INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

VocalTurk: Exploring Feasibility of Crowdsourced Speaker Identification

Susumu Saito[12], *Yuta Ide*[1], *Teppei Nakano*[12], *Tetsuji Ogawa*[1]

[1]Waseda University, Japan
[2]Intelligent Framework Lab, Japan
{susumu, ide, teppei, ogawa}@pcl.cs.waseda.ac.jp

Abstract

This paper presents VocalTurk, a feasibility study of crowdsourced speaker identification based on our worker dataset collected in Amazon Mechanical Turk. Crowdsourced data labeling has already been acknowledged in speech data processing nowadays, but empirical analysis that answer to common questions such as "*how accurate are workers capable of labeling speech data?*" and "*what does a good speech-labeling microtask interface look like?*" still remain underexplored, which would limit the quality and scale of the dataset collection. Focusing on the speaker identification task in particular, we thus conducted two studies in Amazon Mechanical Turk: *i)* hired 3,800+ unique workers to test their performances and confidences in giving answers to voice pair comparison tasks, and *ii)* additionally assigned more-difficult tasks of *1-vs-N* voice set comparisons to 350+ top-scoring workers to test their accuracy-speed performances across patterns of $N = \{1, 3, 5\}$. The results revealed some positive findings that would motivate speech researchers toward crowdsourced data labeling, such as that the top-scoring workers were capable of giving labels to our voice comparison pairs with 99% accuracy after majority voting, as well as they were even capable of batch-labeling which significantly shortened up to 34% of their completion time but still with no statistically-significant degradation in accuracy.
Index Terms: Crowdsourcing, labeling, voice comparison

1. Introduction

Despite the rise of machine-learning-based speech processing, the value of crowd annotators perceived by researchers has gradually changed to a negative impression nowadays. Some recent literature of the speaker recognition even reported that they already outperformed humans [1], implying that those methods no longer need human annotations that are less accurate than they are. To avoid noises in the data labels, some researchers tend to look for other approaches than public crowdsourcing for data collection since most crowdsourcing platforms are known to hold a certain percentage of workers who performs poorly (*e.g.,* spammers) [2].

Nonetheless, we claim to advocate crowdsourcing [3, 4] as still a realistic and reasonable choice for collecting data labels [5, 6]. Yet even today, we still need accessible ways to collect large datasets [7]; we must keep collecting more data as long as we rely on statistical approaches to solve more complicated problems —such as acoustic event detection [8] and low resource automatic speech recognition (ASR) [9] — especially those which considers domain knowledge, etc.

This paper investigates a couple of aspects of crowdsourced labeling that are left unexplored, thereby preventing researchers from using it. First, we address a question of "*are there workers who can label speech data accurately, and how many of them exist?*" As we assume the lack of such knowledge as the primary barrier of using crowdsourcing, collecting statistics of workers' performances (*i.e.,* accuracy and speed/monetary cost) would help us see through how feasible the crowdsourced speech data labeling would be. Our second question is "*What is a good (and the simplest) microtask design for speech data labeling that makes the most of workers' abilities?*" Not only demonstrating the benefit of crowdsourcing, but also sharing a common practice that can be performed easily for taking advantage of the benefit would be another important point. We therefore aim to provide a simple guideline for designing microtasks which encourages many other speech data labeling scenarios refer to it.

Our study was conducted by posting crowdsourcing microtasks. In Amazon Mechanical Turk (MTurk)[1], we asked workers to answer to a set of **human voice comparison questions** as an example for speech data labeling tasks in this paper. The study involved two steps: *i)* an easy test for worker filtering to locate outstanding workers, and *ii)* a more difficult test for benchmarking worker performance among the outstanding workers. We first hired 3,854 unique MTurk workers by broadcasting microtasks which consisted of twenty-four *1-vs-1* voice pair comparison questions. The result demonstrated that a large percentage of workers provided a fair contribution, with 1,329 workers (34.5%) answering perfectly and 87.6% of workers scoring at least 75% answering accuracy. We also discuss spammers' distribution and their typical behavior and tendency in answering. In the second experiment we additionally posted microtasks to 350+ "outstanding" workers among the hired workers, asking to answer forty-six *1-vs-N* voice set comparison questions, with variations of $N = \{1, 3, 5\}$. Our analysis demonstrated that the time spent for answering questions statistically significantly decreased as N increased while high accuracy was maintained (>90%), which implies the batch processing in speech labeling microtasks would result in better accuracy-time efficiency in workers' labeling jobs.

We believe the contribution of this paper would be to motivate speech researchers who needs a large data labels to consider using crowdsourcing for the data label collection, and to further understand crowd workers[2].

2. Related Work

Examples of crowdsourced speech data analysis [10] include assistance in manually correcting ASR output [11, 12], subjective evaluation of sound quality [13, 14, 15, 16], active learning of acoustic models [17, 18], and so on. The feasibility of crowdsourced annotation has also been investigated, concluding that crowdsourcing can perform well for annotating audio data under the desired conditions [19]. In fact, an ITU-T rec-

[1]https://worker.mturk.com/

[2]Our collected worker responses are publicly available at https://interspeech2021.iflab.tokyo/

http://dx.doi.org/10.21437/Interspeech.2021-464

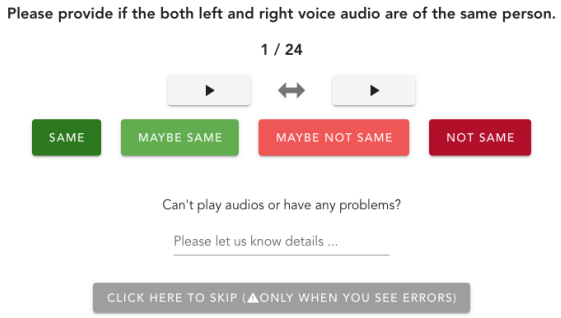

Figure 1: *Microtask interface for worker filtering. Workers were asked to listen to two short voice audios and provide whether they were of same speaker.*

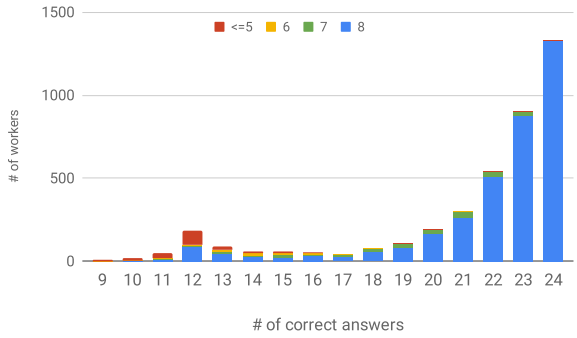

Figure 2: *Histogram of workers by number of correct answers in voice comparison test for worker filtering. Colors indicate number of correct answers for "obvious" questions (i.e., pairs of exactly same audio or voices of speakers of different genders.*

ommendation that assumes the use of MTurk in the subjective evaluation of sound quality has been standardized [16]. The use of crowdsourcing is also being attempted to build models for predicting the quality of human speech perception [20]. In addition, to solve the problem of degradation of annotation quality due to the presence of spammers and low-capacity workers, attempts have been made to evaluate the reliability of crowd workers [21, 22] and to adopt only annotations from reliable workers [17]. This study differs from other studies in that it not only demonstrates the feasibility of pre-selecting promising workers and the appropriateness of using crowdsourcing for annotating audio and acoustic events, but also provides guidelines for efficient user interface design for audio and acoustic event annotation based on a survey of a large number of workers.

3. Worker Filtering

As the first step, we recruited a large number of MTurk workers and asked them to take an easy voice comparison test via a microtask (or "HIT", as called in MTurk in particular). Through the analysis of the collected data, we discuss how many workers do (or do not) have ability of speech data labeling, as well as explore several features they possessed in their answers.

3.1. Settings

In this step, the HIT included two phases of *i)* a short preliminary survey (1–2 minutes; asked workers' age, etc.) and *ii)* a set of 1-vs-1 voice comparison questions (3–4 minutes). As findings for the survey is out of the paper's scope, in below, we only describe the voice comparison part.

Microtask UI Design. The test interface was designed as "1-vs-1" speech data comparison (see Figure 1). A worker is first asked to click two buttons, each of which plays an audio of a few seconds-long human utterance, and then click either of the four buttons —two of them are "*same*" and "*not same*" for stating whether the suggested audio pair was of the same person, and the rest two are "*maybe same*" and "*maybe not same*" for stating answers with their weak confidence. Workers were not allowed to give their answers until they play each audio at least once. Once a HIT is started, each worker was asked to evaluate on 24 pairs in a row. The HIT was automatically submitted when the last comparison pair was finished.

Compared Data. For speech audios, we used public voice data from CMU-ARCTIC speech synthesis databases[3]. We selected four female speakers ("slt", "slp", "eey", and "clb") and four male speakers ("rxr", "rms", "ksp", and "jmk") from all available datasets. The overall numbers of positive ("same") and negative ("not same") examples were divided equally, thus 12 pairs each. For validation purpose, each of both examples contained several "obvious" pairs, which we expect all workers who read instruction can answer correctly: four pairs of the exact same audios in the positive pairs, and four pairs of different genders in the negative pairs. The shown order of the comparison pairs was semi-randomized under constraints of *a)* no consecutive pair involves the same speaker's voice and *b)* neither positive nor negative examples are shown more than three times in a row.

Recruiting Workers. In MTurk, we broadcasted our voice comparison tasks as HITs with no qualification requirement associated. A HIT was priced at $0.80 ($8–$12/hr for 4–6-minute estimated completion time). Workers were instructed that they were allowed to take only one HIT, and that they would have to return if they accepted more than one.

3.2. Results

The experiment was conducted in late February to early March in 2021 and the total spent time was approximately 75 hours. The number of collected valid answers totalled up to 92,496 submitted by 3,854 unique workers, after eliminating records of 179 workers who skipped one or more questions due to system error or spamming actions.

Overall accuracy. Our results indicated that a majority of workers were capable of scoring high accuracy in our voice pair comparison test, while there existed another worker cluster that failed to perform well enough. According to Figure 2, there existed two peaks in the workers' histogram: workers who correctly answered to 24 pairs (100%) had the highest peak with the largest number of 1,329 workers (34.5%), whereas those who only answered to 12 pairs correctly (50%) had the lower peak of 172 workers (4.46%).

The figure also demonstrates, as shown in a red color, a total of 190 workers (4.9%) gave three or more wrong answers out of eight "obvious" questions mostly around the lower peak, who are considered as spammers. Also analyzing workers' confidences together (Figure 3), more workers became less confident in answering to the obvious questions as their overall answering accuracy became lower.

3.3. Discussions

The two peaks in the histogram of Figure 2 imply that there are two types of workers: those who perform carefully and accu-

[3]http://festvox.org/cmu_arctic/

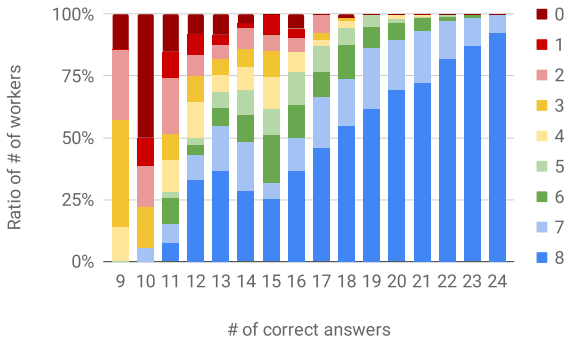

Figure 3: *100% stacked column chart for ratio of workers colored based on how many "obvious questions" they confidently gave answers to. Workers gave less ambiguous answers to obvious questions as their answering accuracy became higher.*

rately, and those who perform poorly (*e.g.,* giving random answers or not understanding instructions); although —by looking at the overall percentage of high-performing workers— collecting crowdsourced labels accurately does seem feasible, the potential spammers cannot be ignored since such a type of workers are known to process a large percent of microtasks in crowd marketplaces [2, 23, 24]. One of reasonable practices would be, as in Figure 3, to filter out workers earlier when they show low confidence in their answers to easy questions.

However, it should also be noted that obvious questions might have not been "obvious" to all workers. The voice pairs of different genders that were included in the four of the obvious questions were chosen by the authors by considering their pitch differences, but that genders make trivial distinction in voices is not always the case. The other four same-audio pairs also might not be easy for some workers, due to the poor audio environment, etc. With these consideration, we cannot define exactly who are spammers and who are not —whatever the worker's attitude is, filtering workers in/out based on the data labeling criteria is possibly the basic and feasible principle.

4. 1-vs-N Task Design Optimization

The second step was also conducted by posting HITs in MTurk, but asking only high-performing workers from the previous test to work on less-easy questions, with more varying interfaces. This experiment aims at benchmarking how accurately crowd workers are capable of labeling speech data in a basic microtask setting, as well as how their performance changes across different microtask interface designs.

4.1. Settings

Microtask UI Design. We designed three microtask interfaces for 1-vs-1, 1-vs-3, and 1-vs-5 voice comparison tasks. The interface for 1-vs-1 comparison is designed similarly to that used in the previous step (as in Figure 1), except that it no longer has "maybe same" and "maybe not same" buttons. A slight modifications were made to the 1-vs-3 and 1-vs-5 comparison interfaces (see Figure 4 for the interface of 1-vs-3 comparison): the button that plays a "target" voice is at the top, followed by three (or five) buttons that play compared voices, each of which is accompanied with a selectable check box for workers to indicate the voice is of the same person as the target voice. Workers are allowed to click 0, 1, or multiple check boxes in each set.

In one HIT, a worker is asked to evaluate voice sets in all the three interfaces in a row (the voice sets or *questions*, is called a

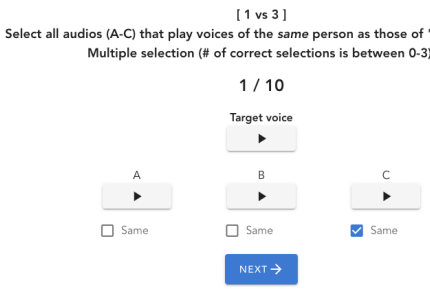

Figure 4: *Microtask interface for 1-vs-3 question batch. Workers were asked to listen to target (reference) voice audio at top, and then listen to three compared voice audios to provide all those which are of same person as that of target audio.*

"question batch" hereafter). Question batches for 1-vs-1, 1-vs-3, and 1-vs-5 involve 30, 10, and 6 voice sets respectively. The three question batches are shown to a worker in a certain order and their questions do not interrupt each other.

Compared Data. Similarly to the previous step, we selected voice audios from CMU-ARCTIC databases. See Table 1 for the list of all audio data used in this experiment. First, we picked six female and six male speakers. For each of the six speakers of a gender, we chose five other utterances of speakers of the same gender from the database; thus created 6×5=30 distinct voice pairs in total. These 30 voice pairs were then used as the 1-vs-1 batch, and were regrouped to create 10 question sets for the 1-vs-3 batch and 6 question sets for the 1-vs-5 batches, under the constraint that every picked speaker is used as the target voice in at least one voice set within the batch. The 30 pairs consisted of 15 positive ("same") and negative ("not same") examples respectively. The number of positive examples in each question in the 1-vs-3 batch ranged from 1 to 5, and that in the 1-vs-5 batch ranged from 0 to 3.

Across all workers, the order of questions was semi-random at inter-batch, and fixed at in-batch. In the first HIT trial, the shown order of the question batches and the speakers' gender during the HIT were determined with round-robin based on the internal worker IDs in our system. The second HIT trial kept the order of the batches, but with questions for the voice sets of the other gender. Each worker was not allowed to work on more than two HITs. This manner enabled us to cancel the order effect caused by the experimental design, to make fair inter-batch and inter-gender comparisons. The in-batch question orders were globally shuffled only once upon production by the authors; in other words, all workers saw the questions in the same order in each batch.

Recruiting Workers. This study was also conducted by posting HITs in MTurk, but with a Qualification requirement associated to allow only the 1,329 "outstanding" workers who

Table 1: *Mapping table between our speaker indices and CMU-ARCTIC database's speaker names and indices.*

Ours		**CMU-ARCTIC's**	
Speaker	Utterance	Gender: Speaker	Utterance
fA, mA	1, 2, 4–6	f: slt, m: aew	a00{11, 12, 14–16}
fB, mB	1–6	f: lnh, m: rms	a00{11–16}
fC, mC	1–5	f: ljm, m: ksp	a00{11–15}
fD, mD	3, 4	f: eey, m: jmk	a00{13, 14}
fE, mE	3, 5, 6	f: clb, m: gka	a00{13, 15, 16}
fF, mF	1–6	f: slp, m: rxr	a00{11–16}

Table 2: *# of workers who completed all comparison sets for both genders, per evaluation order of speakers' gender of provided voices and per N for 1-vs-N question batches.*

Order of N \ gender	male → female	female → male
$1 \rightarrow 3 \rightarrow 5$	29	28
$1 \rightarrow 5 \rightarrow 3$	31	31
$3 \rightarrow 1 \rightarrow 5$	28	31
$3 \rightarrow 5 \rightarrow 1$	30	30
$5 \rightarrow 1 \rightarrow 3$	31	29
$5 \rightarrow 3 \rightarrow 1$	26	30

scored 100% accuracy in the previous worker filtering test. A HIT was priced at $1.80 ($9.00–$13.50/hr for 8–12-minute estimated completion time). Each worker could submit two HITs, but one HIT at a time. Upon completing the second HIT, bonus reward of $1.20 was granted to the worker as an incentive to encourage workers to finish two HITs.

4.2. Results and Discussions

The experiment was conducted in early March in 2021, and the total time spent on the data collection was approximately 32 hours. We collected 32568 answer submissions by 354 outstanding workers who provided valid answers for all comparison sets, with the limitation of "2 HITs (= 92 answers) per worker." Among the workers, per-gender-and-batch numbers of workers varied between 26 and 31 (see Table 2).

Voice Comparison Accuracy. Workers' answering accuracy for 86.7% of all comparison pairs were as high as 90% or above, whereas they struggled to give correct labels to some comparison pairs. The average number of correct answers in a batch was 28.6 pairs out of 30 across all questions and question batches (95.37% average accuracy, $SD = 1.96$). The best accuracy across all questions was 100.0% for the pairs of (mC-3, mA-5) and (fB-2, fC-5), regardless of question batches; presumably, workers could easily tell the difference by the speakers' accents and voice pitches. The worst accuracy was 68.08% scored for (fD-4, fD-3) when asked in the 1-vs-1 batch; although the correct answer is "same", a moderate amount of workers seemed to think the same speaker sounded like a different one.

Majority-Voted Accuracy. We also found the crowdsourced speech data labeling can achieve better accuracy with majority voting, as one of the simplest methods commonly applied for data aggregation [25, 26, 22]. See Figure 5 for the labeling accuracy after applying majority voting across answers posted by 3, 5, and 7 workers. We simulated majority voting by collecting answers posted by multiple workers who have consecutive IDs, for each question batch. In all the question batches starting from 94%–97% overall accuracy without majority voting, the aggregated results could achieve above 97.5% and 99% when asked 3 and 7 outstanding workers, respectively.

Time-Accuracy Efficiency. We found a larger number of compared audios in a question batch would improve workers' time-accuracy efficiency in speech data labeling. Our analysis revealed that the total time workers spent in completing all questions in a batch became statistically significantly shorter regardless of the speakers' gender ($p < 0.001$) as the number of compared voices increased (let spent time for 1-vs-N batch $T(N)$, $T(1)$>$T(3)$ and $T(1)$>$T(5)$ by $p < 0.001$, and $T(3)$>$T(5)$ by $p < 0.01$). See Figure 6 for the average HIT completion times spent per question batch and gender. The best improvement was 33% completion time reduction by the 1-vs-5 question batch for the female voices, where 314.1 seconds required for the 1-vs-1 batch was shortened to 207.2 seconds (both on average). In contrast, our analysis did not show any statistically significant difference in answering accuracy across the batches ($p = 0.53$). This fact implies that high-performing crowdworkers are capable of accurately processing multiple voice pair comparisons at a time, therefore smaller numbers of comparisons in a question may be time-redundant.

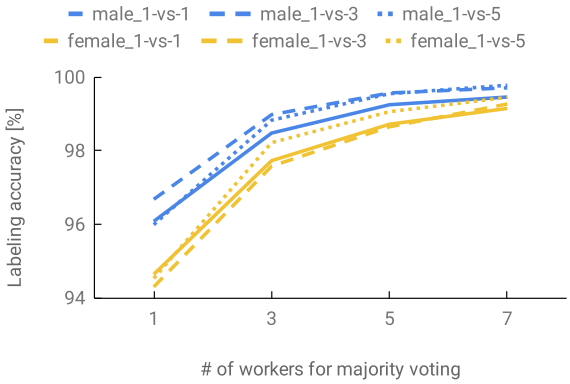

Figure 5: *Labeling accuracy after majority voting with answers posted by different numbers of workers. Even with such simple aggregation method, crowdsourced speech data labeling can achieve as high as 99% or above accuracy.*

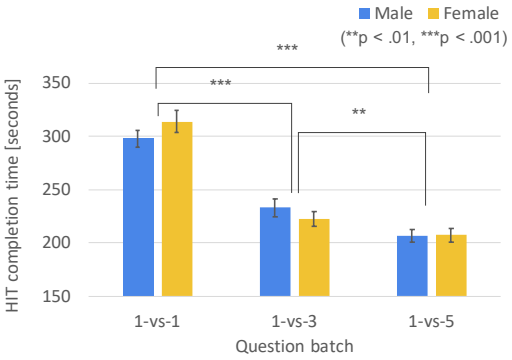

Figure 6: *Workers' average HIT completion times per question batch. HIT completion time became shorter as question batch became larger and all question batches were significantly different from each other, indicating that batch work for voice comparison reduces labeling time per pair.*

5. Conclusions

This paper explored the feasibility of using crowdsourcing for speech data labeling, with our example task of human voice comparison. Our results indicated that a large percentage of workers are supposedly sincere workers whereas some workers were possibly spammers who tended to answer unconfidently. The top-scoring workers could cooperate to obtain ∼99% labeling accuracy with majority voting, and they were even capable of batch-tasking without quality loss. We believe these findings are a reasonable baseline and guideline for speech researchers to crowdsource data labeling.

6. Acknowledgements

This paper is based on results obtained from a project, JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

7. References

[1] J. H. L. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 74–99, 2015.

[2] P. G. Ipeirotis, F. Provost, and J. Wang, "Quality management on amazon mechanical turk," in *Proceedings of the ACM SIGKDD Workshop on Human Computation (HCOMP2010)*, 2010, pp. 64–67.

[3] M. Buhrmester, T. Kwang, and S. D. Gosling, "Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data?" *Perspectives on Psychological Science*, vol. 6, no. 1, pp. 3–5, 2011.

[4] J. de Winter, M. Kyriakidis, D. Dodou, and R. Happee, "Using crowdflower to study the relationship between self-reported violations and traffic accidents," *Procedia Manufacturing*, vol. 3, pp. 2518–2525, 2015, 6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015) and the Affiliated Conferences, AHFE 2015.

[5] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2017)*, 2017, pp. 776–780.

[6] J. Sager, R. Shankar, J. Reinhold, and A. Venkataraman, "VESUS: A Crowd-Annotated Database to Study Emotion Production and Perception in Spoken English," in *20th Annual Conference of the International Speech Communication Association (INTERSPEECH2019)*, 2019, pp. 316–320.

[7] Z. Zhang, N. Cummins, and B. Schuller, "Advanced data exploitation in speech analysis: An overview," *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 107–129, 2017.

[8] T. Virtanen, M. D. Plumbley, and D. Ellis, *Computational Analysis of Sound Scenes and Events*. Springer International Publishing, 2018.

[9] A. R. Syed, A. Rosenberg, and E. Kislal, "Supervised and unsupervised active learning for automatic speech recognition of low-resource languages," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2016)*, 2016, pp. 5320–5324.

[10] M. Eskenazi, G.-A. Levow, H. Meng, G. Parent, and D. Suendermann, *Crowdsourcing for speech processing: Applications to data collection, transcription and assessment*. John Wiley & Sons, 2013.

[11] Y. Gaur, F. Metze, Y. Miao, and J. P. Bigham, "Using keyword spotting to help humans correct captioning faster," in *16th Annual Conference of the International Speech Communication Association (INTERSPEECH2015)*, 2015, pp. 2829–2833.

[12] Y. Gaur, F. Metze, and J. P. Bigham, "Manipulating word lattices to incorporate human corrections," in *17th Annual Conference of the International Speech Communication Association (INTERSPEECH2016)*, 2016, pp. 3062–3065.

[13] F. Ribeiro, D. Florêncio, C. Zhang, and M. Seltzer, "CROWDMOS: An approach for crowdsourcing mean opinion score studies," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2011)*, 2011, pp. 2416–2419.

[14] B. Rainer, M. Waltl, and C. Timmerer, "A web based subjective evaluation platform," in *2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX2013)*, 2013, pp. 24–25.

[15] M. Schoeffler, S. Bartoschek, F.-R. Stöter, M. Roess, S. Westphal, B. Edler, and J. Herre, "webmushra — a comprehensive framework for web-based listening tests," *Journal of Open Research Software*, vol. 6, Feb. 2018.

[16] B. Naderi and R. Cutler, "An open source implementation of ITU-T recommendation p.808 with validation," in *21st Annual Conference of the International Speech Communication Association (INTERSPEECH2020)*, 2020, pp. 2862–2866.

[17] S. Hantke, E. Marchi, and B. Schuller, "Introducing the weighted trustability evaluator for crowdsourcing exemplified by speaker likability classification," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC2016)*, May 2016, pp. 2156–2161.

[18] S. Hantke, Z. Zhang, and B. Schuller, "Towards intelligent crowdsourcing for audio data annotation: Integrating active learning in the real world," in *18th Annual Conference of the International Speech Communication Association (INTERSPEECH2017)*, 2017, pp. 3951–3955.

[19] L. F. Gallardo, R. Z. Jiménez, and S. Möller, "Perceptual ratings of voice likability collected through in-lab listening tests vs. mobile-based crowdsourcing," in *18th Annual Conference of the International Speech Communication Association (INTERSPEECH2017)*, 2017, pp. 2233–2237.

[20] X. Dong and D. S. Williamson, "A Pyramid Recurrent Network for Predicting Crowdsourced Speech-Quality Ratings of Real-World Signals," in *21st Annual Conference of the International Speech Communication Association (INTERSPEECH2020)*, 2020, pp. 4631–4635.

[21] B. Naderi, I. Wechsung, and S. Möller, "Effect of being observed on the reliability of responses in crowdsourcing micro-task platforms," in *2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX2015)*, 2015, pp. 1–2.

[22] B. Naderi, T. Hoßfeld, M. Hirth, F. Metzger, S. Möller, and R. Z. Jiménez, "Impact of the number of votes on the reliability and validity of subjective speech quality assessment in the crowdsourcing approach," in *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX2020)*, 2020, pp. 1–6.

[23] G. Kazai, J. Kamps, and N. Milic-Frayling, "Worker types and personality traits in crowdsourcing relevance labels," in *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM2011)*, 2011, p. 1941–1944.

[24] V. C. Raykar and S. Yu, "Eliminating spammers and ranking annotators for crowdsourced labeling tasks," *Journal of Machine Learning Research*, vol. 13, no. 16, pp. 491–518, 2012.

[25] J. Vuurens, A. P. de Vries, and C. Eickhoff, "How much spam can you take? an analysis of crowdsourcing results to increase accuracy," in *Proc. ACM SIGIR Workshop on Crowdsourcing for Information Retrieval (CIR2011)*, 2011, pp. 21–26.

[26] N. Q. V. Hung, N. T. Tam, L. N. Tran, and K. Aberer, "An evaluation of aggregation techniques in crowdsourcing," in *International Conference on Web Information Systems Engineering*, 2013, pp. 1–15.