Seeing Sound: Investigating the Effects of Visualizations and Complexity on Crowdsourced Audio Annotations

MARK CARTWRIGHT, New York University, USA
AYANNA SEALS, New York University, USA
JUSTIN SALAMON, New York University, USA
ALEX WILLIAMS, University of Waterloo, Canada
STEFANIE MIKLOSKA, University of Waterloo, Canada
DUNCAN MACCONNELL, New York University, USA
EDITH LAW, University of Waterloo, Canada
JUAN P. BELLO, New York University, USA
ODED NOV, New York University, USA

Audio annotation is key to developing machine-listening systems; yet, effective ways to accurately and rapidly obtain crowdsourced audio annotations is understudied. In this work, we seek to quantify the reliability/redundancy trade-off in crowdsourced soundscape annotation, investigate how visualizations affect accuracy and efficiency, and characterize how performance varies as a function of audio characteristics. Using a controlled experiment, we varied sound visualizations and the complexity of soundscapes presented to human annotators. Results show that more complex audio scenes result in lower annotator agreement, and spectrogram visualizations are superior in producing higher quality annotations at lower cost of time and human labor. We also found recall is more affected than precision by soundscape complexity, and mistakes can be often attributed to certain sound event characteristics. These findings have implications not only for how we should design annotation tasks and interfaces for audio data, but also how we train and evaluate machine-listening systems.

CCS Concepts: • **Human-centered computing** → **Computer supported cooperative work**; **Empirical studies in collaborative and social computing**; **Empirical studies in visualization**; • **Information systems** → *Speech / audio search*; • **Applied computing** → *Sound and music computing*;

Additional Key Words and Phrases: Sound Event Detection; Annotation

ACM Reference Format:
Mark Cartwright, Ayanna Seals, Justin Salamon, Alex Williams, Stefanie Mikloska, Duncan MacConnell, Edith Law, Juan P. Bello, and Oded Nov. 2017. Seeing Sound: Investigating the Effects of Visualizations and Complexity on Crowdsourced Audio Annotations. *Proc. ACM Hum.-Comput. Interact.* 1, 2, Article 29 (November 2017), 21 pages. https://doi.org/10.1145/3134664

Authors' addresses: Mark Cartwright, New York University, New York, USA, mark.cartwright@nyu.edu; Ayanna Seals, New York University, New York, USA, ayannaseals@nyu.edu; Justin Salamon, New York University, New York, USA, justin.salamon@nyu.edu; Alex Williams, University of Waterloo, Waterloo, Canada, alex.williams@uwaterloo.ca; Stefanie Mikloska, University of Waterloo, Waterloo, Canada, smiklosk@uwaterloo.ca; Duncan MacConnell, New York University, New York, USA, dom228@nyu.edu; Edith Law, University of Waterloo, Waterloo, Canada, edith.law@uwaterloo.ca; Juan P. Bello, New York University, New York, USA, jpbello@nyu.edu; Oded Nov, New York University, New York, USA, onov@nyu.edu.

2573-0142/2017/11-ART29
https://doi.org/10.1145/3134664

1 INTRODUCTION

Crowdsourcing is a powerful tool for developing large training and evaluation datasets for machine learning in many domains, most markedly computer vision [10, 20, 28, 35, 37]. Many of the advances in the development of computer vision algorithms can be attributed to training powerful machine learning models on large image datasets such as ImageNet [10], which contains millions of images annotated through crowdsourcing.

While crowdsourcing has also been used to gather data for learning problems in the non-speech audio domain [5, 6, 23, 25, 26, 39, 41, 46], machine listening—the sibling field of computer vision—has yet to see the same transformative success. This gap is in a large part due to the discrepancy between the amount of training data available for visual and audio tasks. For example, many image recognition systems are trained using ImageNet [10] which currently contains over 21,000 object categories (i.e., synsets) and 14 million weakly-annotated labels, indicating the presence of objects, and roughly 450,000 strongly-annotated labels, indicating the presence and location (e.g., bounding box) of objects [42]. And state-of-the-art speech-recognition systems are trained on thousands of hours of transcribed speech data [52]. In contrast, the largest strongly-labeled datasets for machine listening contain approximately 10,000 events in less than 50 hours of audio [13, 31]. In recent months, researchers have made progress on this problem by generating large numbers of weak labels [3, 15] with minimal human oversight. However, it is unclear how well algorithms trained on weakly-labeled data will perform in sound event detection (SED) tasks, where the goal is to predict not only the presence but also the start (onset) and end (offset) times of each sound event. Furthermore, a large, strongly-labeled dataset is essential for evaluation.

The problem of data scarcity is compounded by the fact that there has been little research into the design of effective audio-annotation tools compared to other domains [8, 11, 19, 21, 22, 40, 48, 49]. Little is known about the impact of various design choices on the quality and speed of annotations by non-experts. The design principles that work in other domains may not apply to audio annotation tasks. For example, the temporal nature of audio tasks mean that annotators cannot simply listen and label with a “glance-and-click" approach as in image annotation. While video annotation also shares this temporal hindrance [19, 22, 40, 48], the challenge is not nearly as severe as videos can often be described by sampling a small number of frames, making the process akin to image annotation.

In addition, there has been little research into how the complexity and sound class composition of acoustic scenes affect annotation performance, an issue that, despite receiving virtually no attention in the literature, can potentially impact how we train and evaluate machine listening algorithms.

In this paper, we examine two factors—sound visualization (e.g., such as waveform and spectrogram) and acoustic characteristics (e.g., type and density of sound events)—and how they affect the abilities of novices to perform audio annotation tasks accurately and efficiently. In particular, we focus on sound event detection tasks, for which most algorithms require their training data in the form of “strong" annotations of sound events that specify not only the class labels (e.g., “dog barking") but also the locations in time (i.e., onset and offset times) of the sound events. Our experiments address the following questions:

(1) What is the trade-off between reliability and redundancy in crowdsourced audio annotation?
(2) Which sound visualization aid yields the highest quality annotations with and without taking into account time constraints?
(3) What limitations can we expect from crowdsourced audio annotations as a function of soundscape complexity and sound event class?

The answers to these questions provide insights into how one should design interfaces for crowdsourcing audio annotation tasks. In the process of answering these questions, we also raise new

questions regarding the definition of ground-truth annotations and the sufficiency of a CSCW-based approach which aggregates audio annotations from a population of users. Finally, our contribution also include *Scaper*[38], an open-source tool for synthesizing soundscapes and *CrowdCurio Audio-Annotator*, a open-source, web-based tool for crowdsourced audio annotation.

2 RELATED WORK

Sound visualizations in the form of spectrograms are very common in the expert annotation of bioacoustic recordings [9, 36, 44, 47]. Often expert birdcall annotators do not even listen to recordings. Instead, they learn to recognize specific visual patterns in spectrograms, allowing them to annotate 24 hours of audio for one class of events (e.g., calls of a specific bird species) in about one hour [44]. While impressive, it is unclear how such techniques scale to multi-label (i.e. multiple, potentially overlapping sound events) annotation; they may require a separate pass for each class.

Sound visualizations have also been used in crowdsourced annotation with novices. In Whale.fm [39], short recordings of whale vocalizations were presented with spectrogram visualizations, and citizen scientist volunteers matched query whale vocalization to one of 36 candidate vocalizations. However, participants' use of the spectrogram was not evaluated, and it is unclear how this design decision affected the resulting annotations.

There have been two studies that have investigated the use of spectrograms to hasten audio annotation. Lin et al. [27] developed a saliency-maximized audio spectrogram to enable fast detection of sound events by human annotators. They then conducted a study on the effect of this alternative representation on audio annotation quality. In a within-subjects study, 12 annotators unfamiliar with spectrograms labeled two 80-minute files using saliency-maximized spectrograms and two using standard spectrograms. The investigators found a significant increase in F-score with the saliency-maximized spectrograms. Unfortunately, speed and quality were confounded in their analysis, and the investigators did not specify the sound classes of the events.

Truskinger et al. [45] investigated the ability of novices to rapidly visually scan spectrograms for bioacoustic recordings similar to the process of experts [44]. In their study, they summarized 24-second soundscapes in the form of spectrogram flashcards. They presented the flashcards in constant intervals of either 1, 2, or 5 seconds, and for each flashcard, participants indicated the presence of the target sound class (koala vocalizations). In other words, it was a rapid, binary, weak-labeling task and is similar to a later image annotation study [20] which tested even shorter intervals. Truskinger et al. found annotators had equivalent accuracies at both 1 s and 5 s intervals, but accuracy decreased at 1 s intervals. This is a very encouraging study that shows that novices can learn the visual patterns of some audio classes. However, the Koala vocalizations were chosen for their visual saliency and ease of detection, so it is unclear if this approach can generalize to other audio classes.

Lastly, there have also been efforts to speed up the audio annotation using semi-automated approaches [16, 18]. For example, Kim and Pardo [18] developed a human-in-the-loop approach which sped up multi-label annotation two-fold.

In relation to these studies, our work takes a step back and questions assumptions that all of these studies have made, asking at a more fundamental level whether waveforms, spectrograms, or no visualization yields higher quality and faster annotations. We address this primary question in a way that seeks greater generalization by investigating how sound classes and soundscape complexity affect annotation performance. We also seek to understand the redundancy/reliability trade-off in this crowdsourcing task, an important practical issue that has yet to be investigated. Once we have greater fundamental understanding of this problem, we can then choose whether to apply the additional optimizations investigated in these related works in an informed way.

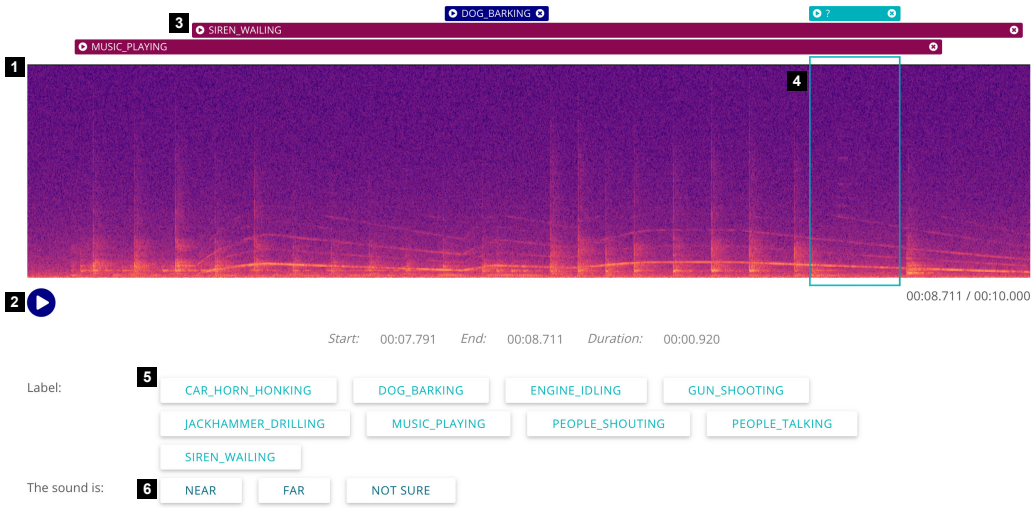


Fig. 1. A screenshot of the Audio-Annotator. Note that the numbered black squares are not part of the interface; rather, they are labels for reference in Section 3.1.

3 TECHNICAL SETUP

In this section, we describe the tools and user interfaces used in our study.

3.1 Audio-Annotator: A Web-based Tool for Audio Annotation

To run our experiment, we needed a web-based audio-annotation tool designed for citizen scientists that met the following design criteria:

- Support for configurable sound visualizations including *waveform*, *spectrogram*, and a blank *no-visualization* control
- Support for multiple, overlapping time-interval audio-event annotations (i.e., strong labels)
- Annotations that do not impede the sound visualization yet clearly display onset and offset times
- An intuitive design for specifying onset/offset times and sound classes
- Compatible with multiple backends for flexibility
- Open source to easily change the design for different tasks as needed

Existing web-based audio-annotation tools [12, 18, 36] did not meet these criteria. Therefore, we developed an open-source audio-annotation tool by extending the popular WaveSurfer.js [17] audio playback and visualization tool to meet our needs. This new tool, which we call the *Audio-Annotator*, is illustrated in Fig. 1 and is built using JavaScript, HTML5, Web-Audio API, and the Materialize CSS framework. The large, colorful, rectangular visualization (Label 1 in Fig. 1) in the middle of the screenshot is the configurable sound visualization. The large scale and prominent placement of the visualization was chosen to assist users in the cognitive process of detecting and identifying sounds. In this example, we have configured it to display a spectrogram visualization (see Section 4.1 for descriptions of this visualization and the others used in the study). For all visualizations, the horizontal axis represents time and the playback cursor moves from left to right. Users can play the audio recording with either the play button (Label 2 in Fig. 1) below the bottom left corner of the visualization or by pressing the space bar. The playback cursor allows one to relate the audio to the visual cues. On the rectangular space occupied by the visualization, users create and

edit their annotation region using a similar interaction paradigm as used in most Digital Audio Workstations (DAWs) (e.g., GarageBand [1], ProTools [2], etc.). This includes the following scenarios and interactions:

- *Seeking to a specific time:* Click the desired location in the x-dimension of the visualization. When combined with space-bar-triggered playback, this interaction enables users to quickly interact with the visualization while listening to set the onset and offset with fluid precision techniques.
- *Creating an annotation:* Click and drag within the visualization rectangle to identify the region in which a sound event occurs. This creates a corresponding annotation region bar (Label 3 in Fig. 1) above the visualization.
- *Deleting an annotation:* Click on the "x" within the annotation's region bar above the visualization.
- *Playing an annotation:* Click on the play button that appears in the annotation's region bar. This button only plays that specific annotation region
- *Selecting an annotation:* Double-click on the annotation region bar. This displays the annotation region's bounding box (Label 4 in Fig. 1) on the visualization and enables the region to be moved or resized.
- *Moving an annotation:* Click on the region bar of a selected annotation and drag it to the desired temporal location.
- *Resizing an annotation:* Click on the left or right edge of a selected region's bounding box and drag the edge to the desired location. This is used to change the onset or offset of the annotation, refining the annotation timing.

After a region is created or selected, it can be assigned a class. These classes are displayed in the grid of buttons beneath the sound visualization. In Fig. 1, this grid is configured to contain both sound event classes (e.g., *jackhammer*; Label 5 in Fig. 1) and proximity classes (e.g., *near*; Label 6 in Fig. 1). Once a sound event class is chosen, it will appear in the corresponding annotation region bar above the visualization. Once a proximity class is chosen, the coloring of the annotation region bar is changed to correspond to the selected proximity class (*red*=near, *blue*=far, *purple*=not sure, *grey*=not yet chosen).

The Audio-Annotator is open-source and available for download at https://github.com/CrowdCurio/audio-annotator.

3.2 CrowdCurio: A Research-based Crowdsourcing Platform

To host the Audio-Annotator and manage the experiment, we incorporated our interface into CrowdCurio[1], a research-oriented crowdsourcing platform [24, 50]. CrowdCurio is powered by two key components: a Project Manager and an Experiment Manager. The Project Manager enables researchers with little technical expertise to create crowdsourcing projects with a set of interfaces. Similarly, the Experiment Manager provides researchers with the means to run A/B tests on projects, coordinate the assignment of subjects to experimental conditions, manage logistical tasks typically associated with online experiments (e.g., the completion of consent forms), and incorporate data collection instruments (e.g., surveys and questionnaires). Additionally, this component enables researchers to continually monitor and evaluate the quality of the data emerging from their crowdsourcing projects. The platform maintains a RESTful API, and both components are interfaced with a Python API client that facilitates the management of projects, experiments, and crowdsourced data.

[1] https://www.crowdcurio.com

3.3 Scaper: A Soundscape Synthesizer

To reliably evaluate the influence of different interventions on the quality of human annotations, we require highly controlled and perfectly annotated audio stimuli (soundscapes). First, we require labels with precise start (onset) and end (offset) times for each sound event. Second, to evaluate the influence of a soundscape's acoustic characteristics on its annotation quality, we need to control the types of sound events that occur in a soundscape and the degree to which they overlap. To achieve this, we generated the soundscapes using Scaper[2] [38], an open-source library we developed to probabilistically synthesize soundscapes from a collection of isolated sound-event recordings. Scaper provides high-level, probabilistic control over the number and types of sound events, their start times, durations, and loudness with respect to a "background" track. The library outputs the synthesized soundscape audio along with an annotation file.

From our own experience annotating soundscapes, we noted that it was harder to label sound events accurately when a soundscape was *complex*, i.e. when it contained many, potentially overlapping, sound events. To evaluate this, we defined two complementary measures of soundscape complexity: *max-polyphony* and *gini-polyphony*, where "polyphony" refers to the number of overlapping sound events at any given moment in time (excluding the background). *Max-polyphony* is the largest polyphony observed in a soundscape. Our hypothesis is that when sound events overlap it becomes harder to identify their sound class and harder to identify their precise onsets and offsets, and so we expect a high polyphony to have a negative effect on annotation quality. *Gini-polyphony* is intended to measure how "concentrated" a soundscape's polyphony is—that is, whether the polyphony is evenly distributed over time (e.g., idling engine and alarm sounds that span the entire soundscape) or whether it is concentrated in a few points in time (e.g., the many sounds of a traffic accident on a typically quiet street). To quantify this, we compute the sound event polyphony at small, fixed, time intervals throughout the soundscape. With these polyphony values, we calculate the Gini coefficient, a statistical measure of the degree of inequality represented in a set of values [53]. The coefficient ranges between 0–1, with zero representing maximal equality (low soundscape complexity) and one representing maximal inequality (high soundscape complexity). After synthesizing a soundscape, Scaper automatically computes its max-polyphony and gini-polyphony, storing it in the annotation file.

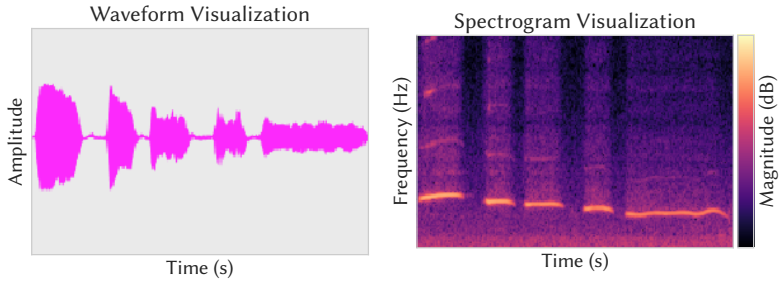
4 METHODS

Our study followed a $3 \times 3 \times 2$ between-subjects factorial experimental design in which we varied the following factors:

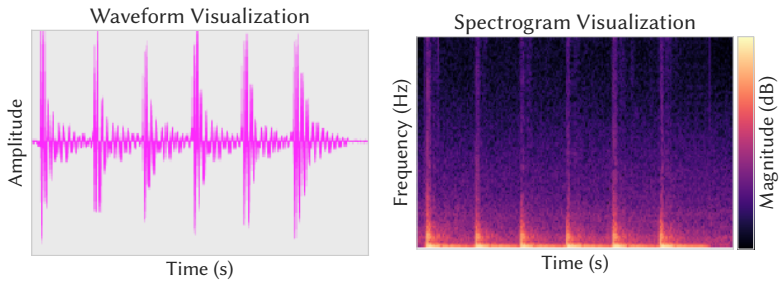
(1) **visualization**: the sound visualization (*no visualization* (the control), *waveform*, or *spectrogram*)
(2) **max-polyphony**: the first measure of the soundscape complexity, binned into three levels (described in Section 4.2)
(3) **gini-polyphony**: the second measure of the soundscape complexity, binned into two levels (described in Section 4.2)

Each of the 18 different experimental conditions were randomly assigned and replicated by 30 participants for a total of 540 participants. For each condition, participants annotated a sequence of 10 soundscapes that were randomly synthesized under constraints to match the max-polyphony and gini-polyphony levels of the condition. To mitigate any ordering effects, the presentation order of the 10 soundscapes within each condition was counterbalanced using Latin squares.

[2] https://github.com/justinsalamon/scaper



(a) Spectrogram and waveform visualizations of a 1.3 second long recording of a person whistling. In the spectrogram, a clear horizontal line is formed at the fundamental frequency of the whistle, clearly denoting a periodic sound. In the waveform, the periodicity is not clear, but the onsets and offsets of the whistle are still clearly defined.



(b) Spectrogram and waveform visualizations of a 1.3 second long recording of a person knocking on a door. In both visualizations, there are vertical lines clearly denoting the onsets of the door knocks. However, the decay of the sound marking the offset is clearer in the waveform visualization.

Fig. 2. Example visualizations.

Before running the full experiment, we tested and refined the Audio-Annotator interface and the methods outlined in this section using both preliminary in-person user testing and a pilot study with the crowdworkers until users could fluidly complete the experimental tasks.

4.1 Sound Visualizations

Sound visualizations are used to communicate acoustic information to a user through images. These are often two-dimensional visualizations in which the x-axis is time and the y-axis measures an application-appropriate variable. In multi-label sound-event annotation, users detect sound events and then annotate the events' onset/offset times and classes. In the signal of an audio recording, sound events typically manifest as spectro-temporal energy patterns. To aid in the detection and annotation of sound events, a sound visualization should help users detect these patterns. In our experiment, we compared three common visualization strategies: *waveform*, *spectrogram*, and a *no-visualization* control strategy.

4.1.1 Waveform. Waveform visualizations are likely the most common sound visualization, and are typically used in audio/video recording, editing, and music player applications. They are

two-dimensional visualizations in which the horizontal axis represents time, and the vertical axis represents the amplitude of the signal. From Fig. 2, we can see that waveforms clearly visualize the amplitude of a monophonic signal, but they make it difficult to parse audio-rate frequency content of a signal, especially when viewing several seconds at time. In our experiment, each pixel represents 12 ms of audio in the horizontal dimension and one of 256 amplitude levels in vertical dimension. Note that while our audio stimuli were sampled at 44.1 kHz (i.e., a temporal resolution of 0.023 ms), WaveSurfer.js achieves this ~500x reduction in temporal resolution using an algorithm that preserves peak amplitudes. At this resolution, the waveform visualization displays the signal's amplitude envelope, which is representative of the short-term energy in a signal over time.

4.1.2 Spectrogram. Spectrograms visualizations are more commonly found in professional audio applications. They are time-frequency representations of a signal that are typically computed using the short-time Fourier transform (STFT). They are two-dimensional visualizations that represent three dimensions of data: the horizontal axis represents time, the vertical axis represents frequency, and the color of each time-frequency pair (i.e., a pixel at a particular time and frequency) indicates its magnitude. In the spectrogram in Fig. 2, the colors on the yellow end of the spectrum represent high magnitudes whereas colors on the purple end of the spectrum represent low magnitudes. In Fig. 2 the spectrograms clearly decompose signals into their component frequencies, but fine amplitude variations can be more difficult to decode in color encodings. This decomposition by frequency likely makes it easier to detect events when multiple events overlap in time. In our experiment, our spectrogram was calculated with a linear-frequency log-magnitude STFT with a frame size of 12 ms, a hop size of 6 ms, and no zero-padding. With our audio stimuli sampled at 44.1 kHz, each pixel in the spectrogram represents the log-magnitude of an 84 Hz-wide frequency band over a time duration of 12 ms.

4.2 Audio Stimuli

One of the potential applications of the Audio-Annotator tool is to annotate sounds captured by an urban acoustic sensor network [34]. Therefore, we based the audio stimuli on soundscapes of urban sound events related to traffic, construction and other types of human activity. We curated a collection of 90 isolated sound events from several sound effects libraries [4, 14, 29], ten per each of the following sound classes: *car horn honking*, *dog barking*, *engine idling*, *gun shooting*, *jackhammer drilling*, *music playing*, *people shouting*, *people talking*, *siren wailing*. For the background, we used Brownian noise rather than a real recording of urban "hum" to avoid the possibility that a real recording could itself contain sound events and bias our experiment.

We synthesized 3000 soundscapes which were 10 s in duration and in which up to 9 sound events were randomly chosen without replacement from the 90 events in our collection. The signal-to-noise ratio of every sound event was chosen uniformly between 2–8 dB relative to the background. We then assigned the soundscapes into 6 complexity groups: 3 levels of max-polyphony $\times$ 2 levels of gini-polyphony. The max-polyphony levels were 0 (maximum polyphony of 1), 1 (maximum polyphony of 2), and 2 (maximum polyphony of 3–4). The gini-polyphony levels were 0 (gini-coefficient 0.5–1) and 1 (gini-coefficient 0–0.5). From the 3000 synthesized soundscapes, we randomly selected 10 soundscapes per each of the six complexity groups, resulting in a total of 60 audio stimuli.

4.3 Procedure

After accepting the task and acknowledging consent, participants completed a hearing screening (using the method described in [7]) to ensure they listened over a device with an adequate frequency

response (e.g., *not* laptop speakers) and followed instructions. Participants were allowed two attempts at passing the screening.

After the screening, participants were asked to complete a questionnaire on prior experiences with sound technology, sound labeling, and music instruments.

Next, participants were shown an instructional video lasting a few minutes that explained how to interpret the sound visualization they were assigned. The video also walked participants through the both the annotation interface and task.

Once participants watched the training video, they began the sound annotation activity using the Audio-Annotator tool. All participants first annotated the same practice soundscape of medium complexity to familiarize themselves with the task. Then, participants annotated the sequence of 10 soundscapes assigned to them. They were instructed to label all of the sounds in the soundscape. They could use as much time as they wanted to complete the task and could also refer to the instructions and video as needed.

Lastly, participants were directed to a questionnaire on their demographics and reflections on the annotation exercise.

4.4 Participants

We recruited 540 participants via Amazon Mechanical Turk. Participants were U.S. based with an approval rate higher than 99 percent. For additional quality control, participants were also required to pass a hearing screening as mentioned in the previous section. Each participant was assigned to one experimental condition. The task completion times and payment were dependent on their max-polyphony level. For max-polyphony levels 0, 1, and 2, the median completion times for an individual annotation task were 0.91 (M=1.22, SD=1.39), 1.71 (M=2.36, SD=2.88), and 2.59 (M=3.51, SD=3.97) minutes; and the payments were \$3.26, \$4.93, or \$5.62 USD respectively. Payment values were based on expected completion times as determined by a pilot experiment and a pay rate of \$7.25 USD per hour. To support differentiated payments based on max-polyphony, three separate HITs were created which indicated in their descriptions that participants were only allowed to complete one of the three HITs (CrowdCurio also enforced this constraint).

4.5 Quality Measures

To analyze our results we used several common binary classification measures: *accuracy*, *precision*, *recall*, and *F-score*. However, in this multi-label, time-series annotation scenario, we use these classification measures on a frame-level as implemented in *sed_eval* [32] and as used in the Detection and Classification of Acoustic Scenes and Events (DCASE) challenge [43]. Accuracy (A), precision (P), recall (R), and F-score (F) are defined as:

$$A = \frac{TP + TN}{TP + TN + FP + FN}, \quad P = \frac{TP}{TP + FP}, \quad R = \frac{TP}{TP + FN}, \quad F = \frac{2 \cdot P \cdot R}{P + R} \tag{1}$$

where TP, TN, FP, and FN are the number of respective *true-positives*, *true-negatives*, *false-positives*, and *false-negatives*. When calculating these measures on a frame-level, each soundscape is segmented into non-overlapping, fixed-length (e.g., 100 ms) time frames, and a frame is considered active if any portion of that frame overlaps with the time interval of an annotation. TP, TN, FP, and FN are then calculated in each frame independently for each sound event class.

The choice of the analysis frame size can greatly affect the results and should be chosen based on the application. For sound event detection in noise monitoring, a 1000 ms resolution is adequate for evaluation. However, for sound event detection in smart vehicles, a finer resolution is likely needed. For reference, a frame size of 1000 ms was used in the most recent DCASE challenge [30].

However, we opted for 100 ms frames in our analysis since we are concerned with detecting the preciseness and limitations of human annotations.

To investigate the reliability and redundancy trade-off, we first examined the agreement between the annotations of study participants. Not only does a measure of agreement give us an indication of how clear the annotation task was, but it can provide insight into how many annotators are required to establish a consensus of ground truth. To calculate agreement, we computed the mean frame-level accuracy (using 100 ms frames) for all pairs of annotations for each condition:

$$agreement_c = \frac{2}{|K_c|(|K_c|-1)} \sum_{i,j \in K_c} A_{c,i,j} \tag{2}$$

where c is the condition index, K_c is the number of participants for condition C, i and j represent the indices of a pair of participants, and $A_{c,i,j}$ is the accuracy score between the annotations of participants i and j for condition c.

To investigate whether annotators agreed upon a *correct* annotation, we examined how the number of annotators affects the quality measures of aggregate annotations in relation to the ground truth annotation by simulating an increasing number of annotators for each condition. To do so for each condition, we first shuffled the order of the condition's 30 annotations, each of which was contributed by a different participant. Then, starting with the first annotation, we progressively added each of the 30 participants' annotations, estimating and evaluating an aggregate annotation at each step. To estimate an aggregate annotation, we converted annotations to their frame-based time-series representation and marked a time frame as active only if at least half of the participants marked it as active. For each additional annotator, this aggregate annotation was estimated and evaluated against the ground-truth, resulting in a curve of quality as a function of the number of annotators. To replicate different possible orderings of annotators, we then repeated this whole process 30 times, each time with a different randomized order of annotators. This process was performed on all conditions.

5 RESULTS

5.1 Reliability and Redundancy Trade-off

5.1.1 Participant agreement. The median agreement over all conditions was 0.93 (M=0.92, SD=0.06). Our data did not conform to the assumptions of a classical ANOVA, therefore we performed a non-parametric Aligned Rank Transform (ART) ANOVA [51] to investigate the effect of *visualization*, *gini-polyphony*, and *max-polyphony* on agreement. We found that gini-polyphony ($F(1, 162) = 55.80$, $p < 0.001$) and max-polyphony ($F(2, 162) = 6.17$ $p < 0.01$) had significant effects on agreement, but visualization ($F(2, 162) = 1.10$, $p = 0.31$) did not. The interactions were not significant. In Table 1, we see that gini-polyphony had the strongest effect on agreement. Therefore, the sound visualization does not affect their agreement with each other, but the complexity of the soundscape does—e.g., longer, overlapping sound events have a negative influence on agreement. This suggests that more participants are needed to come to a consensus annotation for complex soundscapes.

5.1.2 Estimated quality as we increase number of annotators. In Fig. 3, we plotted the mean F-score for conditions as a function of the number of annotators with the simulated data described in Section 4.5. From this graph, it seems that as we increase the number of annotators, we can expect aggregate annotations that are on average more similar to the ground-truth annotations. While the quality doesn't converge over the 30 annotators, we do see a sharp rise in the first 5 annotators, a soft-knee at around annotators 5–10, and a more subtle, linear increase from annotators 10–30. The dotted lines in the graph identify the location at which 90% of the gain between the minimum

Factor	Level	Median Agreement (95% CI)
visualization	no-visualization	0.926 [0.918, 0.936]
	waveform	0.936 [0.914, 0.944]
	spectrogram	0.941 [0.930, 0.956]
max-polyphony	level 0	0.946 [0.932, 0.959]
	level 1	0.940 [0.926, 0.959]
	level 2	0.918 [0.905, 0.930]
gini-polyphony	level 0	0.948 [0.941, 0.960]
	level 1	0.909 [0.884, 0.925]

Table 1. Participant agreement as calculated by mean pairwise accuracy. Note that chance agreement would be 0.5.

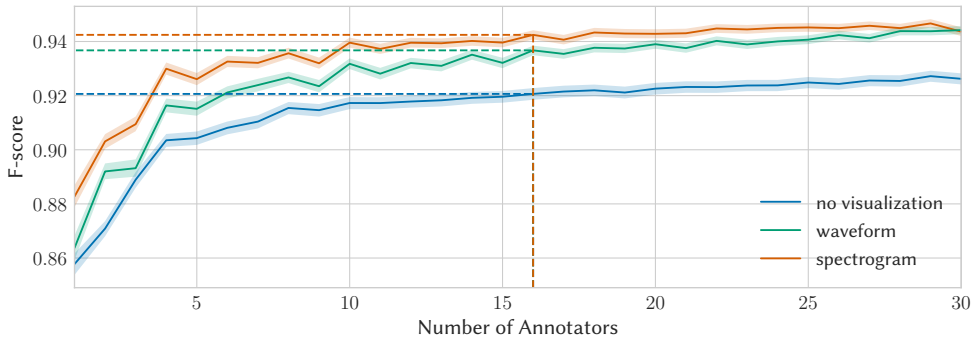


Fig. 3. Mean F-score of consensus annotations for conditions grouped by visualization as the number of annotators is increased. Bands are the 95% CIs of the means. The dotted lines identify the location at which 90% of the gain between the minimum and maximum quality is achieved.

and maximum quality is achieved. From these plots, we can estimate an appropriate number of annotators for a task. For example, these results indicate that there is minimal benefit of recruiting more than 16 annotators regardless of visualization.

5.2 Effect of Visualization On Annotation Quality and Speed

5.2.1 Quality of annotations without time limit. Since we did not limit the time to complete an annotation task, the participants' final annotations are reflective of the best annotations the participants were willing or able to do regardless of the time spent—an estimate of the upper bound on the quality of crowdsourced annotations with our interface, worker pool, and pay scale.

The quality of these annotations with respect to the ground-truth was quantified with the information retrieval measures from Section 4.5 (see Fig. 4). However, these results are more informative if we break the F-score into its precision and recall components. We calculated a one-way ART ANOVA to compare the effect of the visualization on precision and recall, and we find that while there was a significant effect of visualization on precision ($F(2, 5397) = 211.48$, $p < 0.001$),

[3]The box extends from the lower to upper quartile values of the data, with a line at the median and notches representing the 95% CI around the median. Whiskers represents the range of data that extends an additional $1.5 * IQR$ (interquartile range) past the edges of a box. Flier points are the outlying data points past the end of the whiskers.

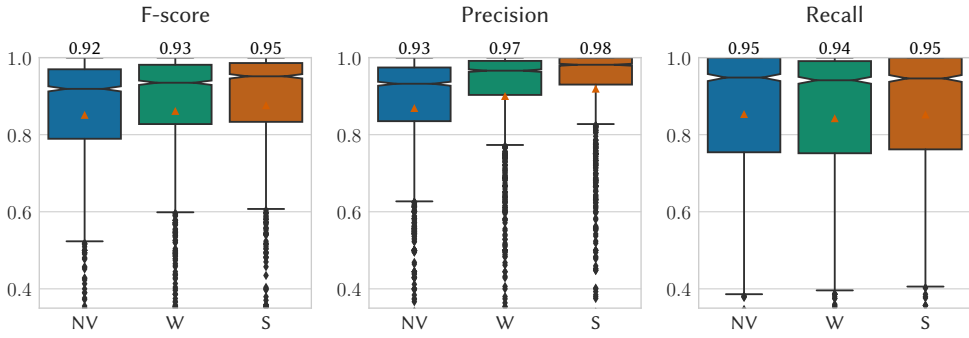


Fig. 4. Final annotation quality in reference to ground-truth labels grouped by visualization. NV=*no visualization*, W=*waveform*, S=*spectrogram*. Triangles indicate means. Median values appear above the plot.[3]

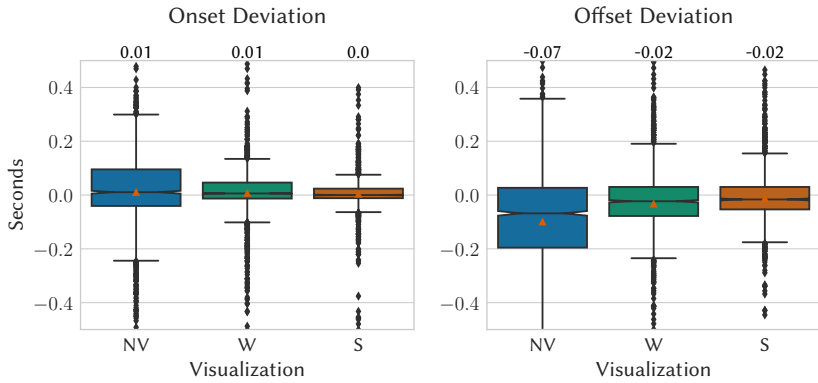


Fig. 5. Annotation onset and offset deviations from ground-truth (e.g., $groundTruth - estimate$) grouped by visualization. NV=*no visualization*, W=*waveform*, S=*spectrogram*. Triangles indicate means. Median values appear above the plot.[3]

there was not a significant effect of visualization on recall ($F(2, 5397) = 1.60,\ p = 0.20$). In a post-hoc Tukey HSD test, we found that the differences in precision between all of the condition groups were statistically significant at $\alpha = 0.05$. The spectrogram had the highest precision, followed by waveform and the no-visualization conditions.

To investigate this result further, we inspected how the onset and offset times of the annotations deviated from the ground truth for the three visualization condition groups. We calculated the difference between the ground-truth onset times and the annotated onset times for all annotated onsets within a ±1s error window of the ground-truth onset time, and we repeated this calculation for offset times. From this analysis, we found that while on average the onsets in the no-visualization conditions were aligned with ground-truth onsets (M=0.01, SD=0.15), the offsets in the no-visualization conditions usually came after the ground-truth offsets (M=-0.1, SD=0.21) (see Fig. 5). Onsets are typically associated with a peak in energy that makes them easier to detect

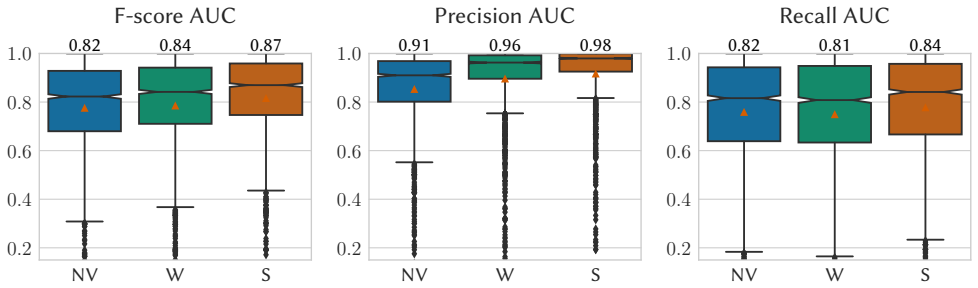


Fig. 6. Annotation speed/quality in reference to ground-truth labels calculated as the area under the curve (AUC) as time increases within an annotation task. NV=*no visualization*, W=*waveform*, S=*spectrogram*. Triangles indicate means.[3]

acoustically, but offset times are often more ambiguous and difficult to estimate. We speculate that annotators may be more conservative in their offset times, selecting a larger annotation region, when they do not have a visual aid to help place them. We also note a clear impact of the visualization on the variance of onset/offset times in the annotations ($p < 0.001$ for all pairs using a Levene test with Bonferroni correction): the variance is largest with no visualization and smallest with the spectrogram visualization. From the data, it seems that despite their clear representation of amplitude, waveform visualizations in general do not aid in annotating onset and offset times more precisely than spectrogram visualizations.

5.2.2 Quality of annotations as task time progresses. To investigate the speed and quality of annotations, we used interaction logs to recreate the annotation state at one second intervals for the first 5 minutes of each annotation task. However, since 96% of the annotations tasks took less than five minutes, we extended or trimmed all annotations to five minutes for comparison. We calculated F-score, precision, and recall at each time step for each annotation. We then jointly quantified the quality and speed by computing the area-under-the-curve (AUC) for each annotation's time-quality curve (see Fig. 6).

We calculated one-way ART ANOVAs to compare the effect of the visualization on the precision and recall AUCs. We found that with the AUCs there was again a significant effect of visualization on precision ($F(2, 5397) = 264.44,\ p < 0.001$), and a post-hoc Tukey HSD test found statistically significant differences between all three condition groups at $\alpha = 0.05$. There was also a significant effect of visualization on recall AUC ($F(2, 5397) = 6.68,\ p < 0.01$) unlike the lack of effect on final annotation recall. In a post-hoc Tukey HSD test, we found that the differences in recall AUC were only statistically significant at $\alpha = 0.05$ with the *spectrogram* pairs. Therefore, the spectrogram visualization not only leads to higher quality annotations than the other visualizations, it also leads to high quality annotations more quickly.

5.2.3 Task learning effects. Since a spectrogram is a complex visualization that most people are likely unfamiliar with, we investigated if it was helpful from the start, or if annotators learned how to effectively use it over the course of the session. To that end, we compared the final annotation quality measures as a function of the order in which the annotation tasks were presented to the annotators.

In Fig. 7, we see a positive trend for the spectrogram as the presentation position increases. Recall that the order of the soundscapes was counterbalanced so that any quality effects dependent on the soundscape itself will be averaged out and not present in these plots. When broken down

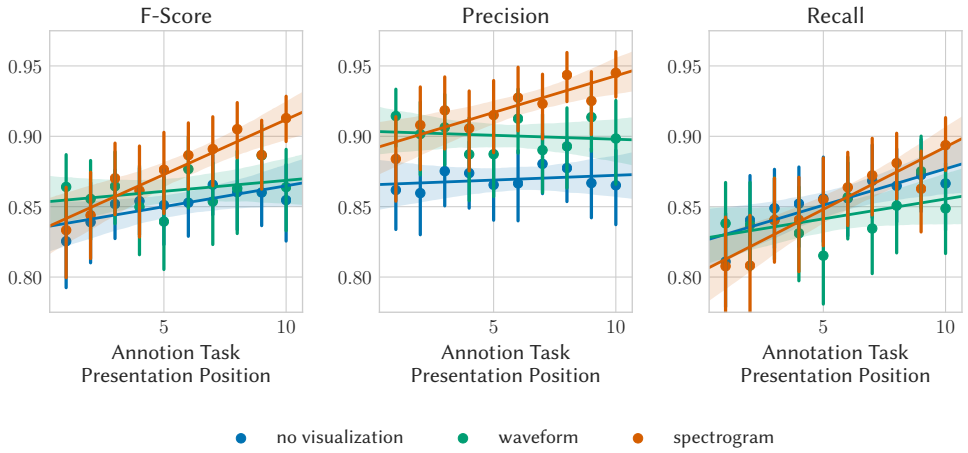


Fig. 7. Means (dots) and regression fits for F-score, precision, and recall of annotations as a function of the annotation task's presentation position. The error bars around the means and the error bands around the regression fits are their respective 95% CIs.

into precision and recall, we see that all of the visualization conditions exhibit a positive trend in recall as the presentation position increases, but only the spectrogram exhibits a positive trend in precision as the presentation position increases. We calculated two-way, repeated measures ART ANOVAs to test the effect of the visualization and presentation position on precision and recall. As expected from our previous analysis there was a significant effect of visualization on precision ($F(2, 537) = 67.45,\ p < 0.001$), but not recall ($F(2, 537) = 0.15,\ p = 0.86$). In addition, there was a significant effect of presentation position on recall ($F(9, 4833) = 4.12,\ p < 0.001$) but not on precision ($F(9, 4833) = 0.37,\ p = 0.95$), and neither had significant interaction effects between presentation position and visualization. Therefore, while there visually appears to be a learning effect manifested in precision when using a spectrogram, this is not statistically significant. However, there is a significant trend in recall for all visualizations, which indicates that with experience annotators improve at identifying sound events regardless of the visualization. This is a very encouraging result since it implies that after a learning period, we can expect annotation quality to be even higher.

5.3 Effect of Soundscape Attributes on Annotation Quality

In this section, we limit our investigation to the spectrogram visualization for simplicity since annotators generated the highest quality annotations with this visualization.

5.3.1 Soundscape complexity. The quality of an annotation is likely dependent on the complexity of the soundscape and its component sound events. To develop a strongly labeled dataset, it is important to establish what the limits of human annotations are for different soundscape complexities.

We calculated two-way ART ANOVAs to test the effect of the two soundscape complexity measures—gini-polyphony and max-polyphony—on the precision and recall of the final annotations generated using the spectrogram visualization. For both factors and their interactions, we found statistically significant effects on both precision and recall ($p < 0.001$ for all). In a post-hoc Tukey HSD test ($\alpha = 0.05$), we found significant differences between all 3 max-polyphony levels for

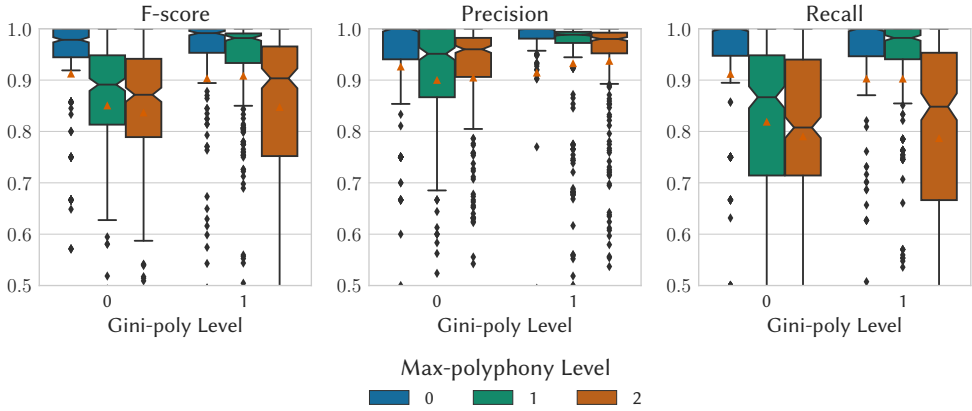


Fig. 8. Final annotation quality for spectrogram conditions broken down by the soundscape complexity factors.[3]

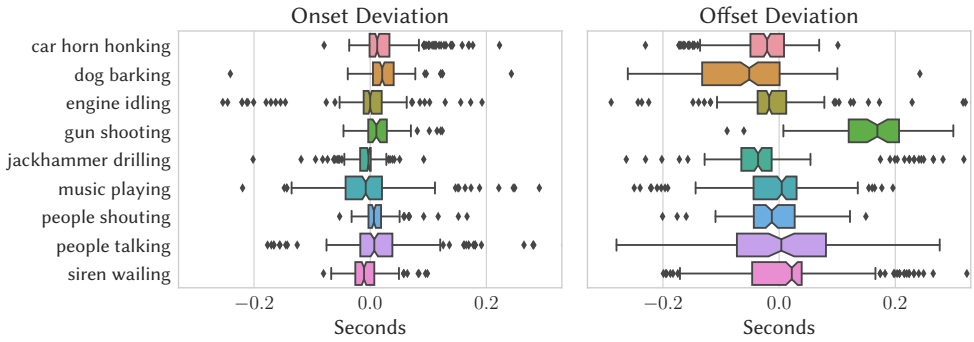


Fig. 9. Spectrogram annotations' onset / offset deviations from ground truth for each sound event class.[3]

recall, but only between level pairs (0,1) and (0,2) for precision. Fig. 8 shows that precision seems to be overall more strongly affected by gini-polyphony than by max-polyphony, and that the effect of the gini-polyphony is stronger when the max-polyphony is higher. This implies that it is harder to precisely annotate soundscapes in which many overlapping events are concentrated in time. That said, the variation in precision due to soundscape complexity is relatively small (max-polyphony $\eta^2 = 0.074$, gini-polyphony $\eta^2 = 0.061$). In contrast, while recall is also affected by both complexity measures and their interaction, it is more strongly affected by max-polyphony (max-polyphony $\eta^2 = 0.22$, gini-polyphony $\eta^2 = 0.029$). The effect of max-polyphony is also overall much greater in recall than in precision. This result is encouraging. It implies that even if annotators are given complex soundscapes, the resulting annotations will be precise even if some sound event annotations are missing.

5.3.2 Sound event class. The perceived onset and offset of a sound event within a soundscape mixture may differ from ground-truth onset and offset times calculated by Scaper. Scaper's ground-truth onset and offset times are more closely correlated to the perceived onset and offset times for

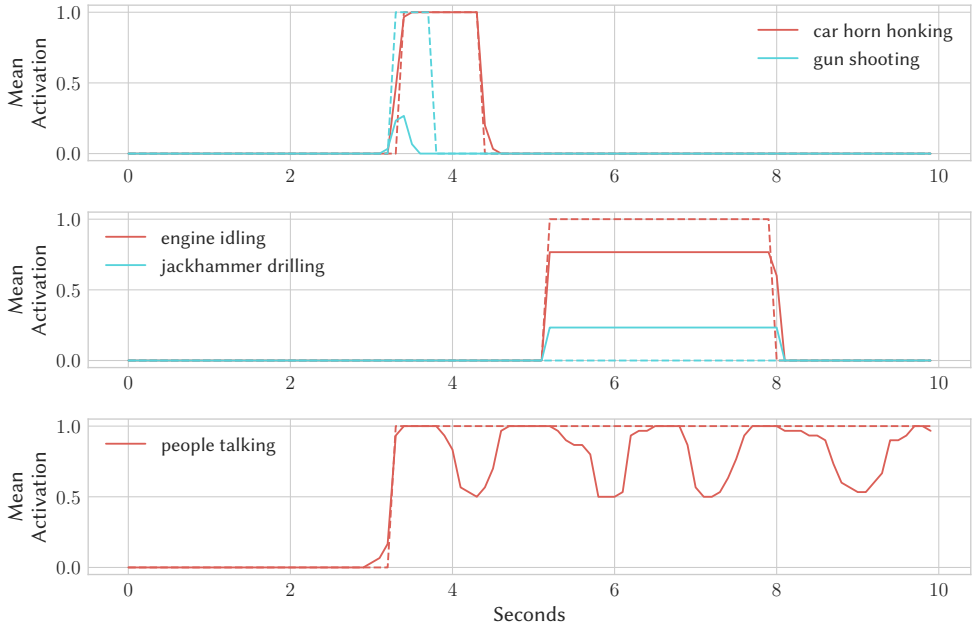


Fig. 10. Mean annotations for three soundscapes using 100 ms frame sizes. Solid line is mean. Dotted line is the ground truth.

a sound event in isolation, but in a mixture sounds may affect the perception of each other through masking phenomena [33]. For example, the sound of a car approaching and passing on a highway has a very long rise and fall in amplitude with a brief period of high amplitude as it is passes. If there are many other sound events also occurring, it may be difficult to precisely annotate the onset and offset (i.e. beginning and ending) times of such a sound event.

To investigate these differences, we examined how the onset and offset times of the annotations deviated from the ground truth for different sound event classes. These deviations were calculated as we did earlier in Section 5.2.1. In Fig. 9, we see that the onsets are all centered around zero for the sound event classes in our study. However, *gun shooting* annotations have offsets (M=0.16, SD=0.086) that are strongly biased from the ground-truth offsets. A gun shot is a loud, brief sound, but it can be accompanied by a long reverberation tail (i.e., a long acoustic decay). While this tail is reflected in the ground-truth, the annotators may not be able to hear this long reverberation tail; it may have been perceptually masked by other sounds.

In addition to small onset and offset errors, annotators may neglect to label a sound event or may confuse it with another. We calculated the sound-class-dependent precision and recall using both 0.1 s and 10 s (i.e. the entire soundscape duration) frame sizes for the spectrogram conditions annotations. For numerical reasons, instead of calculating precision and recall for an individual class on each annotation, we calculated the measures using the TP, FP, and FN counts over all of the spectrogram annotations. At the 0.1 s resolution, we found that *gun shooting* has a low recall score, as we expect from the large positive bias in the offset deviations. We also found that *dog barking* has a low precision score, as we would expect from its large interquartile range and negative bias in offset deviation. However, *gun shooting* also has a low precision score, which cannot be attributed

to its positive bias in offset—it must be caused by other errors such as incorrectly labeling other sounds as *gun shooting*.

When calculating precision and recall with a frame size of 10 s, we are essentially evaluating the annotations as weak labels. At the 10 s resolution, precision scores lower than 1.0 indicate that either annotators labeled a sound event that was not in the soundscape or they attributed a sound event to the wrong class. And recall errors indicate that a sound event was missed or labeled as the wrong class. At this resolution, *dog barking* precision is almost perfect, indicating that much of the errors at the 0.1 s resolution were due to offset deviations. The precision score of *gun shooting* is high at this resolution, but we observed a low recall for this class, which indicates that annotators are missing or mislabeling many of the *gun shooting* events. Unfortunately, because these annotations are multi-label not just multi-class, it is difficult to investigate these types of confusions and missing labels (e.g., a confusion matrix is not applicable).

However, we can see an example of a missed *gun shooting* event in Fig. 10 where we plot the mean annotations computed with a 100 ms frame size. Here it seems that the gunshot's onset co-occurred with *car horn honking*, which likely made it more difficult to detect. In another example, we can clearly see confusion between *jackhammer drilling* and *engine idling*, both of which contain pulsating engine noise. Lastly, the figure also shows how human annotators may segment very precisely, in this case annotating segments of speech around brief pauses. When thresholded at 0.5, the aggregate annotation would still be very similar to the ground-truth. However, the relationship between the "ground-truth" and the human annotations for this example highlights how the chosen frame size affects segmentation and could consequently affect the training and estimated performance of a machine listening algorithm.

6 DISCUSSION AND CONCLUSION

In this work, we contribute to crowdsourcing research by extending the theoretical understanding of, and establishing evidence-based design guidelines for, crowdsourcing audio annotations. Specifically, we sought to (1) explore and quantify the trade-off between reliability and redundancy in crowdsourced audio annotation; (2) identify which sound visualization aids lead to higher quality annotations, with and without taking into account time constraints; (3) examine the limitations of crowdsourced audio annotations resulting from soundscape complexity and sound event class. To achieve these objectives we ran a between-subjects study using a factorial experimental design, in which we varied two aspects of the annotation environment: the sound visualization aids presented to human annotators and the complexity of soundscapes they were asked to label.

Overall, we found that more complex soundscapes resulted in lower annotator agreement and that spectrogram visualizations are superior in helping to produce higher quality annotations at lower cost of time and human labor. We also found that recall is more affected than precision by soundscape complexity and mistakes can be often attributed to certain sound event characteristics such as long acoustic decays.

The findings have practical implications for the design and operation of crowdsourcing-based audio annotation systems:

- When investigating the redundancy/reliability trade-off in the number of annotators, we found that the value of additional annotators decreased after 5–10 annotators and that 16 annotators captured 90% of the gain in annotation quality. However, when resources are limited and cost is a concern, our findings suggest that five annotators may be a reasonable choice for reliable annotation with respect to the trade-off between cost and quality.
- When investigating the effect of sound visualizations on annotation quality, the results demonstrate that given enough time, all three visualization aids enable annotators to identify

sound events with similar recall, but that the spectrogram visualization enables annotators to identify sounds more quickly. We speculate this may be because annotators are able to more easily identify visual patterns in the spectrogram, which in turn enables them to identify sound events more quickly. We also see that participants learn how to use each interface more effectively over time, suggesting that we can expect higher quality annotations with even a small amount of additional training. From this experiment, we do not see any benefit of using the waveform or no-visualization aids when participants are adequately trained on the visualization.

- When examining the effect of soundscape characteristics on annotation quality, we found that while there are interactions between our two complexity measures and their effect on quality, they both affected recall more than precision. This is an encouraging result since it implies that when collecting annotations for complex scenes, we can trust the precision of the labels that we do obtain even if the annotations are incomplete. This also indicates that when training and evaluating machine listening systems, type II errors should possibly be weighted less than type I errors if the ground-truth was annotated by humans and vice versa if the ground-truth was annotated by a synthesis engine such as Scaper.
- We found substantial sound-class effects on annotation quality. In particular, for certain sound classes we found a discrepancy between the perceived onset and offset times when a sound is in a mixture and when a sound is in isolation (our ground-truth annotations). We speculate this effect is caused by long attack and decay times for certain sound classes (e.g., “gun shooting") and that the effect may also be present in other classes exhibiting similar characteristics (e.g., “car passing”), which were not tested. Future research focusing on specific sound classes would be needed to examine such sound class effects rigorously. For crowdsourcing systems’ design and operation purposes, these discrepancies could be accounted for in the training and evaluation of machine listening systems by having class-dependent weights for type II errors.
- Class-dependent errors also seemed to be a result of class confusions and missed detection of events. We speculate that one possible solution to mitigate confusion errors would be to provide example recordings of sound classes to which annotators could refer while annotating. It is also possible that saliency maximization techniques such as the one proposed by Lin et al. [27] could help reduce missed detection of events.

Future research may explore the use of other visual aids and their effect on annotation quality, reliability, and redundancy. Our experimental design choices of waveforms and spectrograms was based on current audio-industry standard practices, which are largely driven to fit the needs of sound-savvy users whose perception of sound is different than those of the average crowd worker. However, annotators’ sound perception and experience using audio systems may affect the utility that different visualizations have for them. Furthermore, other visualizations not explored in this study may be more suitable for non-experts. Future research in which alternative visual aids are used and users’ experience with audio systems is accounted for and varied across conditions can help determine the right fit between crowd workers’ audio experience and crowdsourcing annotation tools. Future research may also investigate incorporating user-dependent annotation aggregation methods, since it is likely that some users produce more reliable annotations than others. In addition, future research may also explore alternative crowdsourcing workflows (e.g. breaking the overall workflow into “find" and “verify" stages), a direction that has been shown to be advantageous for other temporal domains such as video annotation [18]. Finally, in future research we plan to further investigate the relationship of audio and visual stimuli in audio annotation tasks and their connection to findings in studies of perception and cognition.

The goal of this study is to establish best practices for evidence-based crowdsourced sound annotation so that future researchers and practitioners can benefit from our work. While future sound annotation needs may or may not require paid (vs. volunteer) crowd work, we chose the paid crowdsourcing approach in order to achieve results quickly and with relative ease of screening participants for past performance and location. We acknowledge that our findings may not be fully transferable to a volunteer-based environment (e.g. citizen science), but we expect that the findings presented here will not be affected much by differences between paid and unpaid settings. Such differences are likely to impact annotators' motivations and attrition rate, but we don't have reasons to believe that they would impact the efficacy of the visualizations and user interfaces tested in our study.

Our class-dependent analysis and graphs of mean annotations (see Section 5.3.2) raise interesting questions in regards to the definition of "ground truth" annotations. For example, at what time delay do neighboring utterances or jackhammer knocks transition from being one event to separate events? Can we use priming examples or exposure to others' annotations to bias this transition point? Also, should ground truth be defined by the limits of what can be perceived by humans or should we strive to train our machines for super-human accuracy? While the aggregate annotations of a crowd are more accurate than an average individual (see Section 5.1.2), if we strive for truly super-human accuracy then our training sets will need to be synthesized or multi-modal in order to incorporate information not perceptible by humans from the audio modality alone. The answers to these questions are likely application-dependent (e.g. bioacoustic monitoring will have different requirements than noise pollution monitoring), but they are important to consider when designing an audio annotation system.

As a final note, we've released a dataset called the Seeing Sound Dataset which contains all of the soundscapes, final annotations, and participant demographics. This dataset is available at https://doi.org/10.5281/zenodo.887924.

Taken together, the findings of this study make an important step in implementing rigor to the design and operation of crowd-based sound annotation systems, thereby making crowdsourcing a more evidence-based practice. In doing so, we extend the growing body of literature on crowdsourcing to incorporate a domain that has received little research attention so far, despite its importance for other areas of research such as machine learning and urban informatics.

ACKNOWLEDGMENTS

This work was partially supported by National Science Foundation award 1544753 (https://www.nsf.gov/awardsearch/showAward?AWD_ID=1544753). This work also would not be possible without the participation and effort of many workers on Amazon's Mechanical Turk platform.

REFERENCES

[1] Apple Inc. 2017. Apple GarageBand. (2017). http://www.apple.com/mac/garageband/.
[2] Avid Technology, Inc. 2017. Pro Tools. (2017). http://www.avid.com/pro-tools.
[3] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. 2016. Soundnet: Learning sound representations from unlabeled video. In *Proc. of Advances in Neural Information Processing Systems*. 892–900.
[4] BBC. 2017. BBC Sound Effects Library. (2017). https://www.sound-ideas.com/Product/154/BBC-Sound-Effects-Library-Original-CDs-1-60
[5] Mark Cartwright and Bryan Pardo. 2013. Social-EQ: Crowdsourcing an Equalization Descriptor Map. In *Proc. of the International Society for Music Information Retrieval Conference*.
[6] Mark Cartwright and Bryan Pardo. 2015. VocalSketch: Vocally Imitating Audio Concepts. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 43–46. https://doi.org/10.1145/2702123.2702387
[7] Mark Cartwright, Bryan Pardo, Gautham Mysore, and Matthew Hoffman. 2016. Fast and Easy Crowdsourced Perceptual Audio Evaluation. In *Proc. of the International Conference on Acoustics, Speech and Signal Processing*.

[8] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. 2017. Revolt: Collaborative Crowdsourcing for Labeling Machine Learning Datasets. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*.

[9] Cornell Lab of Ornithology. 2017. Raven. (2017). http://www.birds.cornell.edu/brp/raven/RavenOverview.html

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proc. of the IEEE Cnference on Computer Vision and Pattern Recognition*. IEEE, 248–255.

[11] Jia Deng, Olga Russakovsky, Jonathan Krause, Michael S Bernstein, Alex Berg, and Li Fei-Fei. 2014. Scalable multi-label annotation. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 3099–3102.

[12] Thomas Fillon, Joséphine Simonnot, Marie-France Mifune, Stéphanie Khoury, Guillaume Pellerin, and Maxime Le Coz. 2014. Telemeta: An open-source web framework for ethnomusicological audio archives management and automatic analysis. In *Proc. of the International Workshop on Digital Libraries for Musicology*. ACM, 1–8.

[13] Pasquale Foggia, Nicolai Petkov, Alessia Saggese, Nicola Strisciuglio, and Mario Vento. 2015. Reliable detection of audio events in highly noisy environments. *Pattern Recognition Letters* 65 (2015), 22–28.

[14] Freesound. 2017. Freesound. (2017). http://freesound.org/

[15] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio Set: An ontology and human-labeled dartaset for audio events. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*.

[16] Sébastien Gulluni, Slim Essid, Olivier Buisson, and Gaël Richard. 2011. An Interactive System for Electro-Acoustic Music Analysis. In *ISMIR*. 145–150.

[17] Katspaugh. 2017. wavesurfer.js. (2017). https://wavesurfer-js.org/

[18] Bongjun Kim and Bryan Pardo. 2017. I-SED: an Interactive Sound Event Detector. In *Proc. of the International Conference on Intelligent User Interfaces*. ACM, 553–557.

[19] Juho Kim, Phu Tran Nguyen, Sarah Weir, Philip J. Guo, Robert C. Miller, and Krzysztof Z. Gajos. 2014. Crowdsourcing step-by-step information extraction to enhance existing how-to videos. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 4017–4026. https://doi.org/10.1145/2556288.2556986

[20] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, and David A Shamma. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *arXiv preprint arXiv:1602.07332* (2016).

[21] Ranjay A Krishna, Kenji Hata, Stephanie Chen, Joshua Kravitz, David A Shamma, Li Fei-Fei, and Michael S Bernstein. 2016. Embracing error to enable rapid crowdsourcing. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 3167–3179.

[22] Walter S Lasecki, Mitchell Gordon, Danai Koutra, Malte F Jung, Steven P Dow, and Jeffrey P Bigham. 2014. Glance: Rapidly coding behavioral video with the crowd. In *Proc. of the ACM Symposium on User Interface Software and Technology*. ACM, 551–562.

[23] Edith Law and Luis von Ahn. 2009. Input-agreement: a new mechanism for collecting data using human computation games. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*. ACM. https://doi.org/10.1145/1518701.1518881

[24] Edith Law, Conner Dalton, Nick Merrill, Albert Young, and Krzysztof Z Gajos. 2013. Curio: a platform for supporting mixed-expertise crowdsourcing. In *Proc. of the AAAI Conference on Human Computation and Crowdsourcing*.

[25] Jin Ha Lee. 2010. Crowdsourcing Music Similarity Judgments using Mechanical Turk. In *Proc. of the International Society for Music Information Retrieval Conference (Series Crowdsourcing Music Similarity Judgments using Mechanical Turk)*. 183–188.

[26] Mark Levy. 2011. Improving Perceptual Tempo Estimation with Crowd-Sourced Annotations. In *Proc. of the International Society for Music Information Retrieval Conference*. 317–322.

[27] Kai-Hsiang Lin, Xiaodan Zhuang, Camille Goudeseune, Sarah King, Mark Hasegawa-Johnson, and Thomas S Huang. 2012. Improving faster-than-real-time human acoustic event detection by saliency-maximized audio visualization. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2277–2280.

[28] Chris J Lintott, Kevin Schawinski, Anže Slosar, Kate Land, Steven Bamford, Daniel Thomas, M Jordan Raddick, Robert C Nichol, Alex Szalay, and Dan Andreescu. 2008. Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society* 389, 3 (2008), 1179–1189.

[29] Lucasfilm. 2017. Lucasfilm Sound Effects Library. (2017). https://www.sound-ideas.com/Product/435/lucasfilm-sound-effects-library/

[30] Annamaria Mesaros and Toni Heittola. 2016. IEEE AASP Challenge: Detection and Classification of Acoustic Scenes and Events. (2016). http://www.cs.tut.fi/sgn/arg/dcase2016/task-sound-event-detection-in-real-life-audio.

[31] Annamaria Mesaros, Toni Heittola, Antti Eronen, and Tuomas Virtanen. 2010. Acoustic event detection in real life recordings. In *Proc. of the European Signal Processing Conference*. IEEE, 1267–1271.

[32] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. 2016. Metrics for polyphonic sound event detection. *Applied Sciences* 6, 6 (2016), 162.

[33] Brian CJ Moore. 2012. *An introduction to the psychology of hearing*. Brill.
[34] Charlie Mydlarz, Justin Salamon, and Juan Pablo Bello. 2016. The Implementation of Low-cost Urban Acoustic Monitoring Devices. *Applied Acoustics* In Press (2016).
[35] Gabriel Parent and Maxine Eskenazi. 2011. Speaking to the Crowd: Looking at Past Achievements in Using Crowdsourcing for Speech and Predicting Future Challenges. In *Proc. of the INTERSPEECH Conference*. Citeseer, 3037–3040.
[36] Queensland University of Technology's Ecoacoustics Research Group. 2017. Bioacoustics Workbench. (2017). https://github.com/QutBioacoustics/baw-client
[37] Bryan C Russell, Antonio Torralba, Kevin P Murphy, and William T Freeman. 2008. LabelMe: a database and web-based tool for image annotation. *International Journal of Computer Vision* 77, 1 (2008), 157–173.
[38] Justin Salamon, Duncan MacConnell, Mark Cartwright, Peter Li, and Juan Pablo Bello. 2017. 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.
[39] Lior Shamir, Carol Yerby, Robert Simpson, Alexander M von Benda-Beckmann, Peter Tyack, Filipa Samarra, Patrick Miller, and John Wallin. 2014. Classification of large acoustic datasets using machine learning and crowdsourcing: Application to whale calls. *The Journal of the Acoustical Society of America* 135, 2 (2014), 953–962.
[40] Gunnar A Sigurdsson, Olga Russakovsky, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Much ado about time: Exhaustive annotation of temporal data. *arXiv preprint arXiv:1607.07429* (2016).
[41] Jacquelin A Speck, Erik M Schmidt, Brandon G Morton, and Youngmoo E Kim. 2011. A Comparative Study of Collaborative vs. Traditional Musical Mood Annotation. In *Proc. of the International Society for Music Information Retrieval Conference*. Citeseer, 549–554.
[42] Stanford Vision Lab. 2017. ImageNet. (2017). http://image-net.org
[43] Dan Stowell, Dimitrios Giannoulis, Emmanouil Benetos, Mathieu Lagrange, and Mark D Plumbley. 2015. Detection and classification of acoustic scenes and events. *IEEE Transactions on Multimedia* 17, 10 (2015), 1733–1746.
[44] Kyle A Swiston and Daniel J Mennill. 2009. Comparison of manual and automated methods for identifying target sounds in audio recordings of Pileated, Pale-billed, and putative Ivory-billed woodpeckers. *Journal of Field Ornithology* 80, 1 (2009), 42–50.
[45] Anthony Truskinger, Mark Cottman-Fields, Daniel Johnson, and Paul Roe. 2013. Rapid scanning of spectrograms for efficient identification of bioacoustic events in big data. In *eScience (eScience), 2013 IEEE 9th International Conference on*. IEEE, 270–277.
[46] Anthony Truskinger, Haofan Yang, Jason Wimmer, Jinglan Zhang, Ian Williamson, and Paul Roe. 2011. Large scale participatory acoustic sensor data analysis: tools and reputation models to enhance effectiveness. In *Proc. of the IEEE International Conference on E-Science*. IEEE, 150–157.
[47] University of Groningen Sensory Cognition Group. 2017. Soundscape Annotation Tool. (2017). http://www.ai.rug.nl/~vdlinden/annotationtool/index.html
[48] Carl Vondrick, Donald Patterson, and Deva Ramanan. 2013. Efficiently scaling up crowdsourced video annotation. *International Journal of Computer Vision* 101, 1 (2013), 184–204.
[49] Wesley Willett, Jeffrey Heer, and Maneesh Agrawala. 2012. Strategies for crowdsourcing social data analysis. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 227–236.
[50] Charles G Willis, Edith Law, Alex C Williams, Brian F Franzone, Rebecca Bernardos, Lian Bruno, Claire Hopkins, Christian Schorn, Ella Weber, and Daniel S Park. 2017. CrowdCurio: an online crowdsourcing platform to facilitate climate change studies using herbarium specimens. *New Phytologist* (2017).
[51] Jacob O Wobbrock, Leah Findlater, Darren Gergle, and James J Higgins. 2011. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 143–146.
[52] Wayne Xiong, Jasha Droppo, Xuedong Huang, Frank Seide, Mike Seltzer, Andreas Stolcke, Dong Yu, and Geoffrey Zweig. 2016. Achieving human parity in conversational speech recognition. *arXiv preprint arXiv:1610.05256* (2016).
[53] Shlomo Yitzhaki and Edna Schechtman. 2012. *The Gini Methodology: A primer on a statistical methodology*. Vol. 272. Springer Science & Business Media.

Received April 2017; revised July 2017; accepted August 2017