are many factors that affect annotations, many of which have not been adequately investigated. In previous work, we investigated the effects of visualization aids and sound scene complexity on the quality of crowdsourced sound-event annotations. In this paper, we extend that work by investigating the effect of sound-event loudness on both sound-event source annotations and sound-event proximity annotations. We find that the sound class, loudness, and annotator bias affect how listeners annotate proximity. We also find that loudness affects recall more than precision and that the strengths of these effects are strongly influenced by the sound class. These findings are not only important for designing effective audio annotation processes, but also for effectively training and evaluating machine-listening systems.

Index Terms— crowdsourcing, audio annotations, machine listening, sound event detection

1. INTRODUCTION

Machine-listening systems are typically trained and evaluated using human-labeled audio annotations, which are time-consuming to obtain. One way to reduce that time is to parallelize the annotation process over a larger population using crowdsourcing [1, 2, 3, 4]. However, noisy "ground-truth" labels can contribute to what may seem to be limits on model performance. Therefore, to effectively use crowdsourced annotations for training and evaluating machine-listening systems, we need to better understand what factors affect annotations. For example, if we know the annotations of a particular sound class are often precise but incomplete, we should incorporate that uncertainty into training and evaluation, possibly penalizing for false-negatives more than false-positives. In previous work [4], we performed a controlled study which investigated how sound visualizations and sound scene complexity affect the quality of audio annotations for sound-event detection (SED). We know that loudness and auditory masking are complex auditory phenomena that affect what we can hear in the presence of other sounds [5, 6], but how do these manifest in annotations that we use as ground-truth for training and evaluation of models? In this paper, we use the data from our previous study to perform a post-hoc analysis in which we investigate the effect of relative loudness of sound-events on both sound-event proximity (i.e. whether a sound-event is *near* or *far*) and source annotations (i.e. the start time, end time, and sound class of a sound event).

In Section 3.1, we investigate how relative loudness and other factors such as sound class affect sound-event proximity annotations. For many tasks, sound events are more relevant if they are nearby (e.g. SED for autonomous vehicles) or loud in relation to the auditory scene (e.g. SED for noise pollution monitoring). Proximity labels could be used to inform machine-listening models about which sound events are more relevant and should be prioritized. For example, a loud, obtrusive siren could be prioritized over a distant, unobtrusive car alarm by putting more weight on the siren's loss during training. With a similar motivation, researchers have used salience, a bottom-up form of auditory attention [7, 8], to inform machine-listening models [9, 7, 10]. While partial loudness [6] and salience models exist [7, 8], neither can easily be used in SED—salience models typically output time-frequency regions without source associations, and partial loudness models [6] require separated sources as input. However, proximity labels may provide an adequate signal, and when limited to a few classes, the labeling task is lightweight and well-suited for crowdsourcing. In our analysis we seek to understand what factors affect proximity labels and if they can be used as a proxy for loudness annotations.

In Section 3.2, we investigate how relative loudness of sound-events affects both the precision and recall of crowdsourced sound-event source annotations. If we understand how loudness affects these annotations, we can make more informed decisions on how to handle potential annotation and prediction errors for audio recordings in which the relative loudness of sound events is unknown.

The analysis in this work furthers our understanding of the factors that affect crowdsourced audio annotations so we can more effectively use them in the training and evaluation of machine-listening systems.

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

2. METHODS

2.1. The Seeing Sound Dataset

The post-hoc analysis in this paper analyzes the Seeing Sound Dataset [4], which was collected in an earlier $3 \times 3 \times 2$ full-factorial between-subjects study [4]. In these 18 experimental conditions, we varied the sound visualization displayed during annotation (*waveform*, *spectrogram*, *no-visualization*), the *max-polyphony* of the soundscapes, and the *gini-polyphony* of the soundscapes. Max-polyphony and gini-polyphony are two measures of soundscape complexity. Max-polyphony describes the maximum number of overlapping sound events at any point in the soundscape. Gini-polyphony is based on the Gini coefficient [11] of polyphony over time and measures the temporal concentration of polyphony within a soundscape. We grouped our soundscapes into two levels of gini-polyphony (Gini-coeff. ranges (0.5,1] and [0,0.5]) and three levels of max-polyphony (1, 2, and 3–4 overlapping sound events), for a total of 6 complexity conditions. Using the soundscape generation tool Scaper [12], we synthesized 10 monaural soundscapes for each complexity condition, for a total of 60 soundscapes with ground-truth annotations. The soundscapes contained the following sound-event classes: *car horn honking*, *dog barking*, *engine idling*, *gun shooting*, *jackhammer drilling*, *music playing*, *people shouting*, *people talking*, *siren wailing*. We recruited 30 participants from Amazon's Mechanical Turk for each of the 18 conditions, and each participant annotated all 10 of their assigned condition's soundscapes, for a total of 5400 annotated soundscapes. For each soundscape, participants annotated the start and end times of each sound they heard, identified their classes from the list of 10 sound classes, and labeled the proximity of the sounds as *near*, *far*, or *not sure*. There were 15823 of these sound-event / proximity annotations. See [4] for more details.

2.2. Data Preparation

We want to investigate the relationship between sound-event loudness and human-labeled annotations, but sound event detection is typically evaluated for each sound class on the temporal segment level rather than the event level [13, 14]. To ensure that we are measuring the relationship between a single event's loudness and the annotation quality of that event, we filtered the dataset to only include ground-truth and hand-labeled sound-event annotations that meet the following criteria (87.4% of the data):

- There is *only one* ground-truth annotation of the sound class (e.g. "dog barking") in the soundscape.
- There is *at most one* hand-labeled annotation of the sound class in the soundscape.
- At least one of the hand-labeled annotation's boundaries are aligned within 1 s of the ground-truth.

2.3. Relative Loudness Measures

For our analysis, we defined two measures of the relative loudness of a sound event: *Full Event-to-Scene Loudness Ratio (ESLR$_F$)* and *Limited Event-to-Limited-Scene Loudness Ratio (ESLR$_L$)*. Both measures rely on Loudness Units relative to Full Scale (LUFS) [15], a standard measure of perceived loudness used in media broadcasting with log-scaled units similar to decibels. ESLR$_F$ is the difference between the LUFS of the isolated sound event and the LUFS of the the entire soundscape with the sound event removed. ESLR$_F$ captures the loudness of a sound event in relation to all other sounds in the soundscape. We use ESLR$_F$ when investigating the effect of sound-event loudness on proximity annotations, since we believe these proximity annotations are likely judged relative to all sounds in the soundscape. ESLR$_L$ is a similar difference, but rather than using the entire duration of the soundscape in the subtrahend, the duration is limited to that of the sound event in the minuend. Therefore, ESLR$_L$, captures the loudness of a sound event in relation to other simultaneously occurring sounds, but is independent of sounds that are not overlapped with it. We use ESLR$_L$ when investigating the effect of sound-event loudness on sound-event source annotations, which we believe should be independent of non-overlapping events.

3. RESULTS

3.1. Effects of Loudness on Proximity Annotations

Since we do not have ground-truth data for proximity, the goal of this analysis is to try to understand what factors affect this type of crowdsourced annotation and what signal may be contained within them. To simplify our analysis, we discarded the *not sure* proximity labels, which amounted to 4.4% of the data, in order to treat proximity as a binary variable for which $P(Proximity = near) = 0.63$. We grouped proximity labels by the 18 experimental conditions and calculated the Krippendorff's α [16] agreement between the proximity labels given by the 30 participants in each condition. The mean α was 0.31, 95% CI [0.25, 0.41] ($\alpha = 1$ is perfect agreement, and $\alpha = 0$ is random agreement). This low agreement indicates that individuals may have their own interpretations of the discrete proximity labels.

With proximity as our dependent variable, we fit a logistic regression model with independent variables of interest. The variables included *gini-polyphoy*, *max-polyphony*, and *visualization*—i.e. the independent variables in our previous study [4]. We also included *ESLR$_F$* because it is representative of the relative loudness of a sound-event which is a known cue for proximity. While there are other distance cues in auditory perception besides relative loudness (e.g. high-frequency attenuation and ratio of direct to reflected sound) [5], relative loudness alone is an informative cue for familiar sounds, especially when multiple sounds are present [5]. In addition,

sound class was included to investigate class-dependent effects, and *participant ID* was included due to low annotator agreement. The nominal variables (*sound class*, *participant id*) were added to the model using dummy variable coding.

We fit the model with 10-fold cross validation and achieved a mean average precision of 0.80. Therefore, these variables do not explain all of the behavior of the annotators, but they can give us some insight. To measure how much predictive power each variable contributes, we also fit models in which each variable was removed and measured the difference in average precision between the full model and the reduced model. Figure 1 shows that *visualization*, *gini-polyphony*, and *max-polyphony* do not contribute at all to the predictive power of the model. To investigate this further, we ran a chi-square test of independence between each of these variables and proximity. We didn't find a significant relationship for *visualization* ($p = 0.79$). We also calculated the Krippendorff's $\alpha = 0.95$ agreement between visualizations on the empirical $P(Proximity = near)$ of each sound event. From this we can conclude that the visualization does not bias proximity and has little effect on proximity labels when aggregated over the whole population. We also did not find a significant relationship between gini-polyphony ($p = 0.14$) and proximity. There was a significant relationship for max-polyphony ($p < 0.001$), but when used in conjunction with the other variables in the model, this measure did not provide more predictive power.

Figure 1 also shows that while *ESLR$_F$* and *participant id* are both predictive of proximity, *sound class* is much more predictive. Figure 2 explores this relationship, showing that while there is a clear relation between ESLR$_F$ and *near* proximity labels, sound class introduces a strong bias. This can likely be explained by class-dependent distance expectations—we expect *people talking* to occur very close to us and *gun shooting* to occur very far at the same loudness level. In addition, our recordings of *gun shooting* have long reverberation tails which could bias their perceived distance. These results suggest that 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. Lastly, the mean gain in average precision for *participant id* suggests that listeners have individualized thresholds for *near* and *far*. To account for this, annotators should complete a calibration task that estimates their thresholds.

3.2. Effects of Loudness on Source Annotations

3.2.1. Overall annotation quality

We evaluate overall quality of sound-event source annotations with a segment-based method typically used for SED evaluation [13, 14]. This method breaks annotations into non-overlapping 100 ms segments for each sound class. A segment is marked active if it overlaps with the time interval of an annotation for that class. True positives, false positives,

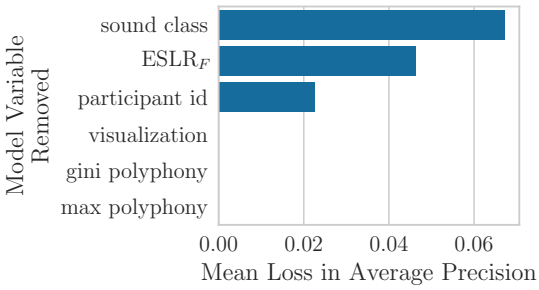

Fig. 1. The mean loss in average precision when a variable is removed from the full logistic regression model predicting $P(Proximity = near)$

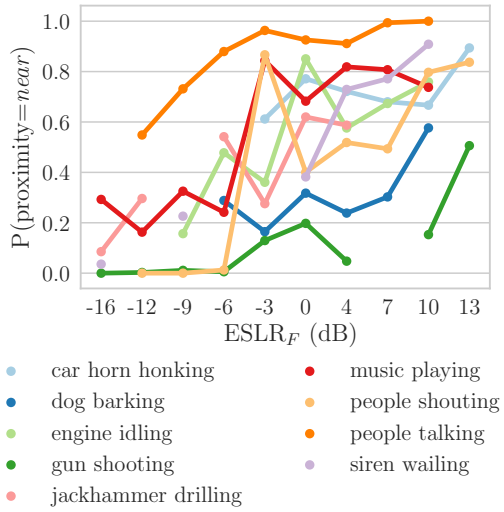

Fig. 2. ESLR$_F$ vs. $P(Proximity = near)$ by sound class. For visual clarity, the data points have been grouped into equally spaced bins. The dots represent the mean of a bin.

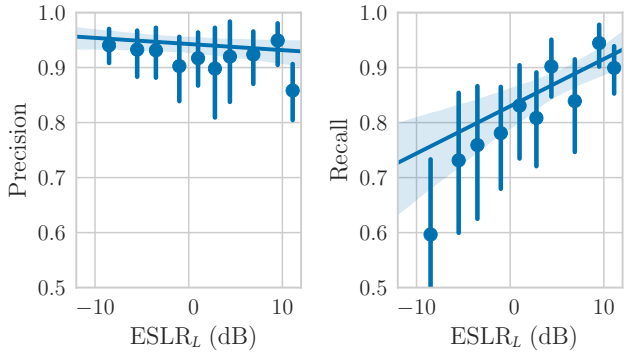

Fig. 3. ESLR$_L$ vs precision (left) and recall (right) of sound-event source annotations. The diagonal line and shaded region in each plot is the robust linear regression fit and 95% CI. For visual clarity, we group the data into ESLR$_L$ decile bins and display only the means (dots) with 95% CIs (vertical lines).

and false negatives are then calculated independently in each segment for each class and aggregated over all users to calculate precision and recall. For easy comparison to our results in [4], we limit our analysis to the *spectrogram* conditions.

Figure 3 shows the relationship of ESLR_L vs. precision and recall. There is a clear, significant correlation between ESLR_L and recall (Spearman $\rho = 0.36$, $p < 0.001$) but not between ESLR_L and precision (Spearman $\rho = 0.11$, $p = 0.13$)—i.e. as the ESLR_L of a sound-event is increased, false negative errors decrease but false positives remain stable. This implies that annotation of sound events will be precise but possibly incomplete. We found a similar result in a previous study in relation to sound scene complexity—max-polyphony had a negative correlation with recall but an insignificant correlation with precision [4]. To relate these two results, we also looked at the relation of ESLR_L to recall at each max-polyphony level. We computed the robust linear regression coefficients for recall regressed on ESLR_L at each max-polyphony level to measure the strength of effect of ESLR_L. The regression coefficients for ESLR_L at max-polyphony levels 0, 1, and 2 were respectively 0.0022, 0.0063, and 0.0077—the effect of ESLR_L on recall increased with max-polyphony. However, when we ran an ANOVA to test the difference of coefficients, we could not reject the null hypothesis of equal coefficients ($p = 0.14$).

We also calculated the Pearson correlation between both onset and offset deviations (i.e. difference between ground-truth and human-labeled times) and ESLR_L. While we found statistically significant correlations, they were weak (onset-deviation Spearman $\rho = 0.19$, $p = 0.007$; offset-deviation Spearman $\rho = -0.17$, $p = 0.017$). When examining the data, we found a trend to slightly overestimate onsets and offsets when sound events are loud and slightly underestimate onsets and offsets when sound events are soft. However, the weakness of these correlations implies that loudness only mildly affects the annotation timing, and the minimal effect on precision and the strong effect on recall is likely caused by missing entire sound events that are soft or masked by other sounds.

3.2.2. *Class-dependent annotation quality*

Finally, we investigate whether the effects on recall are class-dependent. We fit a robust linear regression model with recall regressed on ESLR_L, with a different line fit to each sound class (see Fig. 4). We ran an ANOVA to test the differences of these coefficients and rejected the null hypothesis that all the coefficients were equal ($p < 0.001$). However, after p-value adjustment only a few pairs are different at $p < 0.05$ (noted in Fig. 4). Figure 4 shows loudness affects some classes' annotations (e.g. *engine idling* and *gun shooting*) more than others (e.g. *car horn honking*, *siren wailing*, *people talking*). An explanation could be that some of the unaffected sound classes are designed to catch our attention (*car horn honking* and *siren wailing*). In contrast, *engine idling* is a noisy sound

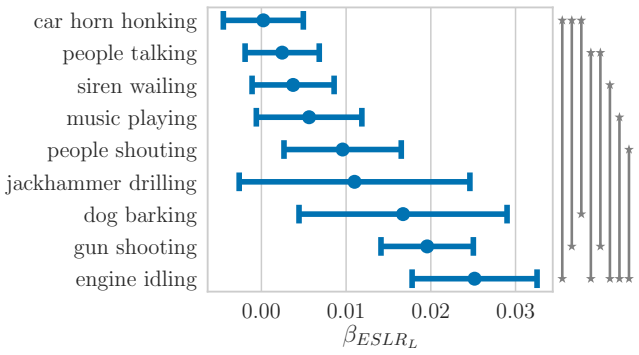

Fig. 4. The robust linear regression coefficients for recall regressed on ESLR_L independently for each sound class. The horizontal lines represent the coefficients' 95% CIs. Vertical lines on the right note significantly different pairs ($p < 0.05$).

that may easily blend into the background unless sufficiently loud, and *gun shooting* has a long reverberation tail not easily perceptible at low sound levels or in the presence of other sound-events. Further experiments are required to investigate and validate the underlying causes of these effects. Nevertheless, the effect of loudness on human sound-event detection is clearly class-dependent—some classes may have more uncertainty in their annotations. This uncertainty could be accounted for when training (e.g. using a Bayesian framework) and when evaluating machine-listening systems (e.g. down-weighting false negatives for some classes). This may be particularly important if trying to match human perception using synthesized data and annotations [13, 12].

4. CONCLUSION

In this work, we performed a post-hoc analysis on crowdsourced audio annotations to investigate the effect of sound-event loudness on sound-event source and proximity annotations. We found that proximity labels are more affected by sound class than by loudness, and less so by annotator bias. These results show that while proximity labels are indicative of relative loudness, care must be taken to account for these other factors. The utility of these labels for training and evaluating machine listening models will be determined in future experiments. Additionally, we found that sound-event loudness affects overall event recall, but only minimally affects precision and onset/offset deviations. Furthermore, these 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 can be accounted for in the training and evaluation of machine listening systems. Overall, the results from this post-hoc analysis further our understanding of crowdsourced audio annotations and enable researchers to more effectively collect audio annotations and build machine-listening systems.

5. REFERENCES

[1] Lior Shamir, Carol Yerby, Robert Simpson, Alexander M von Benda-Beckmann, Peter Tyack, Filipa Samarra, Patrick Miller, and John Wallin, "Classification of large acoustic datasets using machine learning and crowdsourcing: Application to whale calls," *The Journal of the Acoustical Society of America*, vol. 135, no. 2, pp. 953–962, 2014.

[2] Anthony Truskinger, Haofan Yang, Jason Wimmer, Jinglan Zhang, Ian Williamson, and Paul Roe, "Large scale participatory acoustic sensor data analysis: tools and reputation models to enhance effectiveness," in *Proc. of the IEEE International Conference on E-Science*. 2011, pp. 150–157, IEEE.

[3] Mark Cartwright and Bryan Pardo, "Vocalsketch: Vocally imitating audio concepts," in *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*. 2015, pp. 43–46, ACM.

[4] M. Cartwright, A. Seals, J. Salamon, A. Williams, S. Mikloska, D. MacConnell, E. Law, J.P. Bello, and O. Nov, "Seeing sound: Investigating the effects of visualizations and complexity on crowdsourced audio annotations," *Proceedings of the ACM on Human-Computer Interaction*, vol. 1, no. 2, 2017.

[5] Brian CJ Moore, *An introduction to the psychology of hearing*, Brill, 2012.

[6] B.R. Glasberg and B.C.J. Moore, "A model of loudness applicable to time-varying sounds," *Journal of the Audio Engineering Society*, vol. 50, no. 5, pp. 331–342, 2002.

[7] Emine Merve Kaya and Mounya Elhilali, "Modelling auditory attention," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 372, no. 1714, 2017.

[8] Christoph Kayser, Christopher I Petkov, Michael Lippert, and Nikos K Logothetis, "Mechanisms for allocating auditory attention: an auditory saliency map," *Current Biology*, vol. 15, no. 21, pp. 1943–1947, 2005.

[9] Ozlem Kalinli, Shiva Sundaram, and Shrikanth Narayanan, "Saliency-driven unstructured acoustic scene classification using latent perceptual indexing," in *Multimedia Signal Processing, 2009. MMSP'09. IEEE International Workshop on*. 2009, pp. 1–6, IEEE.

[10] Malcolm Slaney, Trevor Agus, Shih-Chii Liu, Merve Kaya, and Mounya Elhilali, "A model of attention-driven scene analysis," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*. 2012, pp. 145–148, IEEE.

[11] Shlomo Yitzhaki and Edna Schechtman, *The Gini Methodology: A primer on a statistical methodology*, vol. 272, Springer Science & Business Media, 2012.

[12] Justin Salamon, Duncan MacConnell, Mark Cartwright, Peter Li, and Juan Pablo Bello, "Scaper: A library for soundscape synthesis and augmentation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 2017.

[13] Tuomas Virtanen, Annamaria Mesaros, Toni Heittola, and Aleksandr Diment, "IEEE AASP challenge: Detection and classification of acoustic scenes and events," 2017, http://www.cs.tut.fi/sgn/arg/dcase2017/.

[14] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, pp. 162, 2016.

[15] EM Grimm, R Van Everdingen, and MJLC Schpping, "Toward a recommendation for a european standard of peak and lkfs loudness levels," *SMPTE motion imaging journal*, vol. 119, no. 3, pp. 28–34, 2010.

[16] Andrew F Hayes and Klaus Krippendorff, "Answering the call for a standard reliability measure for coding data," *Communication methods and measures*, vol. 1, no. 1, pp. 77–89, 2007.