Necessity of Voice Sample Selection in Qualification Tests for Crowdsourced Subjective Audio Quality Evaluation*

Takuma Yabe*, Moe Yaegashi*, Teppei Nakano*†, Tetsuji Ogawa*
*Department of Communications and Computer Engineering, Waseda University, Tokyo, Japan
†Intelligent Framework Lab, Tokyo, Japan

***Abstract*—This study examines the impact of voice sample selection in worker qualification tests for subjective audio quality evaluation via crowdsourcing. As synthesized speech quality continues to improve, distinguishing superiority based solely on absolute evaluations, such as the Mean Opinion Score (MOS), has become increasingly challenging. While pairwise comparison provides greater reliability, evaluating a large number of systems and samples remains a challenge, making crowdsourcing a practical approach. However, ensuring reliable evaluations requires screening workers through qualification tests. This study investigates whether the difficulty of evaluating speech samples—quantified by variability in MOS scores—affects the effectiveness of qualification tests. Specifically, we compare qualification tests using easy-to-evaluate samples (consistent MOS scores) and difficult-to-evaluate samples (highly variable MOS scores). Contrary to expectations that difficult-to-evaluate samples would yield better-qualified workers, our experiments, using criteria such as task comprehension, response confidence, and consistency, revealed no significant differences in subjective evaluation performance. These findings suggest that meticulous selection of voice samples for qualification tests may not be necessary, potentially simplifying test design while maintaining evaluation reliability.**

***Index Terms*—Crowdsourcing, qualification test, pairwise comparisons, subjective evaluation of audio quality**

I. Introduction

With recent advancements in speech synthesis technology, generating high-quality speech that closely resembles human vocalizations has become increasingly feasible. Consequently, the comparative evaluation of speech synthesis methods requires greater precision. One of the most widely used approaches for subjective evaluation involves the Mean Opinion Score (MOS) [1], where listeners rate speech samples on a five-point scale. However, since MOS is an absolute evaluation metric, it becomes difficult to distinguish the relative quality of the many high-quality synthesized speech samples available today [2].

To address this limitation, extensive research has focused on using high-precision deep learning models for the automatic and accurate estimation of MOS values [3]–[11]. However, the accuracy of these models is heavily dependent on the quantity and quality of training data (i.e., speech samples paired with MOS ratings), often leading to inconsistencies in system rankings.

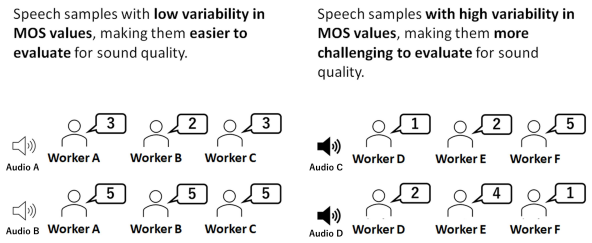

Fig. 1. Method for constructing speech sample pairs in qualification tests for subjective audio quality evaluation: This study investigates whether listening abilities of workers who pass test vary depending on how speech sample pairs are selected—specifically, whether pairs consist of speech samples that are easy to evaluate in terms of audio quality (left figure) or those that are more challenging (right figure). Difficulty of evaluating audio quality is assessed based on variability of pre-assigned MOS values.

In contrast, pairwise comparison, a relative evaluation method that assesses the quality difference between two speech samples, is often considered more reliable, particularly for high-quality samples. However, as the number of systems or samples increases, the number of required comparisons grows exponentially, placing a significant burden on evaluators. To mitigate this challenge, crowdsourcing has been adopted as an efficient means of conducting large-scale evaluations in a short time and at a low cost [12]–[17]. Despite its advantages, crowdsourcing introduces challenges such as insincere or malicious responses aimed at maximizing rewards, as well as inconsistent or biased judgments based on unintended evaluation criteria. These unreliable responses, along with the workers who provide them, introduce noise into the evaluation process, making it necessary to identify and exclude them from analysis [18]–[23].

To address this issue, it is crucial to design qualification tests that effectively screen workers with the necessary listening abilities. While previous research has examined the impact of worker selection criteria on the results of audio quality evaluations [24], little attention has been given to how the selection of voice samples used in qualification tests influences the worker selection process.

This study investigates whether the inherent difficulty of

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

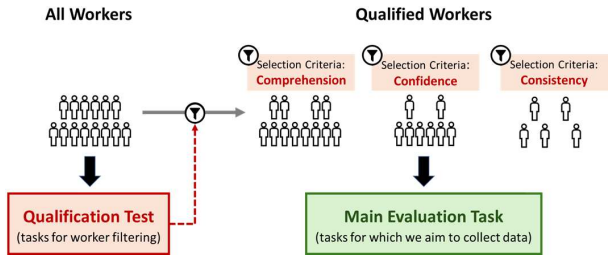

Fig. 2. Overview of crowdsourced subjective evaluation experiment. Qualification test filters participants based on three criteria: comprehension of evaluation criteria, confidence in their responses, and response consistency. These filters can be applied either individually or in combination.

assessing audio quality in the voice samples, which form voice pairs in qualification tests based on pairwise comparisons, affects the outcome of worker selection and, consequently, the reliability of the final subjective evaluations conducted by the qualified workers. Specifically, we utilize speech samples that have been assigned MOS values by multiple evaluators in prior studies and conduct pairwise comparisons using two types of voice pairs, as illustrated in Fig. 1: (1) pairs with high MOS variability (i.e., samples that are expected to be difficult to assess) and (2) pairs with low MOS variability (i.e., samples that are expected to be easier to assess). Intuitively, more challenging samples are expected to lead to the selection of workers with superior listening skills, potentially resulting in observable differences between the aforementioned two conditions. However, if no significant differences are observed, this would suggest that careful selection of voice samples for qualification tests may not be necessary, thereby simplifying task design. In particular, this would allow for the use of speech samples without pre-assigned MOS values. The findings of this study are expected to provide valuable insights into the design of worker qualification tests for crowdsourced subjective audio quality evaluations.

The remainder of this paper is organized as follows. Section II discusses the criteria for selecting workers with the desired listening abilities. Section III describes the user interface presented to workers in the crowdsourcing environment. Section IV outlines the design of the audio quality evaluation experiment and the insights gained. Finally, Section V presents the conclusions of this study.

II. Worker Selection Criteria

As illustrated in Fig. 2, we implement qualification tests designed to select workers with the listening skills required by the requester. Only those who pass the test proceed to the actual evaluation task [25].

The qualification test employs three selection criteria, identified as effective in a preliminary study [24]:

1) **Comprehension of evaluation criteria**: Does the worker accurately understand the purpose of the evaluation, including the requester's intent?
2) **Confidence in responses**: How confident is the worker in their evaluations?
3) **Consistency in responses**: Does the worker provide consistent, non-contradictory answers?

To incorporate these criteria into the pairwise comparison-based qualification test, comprehension of evaluation criteria is assessed by including a sample in the comparison pairs that should receive the lowest score according to the requested criterion (e.g., the most degraded audio when evaluating sound quality). Confidence in responses is measured by requiring workers to report their level of certainty in their evaluations. Consistency is evaluated by presenting the same sample pairs multiple times and verifying whether workers provide consistent answers.

These selection criteria can be applied either individually or in combination. When multiple criteria are used concurrently, only workers whose qualification test results satisfy all of the specified requirements are selected. It is important to note that the selection outcome is independent of the order in which the criteria are applied.

Assessing workers' comprehension of evaluation criteria using the lowest-quality samples effectively filters out those who misunderstand the evaluation criteria. Combining this with confidence measurement facilitates the selection of workers with high listening skills, as they can reliably distinguish between speech samples with substantial quality differences. While assessing consistency further enhances the reliability of selected workers, it is important to note that repeated sample presentations may increase both the cost and duration of the qualification test.

III. Utilization of Crowdsourcing and Its User Interface

Amazon Mechanical Turk (MTurk) was used as the crowdsourcing platform for this study, where each microtask assigned to workers is referred to as a Human Intelligence Task (HIT). Each HIT in this study involved a comparative evaluation of the sound quality of two audio samples generated by different methods. In practical implementations, the qualification test and the task for qualified workers are typically posted as separate HITs. However, in this study, they were combined into a single, continuous HIT to streamline the analysis of the collected responses. Workers were informed in advance that the task would consist of two parts: the first part, corresponding to the qualification test, was labeled "Easy Tasks," while the second part, intended for qualified workers, was labeled "Hard Tasks." They were also notified about the transition point between these phases.

To mitigate potential declines in concentration caused by an increased number of comparative evaluations within a single HIT, the following measures were communicated in advance: *i)* if workers exhibited signs of reduced attention during the Easy Tasks (qualification test), the entire HIT could be rejected; and *ii)* if workers were judged to have completed the Hard Tasks (qualified worker test) with sincere effort, they would receive a one-time bonus of $0.50 as a reward for their diligence.

Fig. 3. Example of user interface presented to workers.

Figure 3 illustrates the user interface (UI) used for pairwise comparisons in the qualification test. Since both the qualification test and the main test for qualified workers were conducted sequentially, the UI displays "[Easy Task]" at the top. However, this label is unnecessary when the tests are administered separately in practical applications. In this UI, workers compare two audio samples, each associated with an English sentence displayed above the playback buttons, and determine which sample has less distortion. Workers first listen to the audio samples by clicking the playback buttons labeled "Voice A" and "Voice B" in the middle section, then select the sample they perceive as having less distortion. In practice, workers choose from options such as "Definitely A," "Maybe A," "Definitely B," or "Maybe B," based on their confidence level. The interface ensures that responses can only be submitted after both audio samples have been played. To address potential playback issues, an issue report form and a skip button are provided at the bottom. The skip button is enabled only if the issue report form has been completed.

For the qualified worker test, the UI remains largely unchanged, except for two modifications: the task label changes from "[Easy Task]" to "[Hard Task]," and the progress indicator updates from "1 / 12" to "1 / 18."

This experiment was conducted using the Tutti framework [25], which facilitates crowdsourcing operations from UI development to task management.

IV. Subjective Audio Quality Evaluation Experiment

A. Objective of Experiment

This experiment investigates how the difficulty of evaluating audio quality in the qualification test affects the effectiveness of worker selection, as reflected in their performance in the subsequent test for qualified workers. The evaluation difficulty of an audio sample is quantified based on the variability in its MOS values assigned by multiple raters. Samples with consistent MOS values are considered easier to assess, whereas those with greater variability are deemed more challenging.

To quantify this, the variance s of the MOS values was computed for each sample, and qualification tests were designed using samples grouped according to their s value:

1) A qualification test using comparison pairs of audio samples with $s \leq 0.5$ (relatively easy-to-evaluate samples).
2) A qualification test using comparison pairs of audio samples with $s > 0.5$ (relatively difficult-to-evaluate samples).

The threshold for s was determined empirically.

It is important to note that this study does not focus on the ease of distinguishing between samples within a pair (i.e., whether there is a significant MOS difference between them) but rather on the intrinsic difficulty of evaluating the individual audio samples. The underlying assumption is that reliable judgments can still be made in pairwise comparisons, even when the individual samples are challenging to evaluate.

For example, if workers selected using more difficult samples (condition 2) demonstrate better performance in the test for qualified workers compared to those selected using easier samples (condition 1), this would highlight the importance of incorporating more challenging samples in the qualification test. However, this would also necessitate an additional step to assess the evaluation difficulty of the samples in advance.

B. Speech Materials

The audio samples for both the qualification and qualified worker tests were sourced from the Voice Conversion Challenge 2018 (VCC2018) [26]. This dataset consists of speech samples generated by various voice conversion systems, each evaluated by up to four raters who assigned MOS values.

It is important to note that raters exhibit variability in how frequently they assign the highest MOS score (5). For instance, some raters frequently assign a score of 5, while others reserve it for only a select few samples. The score of 5 assigned by the latter type of rater carries greater significance. To account for variations in evaluation scales across different raters, the MOS values in the VCC2018 dataset were standardized per listener ID using the transformation $Z_n^{(r)} = (X_n^{(r)} - \mu^{(r)})/\sigma^{(r)}$, where $X_n^{(r)}$ and $Z_n^{(r)}$ denote the original and standardized MOS for sample n rated by rater r, and $\mu^{(r)}$, $\sigma^{(r)}$ are the mean and standard deviation of scores given by rater r. This normalization ensured that all raters' scores had zero mean and unit variance, enabling consistent comparison across raters.

For this study, four systems were selected from VCC2018, including the two with the highest and the two with the lowest average MOS scores. These selected systems were consistently used across both the qualification test and the test for qualified workers to ensure uniformity in evaluation conditions.

C. Qualification Test

Each of the two types of qualification tests, categorized by s, consisted of 12 audio sample pairs generated from four voice conversion systems, as illustrated in Fig. 4.

The qualification test was structured as follows:

1) Comparisons between audio samples generated by the system with the lowest average MOS value and natural speech (to assess comprehension of intent and confidence in responses): 3 pairs.

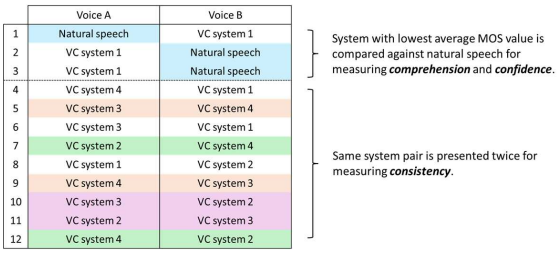

Fig. 4. System pairs used to generate speech samples for qualification test. Speech samples were presented under two conditions: one in which MOS values of individual speech samples in pair vary, and another in which they remain consistent.

2) Comparisons between all possible combinations of the remaining three systems (excluding the system with the lowest MOS value), with each pair presented twice (to assess response consistency): ${}_3C_2 \times 2 = 6$ pairs.
3) Comparisons between the system with the lowest average MOS value and each of the other three systems: 3 pairs.

The three comparison pairs between the system with the lowest average MOS value and natural speech were always presented consecutively at the beginning of the qualification test. Workers who fail this seemingly straightforward comparative evaluation are more likely to exhibit malicious intent. Therefore, the qualification test was designed to promptly identify and exclude such workers from the evaluation process, considering practical deployment.

The worker selection process follows the methodology described in [24]. Specifically, for selection based on comprehension of evaluation criteria, workers who indicated that the natural sound had superior quality in all of the first three comparison pairs were chosen. For selection based on response confidence, workers who answered "Definitely" to all of the first three comparison pairs, regardless of accuracy, were selected. For selection based on response consistency, workers whose agreement rate across all comparison pairs presented twice was 70% or higher were chosen.

D. Test for Qualified Workers

The test for qualified workers was administered to individuals who successfully passed the qualification test. The content of this test remained consistent, irrespective of the conditions applied to the audio samples in the qualification test. Specifically, it required participants to compare all possible pairs of the four voice conversion systems, with each pair evaluated three times, resulting in a total of ${}_4C_2 \times 3 = 18$ comparisons. Notably, the selection of audio samples for this test did not consider the variability in their pre-assigned MOS values.

E. Evaluation Metrics

We collected responses from 521 unique workers, with 261 completing a qualification test using speech samples with $s \leq 0.5$ (i.e., relatively easy-to-evaluate samples), followed by a common test for qualified workers. The remaining 260 completed a qualification test using samples with $s > 0.5$ (i.e., relatively difficult-to-evaluate samples), followed by the same qualified-worker test. Thus, although all workers completed both the qualification and the main evaluation phases within a single continuous HIT, they were divided into two groups according to the type of samples used in the qualification test.

Variance	All Workers (no selection)	Selection Criteria			
		Comprehension	Comprehension & Confidence	Consistency	Comprehension & Consistency
$s \leq 0.5$	**0.400** (261)	**0.592** (124)	**0.707** (68)	**0.701** (41)	**0.829** (26)
$s > 0.5$	**0.416** n.s. (260)	**0.670** n.s. (119)	**0.738** n.s. (58)	**0.662** n.s. (51)	**0.754** n.s. (39)

Fig. 5. Effectiveness of qualification test: Rank correlation values with expert evaluations (shown as upper numbers), along with number of selected workers (indicated in parentheses), for potential difficulty of evaluating speech samples (variance of pre-assigned MOS values) and various selection criteria.

To evaluate the effectiveness of worker selection, we analyzed the results of the test for qualified workers. Confidence-based scoring was applied, where responses marked as "definitely" received 3 points and those marked as "maybe" received 1 point. Systems were then ranked according to their total scores, and these rankings were compared with those derived from expert evaluations (i.e., the ground truth). The correlation between worker-based and expert-based rankings was used as an indicator of the effectiveness of the worker selection process. The ground truth for the superiority of speech pairs was established through consensus between two researchers specializing in speech synthesis.The ground truth for the superiority of speech pairs was established through consensus between two researchers specializing in speech synthesis. While some variability in utterance-level ratings is expected even among experts, system-level rankings are more robust and less susceptible to such fluctuations. Furthermore, the expert rankings were consistent with those derived from the MOS values provided in the original dataset introduced in IV-B, further supporting their reliability as ground truth. Rank correlation values were computed at both the utterance and system levels, with the latter obtained by aggregating evaluation results from the utterance level. However, this study assesses the evaluation abilities of qualified workers based on system-level rank correlation, as it is considered more reliable.

F. Experimental Results

Rank correlations were computed for each worker, averaged based on the selection criteria, and analyzed for significant differences using analysis of variance(ANOVA) [27], a widely used statistical method for comparing group means. The analysis was implemented using the Python library SciPy [28]. Figure 5 presents the rank correlation values and the number of selected workers under different selection criteria—no selection, comprehension level, comprehension level plus confidence, response consistency, and comprehension level plus response consistency—across two conditions of speech sample evaluation difficulty: $s \leq 0.5$ and $s > 0.5$.

The primary analysis revealed no significant differences in subjective evaluation performance (measured by rank corre-

lation values) between the $s \leq 0.5$ and $s > 0.5$ conditions, regardless of the selection criteria applied. This finding challenges the intuitive assumption that workers passing qualification tests with more difficult-to-evaluate speech samples would exhibit superior evaluation performance. Consequently, carefully selecting speech samples for qualification tests may not necessarily improve evaluation outcomes, suggesting that experiments can be conducted without pre-assessing the difficulty of speech sample evaluations. This result has important implications for simplifying the design of qualification tests.

Furthermore, under both the $s \leq 0.5$ and $s > 0.5$ conditions, all selection criteria—task comprehension, response confidence, response consistency, and their combinations—led to significant improvements in rank correlation compared to the no-selection baseline. Notably, selecting workers based on both task comprehension and response consistency achieved the highest listening ability. However, this approach also resulted in a trade-off, reducing the number of selected workers to approximately one-tenth of the original pool.

V. Conclusion

This study investigated whether the inherent difficulty of evaluating individual speech samples in a paired comparison-based qualification test for subjective audio quality assessment affects the test's effectiveness. Experimental results indicated no significant differences in the evaluation performance of qualified workers between conditions involving difficult-to-evaluate and easy-to-evaluate speech samples, regardless of the selection criteria applied. These findings suggest that meticulous selection of speech samples for qualification tests may not necessarily enhance evaluation outcomes, potentially simplifying the design process for such tests.

References

[1] G. Maniati, A. Vioni, N. Ellinas, K. Nikitaras, K. Klapsas, J. S. Sung, G. Jho, A. Chalamandaris, and P. Tsiakoulis, "SOMOS: The samsung open MOS dataset for the evaluation of neural text-to-speech synthesis," in *Proc. Interspeech 2022*, Sep. 2022, pp. 2388–2392.

[2] S. L. Maguer, S. King, and N. Harte, "The limits of the mean opinion score for speech synthesis evaluation," *Computer Speech & Language*, vol. 84, p. 101577, 2024.

[3] Y. Choi, Y. Jung, and H. Kim, "Deep MOS predictor for synthetic speech using cluster-based modeling," in *Proc. Interspeech 2020*, Oct. 2020, pp. 1743–1747.

[4] E. Cooper, W.-C. Huang, T. Toda, and J. Yamagishi, "Generalization ability of MOS prediction networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022)*, 2022, pp. 8442–8446.

[5] S.-W. Fu, Y. Tsao, H.-T. Hwang, and H.-M. Wang, "Quality-Net: An end-to-end non-intrusive speech quality assessment model based on BLSTM," in *Proc. Interspeech 2018*, Sep. 2018, pp. 1873–1877.

[6] W.-C. Huang, E. Cooper, J. Yamagishi, and T. Toda, "Ldnet: Unified listener dependent modeling in MOS prediction for synthetic speech," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022)*, 2022, pp. 896–900.

[7] C.-C. Lo, S.-W. Fu, W.-C. Huang, X. Wang, J. Yamagishi, Y. Tsao, and H.-M. Wang, "MOSNet: Deep learning-based objective assessment for voice conversion," in *Proc. Interspeech 2019*, Sep. 2019, pp. 1541–1545.

[8] B. Patton, Y. Agiomyrgiannakis, M. Terry, K. Wilson, R. A. Saurous, and D. Sculley, "AutoMOS: Learning a non-intrusive assessor of naturalness-of-speech," *arXiv:1611.09207*, 2016.

[9] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, "UTMOS: Utokyo-SaruLab system for VoiceMOS challenge 2022," in *Proc. Interspeech 2022*, 2022, pp. 4521–4525.

[10] W.-C. Tseng, C. yu Huang, W.-T. Kao, Y. Y. Lin, and H. yi Lee, "Utilizing self-supervised representations for MOS prediction," in *Proc. Interspeech 2021*, 2021, pp. 2781–2785.

[11] W.-C. Tseng, W.-T. Kao, and H. yi Lee, "DDOS: A MOS prediction framework utilizing domain adaptive pre-training and distribution of opinion scores," in *Proc. Interspeech 2022*, 2022, pp. 4541–4545.

[12] J. Parson, D. Braga, M. Tjalve, and J. Oh, "Evaluating voice quality and speech synthesis using crowdsourcing," in *Proc. International Conference on Text, Speech, and Dialogue (TSD 2013)*, 2013, pp. 233–240.

[13] S. Buchholz, J. Latorre, and K. Yanagisawa, "Crowdsourced assessment of speech synthesis," in *Crowd-sourcing for Speech Processing: Applications to Data Collection Transcription and Assessment*, 2013, pp. 173–214.

[14] T. M. Byun, P. F. Halpin, and D. Szeredi, "Online crowdsourcing for efficient rating of speech: A validation study," *Journal of Communication Disorders*, vol. 53, pp. 70–83, 2015.

[15] 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 *Proc. 12th International Conference on Quality of Multimedia Experience (QoMEX 2020)*, 2020, pp. 1–6.

[16] B. Naderi, R. Z. Jiménez, M. Hirth, S. Möller, F. Metzger, and T. Hoßfeld, "Towards speech quality assessment using a crowdsourcing approach: Evaluation of standardized methods," *Quality and User Experience*, vol. 6, pp. 1–21, 2020.

[17] B. Naderi, S. Möller, and R. Cutler, "Speech quality assessment in crowdsourcing: Comparison category rating method," in *Proc. 13th International Conference on Quality of Multimedia Experience (QoMEX 2021)*, 2021, pp. 31–36.

[18] A. G. Parameswaran, H. Garcia-Molina, H. Park, N. Polyzotis, A. Ramesh, and J. Widom, "Crowdscreen: Algorithms for filtering data with humans," in *Proc. 2012 ACM SIGMOD International Conference on Management of Data (SIGMOD 2012)*, 2012, pp. 361–372.

[19] S. Jagabathula, L. Subramanian, and A. Venkataraman, "Reputation-based worker filtering in crowdsourcing," in *Proc. 27th International Conference on Neural Information Processing Systems (NIPS'14)*, 2014, pp. 2492–2500.

[20] M. Ashikawa, T. Kawamura, and A. Ohsuga, "Quality improvement by worker filtering and development in crowdsourcing," *Web Intelligence*, vol. 14, no. 3, pp. 229–244, Aug. 2016.

[21] C. Li, V. S. Sheng, L. Jiang, and H. Li, "Noise filtering to improve data and model quality for crowdsourcing," *Knowledge-Based Systems*, vol. 107, pp. 96–103, 2016.

[22] B. Naderi and S. Möller, "Application of just-noticeable difference in quality as environment suitability test for crowdsourcing speech quality assessment task," in *Proc. 12th International Conference on Quality of Multimedia Experience (QoMEX 2020)*, 2020, pp. 1–6.

[23] A. Yamamoto, T. Irino, S. Araki, K. Arai, A. Ogawa, K. Kinoshita, and T. Nakatani, "Effective data screening technique for crowdsourced speech intelligibility experiments: Evaluation with IRM-based speech enhancement," in *Proc. 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 2022)*, 2022, pp. 1405–1411.

[24] M. Yaegashi, S. Saito, T. Nakano, and T. Ogawa, "Do you know how humans sound?: Exploring a qualification test design for crowdsourced evaluation of voice synthesis quality," in *Proc. 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 2022)*, 2022, pp. 1–6.

[25] S. Saito, Y. Ide, T. Nakano, and T. Ogawa, "Vocalturk: Exploring feasibility of crowdsourced speaker identification," in *Proc. Interspeech 2021*, 2021, pp. 1723–1727.

[26] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z. Ling, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2018)*, 2018, pp. 195–202.

[27] R. A. Fisher, *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd, 1925.

[28] P. Virtanen *et al.*, "Scipy 1.0: fundamental algorithms for scientific computing in python," *Nature Methods*, vol. 17, no. 3, pp. 261–272, 2020.