DOES PAID CROWDSOURCING STILL PAY OFF? SIFTING THROUGH ANNOTATOR NOISE IN CROWDSOURCED AUDIO LABELS

Manu Harju, Irene Martín-Morató, Annamaria Mesaros

Signal Processing Research Centre, Tampere University, Finland
{manu.harju, irene.martinmorato, annamaria.mesaros}@tuni.fi

ABSTRACT

Paid crowdsourcing has emerged as a popular method for annotating diverse data types such as images, text, and audio. However, the amount of carelessly working annotators has increased as platforms have become more popular, leading to an influx of spam workers that answer at random, which renders the platforms unusable. This paper documents our attempt to annotate the DESED dataset using Amazon's Mechanical Turk, and failing to obtain any useful data after two attempts. Our observations reveal that while the number of workers performing the tasks has increased since 2021, the quality of obtained labels has declined considerably. After successful trials for annotating audio data in 2021 and 2022, in 2024 the same user interface annotation setup predominantly attracted spammers. Given the consistent task setup and similarity to previous attempts, it remains unclear whether the workers are inherently subpar or if they are intentionally exploiting the platform. The bottom line is that despite spending a considerable amount of money on it, we obtained no usable data.

Index Terms— Data annotation, crowdsourcing

1. INTRODUCTION

Crowdsourcing is a collaborative online process where a group of individuals with different skills, knowledge, and backgrounds is participating to work on some *task*. Tasks are usually surveys, data annotations, description collections, or other such assignments which are difficult for computers but easy for humans [1]. Amazon Mechanical Turk (AMT) uses the term Human Intelligence Task (HIT) for a single annotation/answer. Crowdsourcing involves two key roles: *requesters*, who create data-collection tasks, and *workers*, who complete those tasks. In paid crowdsourcing, requesters compensate workers for their completed assignments. One benefit of using paid crowdsourcing platforms is their vast pool of workers. However, since the work is done by humans with varying abilities and backgrounds, the crowdsourced results are likely to contain some amount of errors. Some errors are simply mistakes, but there are also workers aiming to collect the task rewards without caring much about their work.

The quality of the crowdsourcing results can be improved by (1) taking more control of the data collection process itself, and (2) using different postprocessing and aggregation methods. The former means checking the correctness of some part of the annotations, and rejecting incorrect ones and possibly banning the workers from taking more tasks. In case of a label assigning task, the latter can be done e.g. by directly optimizing the labels or through estimating the reliabilities of the individual annotators. The study in [1] presents a good overview of different aggregation methods. In practice, the two approaches should be used together, but often the purpose of using crowdsourcing is lost if keeping the annotation process clean requires too much effort.

The general setting in the crowdsourcing platforms makes the workers to do *invisible labour*, meaning that part of the time spent on the platform does not generate any income. Invisible labour includes e.g. rejected work, finding new tasks, interacting with requesters, and managing payments [2]. A study from 2018 reports that the average requester on Amazon Mechanical Turk paid $11/hour. However, lower-paying requesters were publishing more work, and as an effect the median wage for workers was approximately $2/hour [3]. Due to the factors explained above, working on microtasks can be difficult to make profitable.

There has been some development of guidelines for requesters on how to make their tasks ethical, e.g by having clear instructions and examples of good answers for the task, and reasonable reward for the tasks. Furthermore, Hiippala *et.al.* argue that human errors should not be a reason for rejection [2]. This creates a problem for the requester: how to recognize when something is a human error and not a bad-faith answer? To be sure to stay on the fair side, the requester should only reject the most obvious cases, e.g. tasks done in too short time. This, in turn, opens up the opportunity for the workers to exploit the requesters by doing the task carelessly or simply bypassing the task and instead providing a response that *seems* correct. The study in [4] shows that the amount of bad survey data has risen from 2% in 2013 and 5% in 2018 to almost 89% in 2022; the authors bring up the same question of how to distinguish a bad-faith answer. They also note that the workers were likely to either co-operate closely with each other or use multiple accounts, as some of the answers were too similar.

In audio, paid crowdsourcing has been used for creating datasets of speech transcriptions [5], audio captions [6], positive and negative audio-caption pairings [7]. However, hearing and classification of sounds are subjective, and e.g. the annotation context and the worker's personal background affects the recognized sounds [8]. Furthermore, requesters can only recommend but cannot control the environment and equipment the workers are using for the tasks, making the distinction between a human error and a bad-faith answer even more trickier.

This paper documents the efforts we made in 2024 to annotate part of the DESED [9] data for the DCASE 2024 Sound Event Detection Task. Our previous work has repeatedly shown that it is possible to obtain reliable annotations for sound events. We started with a study using synthetic data [10]; as the process was shown to work, we moved on to annotate real data [11]. Unfortunately, it seems that the process is no longer working as expected. The contributions of

This work was supported by Academy of Finland grant 332063 "Teaching machines to listen".

the paper are: (1) we analyze the quality of annotations obtained through paid crowdsourcing, observing that it has decreased considerably in a few years, and (2) we show that multi-annotator competence estimation (MACE) [12] is robust against bad-faith annotators, even in large quantities.

2. COLLECTING THE DATA

2.1. Annotation setup

In all the annotation experiments in this paper we followed the procedure presented in [10]. The main idea was to break down the complicated task of annotating onset and offset times beside the class labels of sound events into a simple tagging task of highly overlapping sound clips. Afterwards, the temporally weak annotations could be aggregated with the temporal information. The sound clips used in the experiments were 10-second clips cut out from longer pieces of audio. The start times for the clips were increased one second at a time, such that two consecutive clips have nine seconds of overlapping audio. Each clip was annotated by multiple workers, 5 in the previous work, and 3 in the current experiments. We opted for the lower number now due to the high number of clips to annotate and therefore high cost. As a consequence of the overlap, each one-second segment of audio was included in a total of 50 annotation tasks (30 in 2024). Each annotator's competence value was estimated using MACE [12], and the labels were reconstructed by taking weighted averages over all the opinions that included each one-second segment using the competences as weights [11].

For the first experiment in 2021, the audio was generated by using the isolated sound events from UrbanSound8k [13]. The events were sampled from six classes, and the synthesized dataset consisted of 20 3-minute long files [11]. For the following experiments for MAESTRO Real [14] in 2021 and 2022 we used data recorded from five different scenes of the TUT Acoustic Scenes 2016 dataset [15]; for each scene we used six event classes. Due to some overlap in the classes, the total number of classes of the resulting dataset is 17, but in the HITs the tasks were presented per acoustic scene, i.e. with only six classes to tag. Finally in January 2024, we aimed to annotate 556 files for the evaluation set of the Sound Event Detection task in DCASE 2024 Challenge. For this last annotation task, the target annotation length was 10 seconds; in order to cover this length, due to the annotation method explained above, the source files were 28 seconds long, including 9 extra seconds on each side of the target segment. Furthermore, the number of event classes was ten instead of the six used in previous experiments.

We verified that using three annotators per file instead of five is sufficient by sampling annotations using the data from MAESTRO Real experiments. Using only three randomly selected annotators per file gave similar results as the reconstruction based on five annotators.

2.2. Task description

The task layout used to collect the annotations contained an audio player, a short list of instructions, and a selector for the event classes. The instructions advised doing the experiment in a quiet environment and with good quality headphones. It was mentioned that the annotators could playback the audio as many times they wanted. The annotators were asked to select all the sound event classes they can recognize in the clip from the given list.

In all experiments the files were divided into 15 different batches based on their start time. The first batch contained all the clips with start times 0, 15, 30, ..., the second batch with start times 1, 16, 31, ..., and so on. By this construction, the gap between two clips in a batch is always at least 15 seconds.

We required workers to have at least 1000 completed HITs and at least 90% approval rate. In practice, we accepted almost all annotations. The annotations completed in shorter time than the sound clip were taken into closer inspection, and the ones tagging clearly incorrect labels were rejected. However, the rejected tasks annotators were not banned from taking more tasks. One thing we noticed in the last experiment was that the workers deduced this and simply spent more time on the task such that these "too fast" annotations were not anymore present in the later batches.

2.3. Two attempts

In the first DESED annotation (DESED/A1), we introduced fields for the annotator confidence: for any positive label assigned, the annotator had to specify how confident they were about the label. The confidence was given on a six-step scale from 50% to 100% with 10% increments. The scope was to study the relationship between estimated competence and self-evaluated confidence of the annotators.

After the data collection we noticed that the competence estimation resulted in a very skewed distribution, where most of the competence values were centered close to zero. Furthermore, the aggregated labels for most of the classes did not agree very well with the reference annotation[1], and aggregating the annotations using the previously used method resulted in useless data. There was a large number of annotators doing only a few tasks, hinting that the task setup was too complicated and driving the workers away. The number of available HITs was approximately three-fold compared to the earlier experiments, but the highest number of files annotated by a single worker in DESED/A1 was 112.

We do not know what caused the high number of bad annotations. Based on the task setup, there are two possible factors. First, the number of classes was increased from six to ten, making an individual task more complex. Second, the annotators had to answer the question about confidence for each positive label, which adds to the annotators' work load. We also hypothesized that the reason for such a unusual competence distribution was that the data was too sparse for MACE to handle, due to the high number of annotators doing few HITs. We decided to repeat the process without the confidence question. For the second DESED annotation (DESED/A2) we reverted to the basic task layout to see if there was any difference without the question about confidence. Unfortunately, in terms of label quality, we ended up with similar results as in DESED/A1.

3. SIFTING THROUGH THE DATA

3.1. Analysis of the outcome

We started the analysis of DESED/A1 with the standard approach by estimating the annotators' competence values using MACE. The competences can vary from 0 to 1, and according to MACE the vast majority of the annotators had extremely low competence values: the median competence was 0.09 and the fraction of annotators with a competence value smaller than 0.01 was over 19%. Figure 1 shows the histograms of the competence values in both experiments. In DESED/A1 the amount of workers annotating at most 5 files was 51%; this number decreased to 36% in DESED/A2. However, the

[1] Reference annotation available as manually annotated strong labels [9]

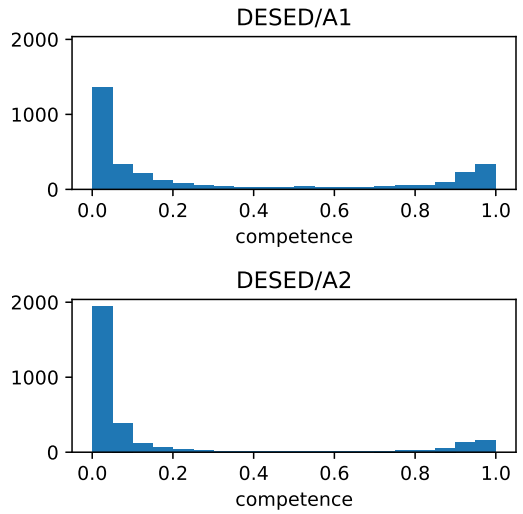

Figure 1: Histograms of the estimated competences.

competence distribution in DESED/A2 was even more skewed than in DESED/A1. The numbers suggest that removing the question about confidence made the task more attractive for workers, but maybe also less engaging.

The annotators often disagreed with each other. In terms of aggregation, having only three opinions per file instead of five accentuated the problem caused by disagreements. 62% of the clips in DESED/A1 had completely disjoint class labels from the three annotators, and this number increased to 87% in DESED/A2. We observed that annotators also disagreed with themselves: there were 20 annotators who annotated the same file in both DESED/A1 and DESED/A2, and in 15 of the cases they assigned completely different sets of labels. The inconsistencies can be due to changes in the circumstances, but either the sounds are very hard to recognize, or the workers did not perform the task genuinely. Nevertheless, these findings illustrate the randomness of the annotator behavior overall.

Table 1 shows the dataset sizes and numbers of workers, as well as the time of the data collection. Due to the long gap between the last MAESTRO Real annotation and DESED/A1, it is understandable that 87% of the worker accounts were new in our experiments. However, between the two DESED experiments there was less than a five months gap, and still almost half of the annotator accounts in DESED/A2 were completely new to our tasks.

3.2. Tagging precision and MACE

For some of the scenes there exist temporally strong labels. We converted the available labels into tags of the annotated clips to measure each annotator's tagging performance. The tagging performances over the workers in different scenes are shown in Table 2. To check the overall quality of the answers, we also calculate the average precision over HITs. With this, precision in DESED/A1 and A2 drops to 43.2 and 43.9, respectively. This indicates that the workers completing more tasks are not producing the better labels.

The worker competence is computed based on the tagging task, and we expect a connection between the competence and tagging performance. In Fig. 2 we show the precision on the individual and combined experiments. The annotators are divided into equally-sized groups based on their competence values, with the bin borders

Scene	Date	#Clips	#Workers	Acc. workers
Synthetic	3/2021	3420	680	680
City center	6/2021	3544	717	1154
Residential area	6/2021	3429	861	1517
Cafe/restaurant	9/2022	3273	1554	2870
Grocery store	9/2022	2840	1509	3450
Metro station	9/2022	3418	1641	3832
DESED/A1	1/2024	10545	3295	6711
DESED/A2	6/2024	10545	3059	8125

Table 1: Annotation dates and numbers of individual sound clips and annotators. The last column shows the cumulative number of workers that participated in our data collections.

Scene	F-score	Precision	Recall
Synthetic	72.3	89.3	62.8
City center	50.4	60.9	45.6
Residential area	51.0	57.2	50.0
DESED/A1	37.0	50.9	31.4
DESED/A2	37.8	51.4	31.7

Table 2: Average tagging scores in different experiments.

marked on the x-axis. The competence quantiles are very skewed: in DESED/A1 3/5 of the workers have a competence less than 0.17, and in DESED/A2 4/5 of the workers have competence less than 0.15. Combining the data from the two flattens the competence distribution, but adds a few outliers in the plot. These results indicate that MACE is still able to identify the better performing workers despite the vast amount of noise in the annotation. Furthermore, combining the data seems to improve the MACE output. Unfortunately we do not have enough reliable annotations even when the experiments are combined.

3.3. Comparing aggregated labels against reference data

We compare the reconstructed soft labels to the reference data using macro soft F-score [16] to avoid the problem of choosing the threshold value for binarizing the data. Table 3 shows the F-scores for the scenes we have a reference annotation available. The scores for both DESED experiments are similar to each other, but also extremely low. Furthermore, when the annotation data is combined from the two experiments, the standard method results in worse labels than either of the experiments alone. Table 3 also includes the average competence C_{avg} evaluated using MACE for each annotation set. The average competence is not telling the whole truth, as the weighting is in practice determined by the differences between the competences related to a single segment. This can also be seen in the combined case, where the average competence is as high as 0.58.

For further analysis, we can inject the reference annotation into the competence estimation along with the collected labels to obtain a competence value C_{ref} for the reference labels. If the reference labels mostly agree with the annotations, the annotators and the reference should have a high competence. Similarly, if the reference labels are mostly different from the annotated labels, MACE interprets the reference as an annotator submitting random labels, resulting in a low competence value. Combining the data from the two experiments improves the MACE results in terms of higher C_{avg} and C_{ref}, but does not change much the competence distribution.

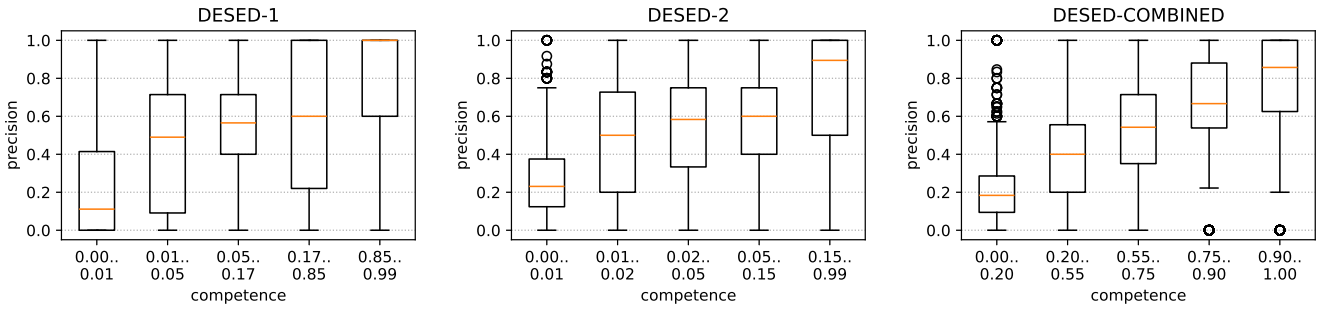

Figure 2: Tagging precision for DESED/A1, DESED/A2, and the combined data. Workers are grouped by their competence values into equally-sized groups. The skewness of the competence value distributions in the two DESED experiments can be seen in the bin borders.

Scene	F_M	C_{avg}	C_{ref}
Synthetic	63.8	0.73	0.89
City center	45.4	0.43	0.68
Residential area	39.3	0.53	0.57
Cafe/restaurant	-	0.43	-
Grocery store	-	0.42	-
Metro station	-	0.34	-
DESED/A1	31.2	0.31	0.35
DESED/A2	31.0	0.17	0.36
DESED/A1 + A2	29.1	0.58	0.61

Table 3: Soft macro F-scores F_M of the reconstructed sound event labels, average competences of the annotators C_{avg}, and the reference annotation competences C_{ref} when injected into the datasets. For the three scenes without a reference annotation available, only the average competence is shown.

3.4. Competence value clamping

The reconstructed soft value for a segment is a weighted average of the annotators labels, using the competence values as weights. For DESED, MACE estimated a majority of the annotators to have a competence value close to zero; this might cause some instabilities in the label reconstruction, if all the annotators for a given segment have very low competences. Furthermore, MACE uses a stochastic method, resulting in fluctuation in the output values. However, the small differences in the competence values can result in unexpectedly large differences when weighting the labels, while, intuitively, if the annotators are equally bad, they should have equal weights.

As an additional experiment, we assume that all low-competent annotators are equally bad, and clamp the competence values of the lowest ranking annotators to a small fixed value. We use 10^{-4} as the competence value, and set it as the competence of the worst 50% and 75% of the annotators. Table 4 shows the comparison between the labels generated from the original competence values and labels generated from the partially clamped competences, as well as using equal weights for all annotators. The standard method is better than not using any weighting, but using the MACE-estimated competences results in a lower F-score than resetting the competences of the lowest ranking annotators, to different degree for DESED/A1 and DESED/A2; furthermore, while combining the DESED/A1 and DESED/A2 annotations shows no benefit with the standard procedure, resetting the lowest competences to the same small value produces the best scoring soft labels. While having more data in DESED/A1+A2 results in a wider distribution of competences and better correspondence with precision, according to Fig. 2, the underlying problem of bad quality labels remains unchanged.

Scene	Original	R50	R75	EQ
DESED/A1	31.2	32.4	34.3	21.8
DESED/A2	31.0	31.9	32.8	22.0
DESED/A1 + A2	29.1	33.8	34.5	21.8

Table 4: Soft macro F-scores for the reconstructed labels and the effect of competence value resetting. In R50 and R75, the lowest-competent 50% and 75% of the annotators, respectively, have competence value reset to 10^{-4}. EQ denotes equal competences.

3.5. Discussion

It is difficult to draw the border between a bad-faith answer and a simple mistake, especially when the task involves human hearing. The problem of bad-quality answers is not platform specific [17], and hence not limited to our experience in using AMT. At the time of our first annotation experiments, there were already discussions about the data quality in paid crowdsourcing [18, 19, 20]. However, in our previous annotation experiments, the amount of low quality work did not hamper significantly the end result quality, unlike now. Based on this work, it seems that MACE is able to identify the annotators producing good quality labels. The problem arises, though, when there are no reliable annotations for a segment, in which case the output annotation ends up having noisy labels.

We speculated that asking annotators' confidence made the annotation somehow annoying or more difficult, causing workers to abandon it after a few HITs. Removing the confidence question indeed decreased the number of annotators who only completed a few HITs and increased the average task count of the workers, but it did not improve the label quality.

4. CONCLUSIONS

This paper presented a detailed analysis of the labels produced by a crowdsourcing process. The approach was to collect temporally strong labels by dividing the work into simpler subtasks of weak labeling, a method previously proven to work. Our conclusion is that the quality of crowdsourced work has worsened considerably, rendering the process unusable. It is hard to pinpoint the reason for this decrease in quality, with potential causes being the influx of workers gaming and exploiting the process, the perceived unfair difficulty/payment ratio of the task, etc. It may be possible to collect sufficiently good labels by simply using more workers, but the process gets prohibitively expensive, driving researchers to return to doing manual annotation themselves.

5. REFERENCES

[1] Y. Zheng, G. Li, Y. Li, C. Shan, and R. Cheng, "Truth inference in crowdsourcing: is the problem solved?" *Proc. VLDB Endow.*, vol. 10, no. 5, pp. 541—552, 2017.

[2] T. Hiippala, H. Hotti, and R. Suviranta, "Developing a tool for fair and reproducible use of paid crowdsourcing in the digital humanities," in *Proceedings of the 6th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Gyeongju, Republic of Korea: International Conference on Computational Linguistics, 2022, pp. 7–12.

[3] K. Hara, A. Adams, K. Milland, S. Savage, C. Callison-Burch, and J. P. Bigham, "A data-driven analysis of workers' earnings on amazon mechanical turk," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, ser. CHI '18. New York, NY, USA: Association for Computing Machinery, 2018, pp. 1—14.

[4] C. C. Marshall, P. S. Goguladinne, M. Maheshwari, A. Sathe, and F. M. Shipman, "Who broke amazon mechanical turk? an analysis of crowdsourcing data quality over time," in *Proceedings of the 15th ACM Web Science Conference 2023*, New York, NY, USA, 2023, pp. 335–345.

[5] N. Pavlichenko, I. Stelmakh, and D. Ustalov, "Crowdspeech and vox DIY: Benchmark dataset for crowdsourced audio transcription," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

[6] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: an audio captioning dataset," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 736–740.

[7] H. Xie, K. Khorrami, O. Räsänen, and T. Virtanen, "Crowdsourcing and evaluating text-based audio retrieval relevances," in *Proceedings of the 8th Detection and Classification of Acoustic Scenes and Events 2023 Workshop (DCASE2023)*, M. Fuentes, T. Heittola, K. Imoto, A. Mesaros, A. Politis, R. Serizel, and T. Virtanen, Eds., 2023, pp. 226–230.

[8] C. Guastavino, *Everyday Sound Categorization*. Springer International Publishing, 2018, pp. 183–213.

[9] R. Serizel, N. Turpault, A. Shah, and J. Salamon, "Sound event detection in synthetic domestic environments," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2020.

[10] I. Martín-Morató and A. Mesaros, "Strong labeling of sound events using crowdsourced weak labels and annotator competence estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 902–914, 2023.

[11] I. Martín-Morató, M. Harju, and A. Mesaros, "Crowdsourcing strong labels for sound event detection," in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2021, pp. 246–250.

[12] D. Hovy, T. Berg-Kirkpatrick, A. Vaswani, and E. Hovy, "Learning whom to trust with MACE," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, L. Vanderwende, H. Daumé III, and K. Kirchhoff, Eds., Atlanta, Georgia, 2013, pp. 1120–1130.

[13] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proceedings of the 22nd ACM International Conference on Multimedia*, ser. MM '14. New York, NY, USA: Association for Computing Machinery, 2014, pp. 1041—1044.

[14] I. Martín-Morató, M. H. Harju, and A. Mesaros, "MAESTRO Real - Multi-Annotator Estimated Strong Labels," Feb. 2023. [Online]. Available: https://doi.org/10.5281/zenodo.7244360

[15] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in *2016 24th European Signal Processing Conference (EUSIPCO)*, 2016, pp. 1128–1132.

[16] M. Harju and A. Mesaros, "Evaluating classification systems against soft labels with fuzzy precision and recall," in *Proceedings of the 8th Detection and Classification of Acoustic Scenes and Events 2023 Workshop (DCASE2023)*, 2023, pp. 46–50.

[17] C. Schild, L. Lilleholt, and I. Zettler, "Behavior in cheating paradigms is linked to overall approval rates of crowdworkers," *Journal of Behavioral Decision Making*, vol. 34, no. 2, pp. 157–166, 2021.

[18] R. Kennedy, S. Clifford, T. Burleigh, P. D. Waggoner, R. Jewell, and N. J. G. Winter, "The shape of and solutions to the MTurk quality crisis," *Political Science Research and Methods*, vol. 8, no. 4, pp. 614—629, 2020.

[19] M. Chmielewski and S. C. Kucker, "An MTurk crisis? Shifts in data quality and the impact on study results," *Social Psychological and Personality Science*, vol. 11, no. 4, pp. 464–473, 2020.

[20] M. Dupuis, K. Renaud, and R. Searle, "Crowdsourcing quality concerns: An examination of amazon's mechanical turk," in *Proceedings of the 23rd Annual Conference on Information Technology Education*, ser. SIGITE '22, New York, NY, USA, 2022, pp. 127—129.