Perceived azimuth-based creditability and self-reported confidence for sound localization experiments using crowdsourcing

Shinnosuke Takamichi^{1,*} and Daisuke Morikawa^{2,†}

¹Graduate School of Information Science and Technology, The University of Tokyo, Japan

²Faculty of Engineering, Toyama Prefectural University, Japan

(Received 10 August 2018, Accepted for publication 22 October 2018)

Keywords: Sound localization, HRTF, Crowdsourcing, Crowdworkers' creditability

PACS number: 43.66.Pn, 43.66.Qp [doi:10.1250/ast.40.142]

1. Introduction

Accuracies of sound localization in binaural listening techniques should be evaluated based on listening tests performed on a large number of listeners. However, in actual tests, only a small number of listeners participate because the experiments are time-consuming. Meanwhile, large-scale collection and evaluation (e.g., [1]) became possible through crowdsourcing platforms. The use of these platforms for experiments in binaural listening is promising, but crowds (crowdworkers) often include low-quality participants who return unnatural answers. To address this problem, we propose two criteria to control the quality of participants: *perceived azimuth-based creditability* and *self-reported confidence*. The experimental results demonstrate that our criteria can identify and remove low-quality participants.

2. Design of sound localization experiments using crowdsourcing

We designed web-based horizontal sound localization experiments that are compatible with any web browser. The web page has an illustration of the head (Fig. 1), buttons to start and finish the evaluation, and a button to give an answer on the perceived azimuth. Participants recruited from the crowdsourcing services press the button to start and listen to the audio samples. The next audio sample will not be played until the participants answer the perceived azimuth. Replaying the sample and changing the answer are not allowed. We do not distinguish in-head and out-of-head sound localizations.

3. Proposed methods for controlling crowdworkers' quality

Because the experiments with crowds allow all kinds of participants and conditions, they include low-quality participants who return unnatural answers. To control participants' quality, we propose two criteria: perceived azimuth-based creditability and self-reported confidence.

3.1. Perceived azimuth-based creditability

One way of estimating participants' quality is by including *gold standard data* when the correct answer is known [2]. But for perceived azimuth, such gold standard data is unobservable. Therefore, assuming that a left-right error

rarely occurs in the in-house experiments, we calculate the left-right error rate and set the creditability score by subtracting the ratio from 1.

3.2. Self-reported confidence

We apply Oyama *et al.*'s method [3] proposed for image labeling task. After conducting sound localization tests, we ask the crowdworkers: "Please estimate the degree of your confidence from 1 (absolutely not confident) to 7 (absolutely confident)." We inform the participants that the reported confidence score would not affect their payment for participating in the tests.

4. Experimental evaluation

4.1. Experimental conditions

Audio samples used in our sound localization tests were 48 kHz-sampled white noise signals convoluted by HRTF (Head Related Transfer Function). The duration of the stimulus was 3 s. A 30 ms linear taper window was applied at the beginning and end of the stimuli. The HRTF of head and torso simulator (4128-C, Brüel & Kjær) was used in this experiment. The distance from the center of the head to the loudspeaker was 1.5 m, and the interval was 30 degrees in the azimuth angle.

The test set had 29 samples that included 24 samples (2 samples per azimuth) for evaluation and 5 samples for user training. The user training was done before the evaluation, and audio samples of head-center, 0, 90, 180, 270 deg. were presented to the participants. The number of participants were 145, and the total number of evaluated samples were 3,480. To recruit the native dialect speakers, we used the crowdsourcing service Lancers*, which is one of the biggest crowdsourcing platforms in Japan. Hired speakers were paid approximately \$0.85 per test. The concordance rate, which means that the presented and the perceived azimuths matched, was used as a localization performance. We also calculated the rate allowed ± 30 deg. errors, except for 0, 90, 180, and 270 deg. of the target azimuths.

4.2. Effects of perceived azimuth-based creditability

We investigated the effectiveness of perceived azimuth-based creditability on localization performance. Figure 2 shows the results of all participants and those who earned a creditability score of 1.0. "All" indicates that all results of all

*e-mail: shinnosuke.takamichi@ipc.i.u-tokyo.ac.jp

†e-mail: dmorikawa@pu-toyama.ac.jp

*http://www.lancers.jp

Fig. 1 Illustration of head. All crowdworkers see this figure and give an answer on his/her perceived azimuth. “左,” “右,” and “あなた” mean “left,” “right,” and “You” in English, respectively.

Fig. 2 Perceived azimuth-based creditability (x-axis) vs. localization performance (y-axis). The number of participants with “confidence score = 1.0” is 117.

Fig. 3 Perceived azimuth-based creditability (x-axis) vs. localization performance (y-axis).

participants were used in calculation. As can be seen, localization performance improved after introducing creditability.

In Fig. 3, creditability and localization performance (with ± 30 deg. errors) are plotted. We can find correlation between the creditability and performance, and low-quality participants can be removed by selecting only the participants who scored 1.0.

Fig. 4 Self-reported confidence scores (x-axis) vs. localization performance (y-axis). The number of participants for each score 1-through-7 is 4, 36, 49, 22, 25, 7, and 2, respectively.

4.3. Effects of self-reported confidence

Similarly, we investigated the effectiveness of self-reported confidence scores on localization performance. Figure 4 shows correlation between the score and performance. When the confidence score changes from 1 to 5, localization performance improves monotonically. However, localization performance with the confidence score of 6 is lower than that with the score of 5. On the other hand, there is no clear relationship between the score and performance allowed ± 30 deg. (However, when the score is 7, the performance allowed ± 30 deg. becomes worse). From these results, we can conclude that extracting participants with a medium confidence score (3-through-5) is effective for quality control.

5. Conclusion

This paper proposed perceived azimuth-based creditability and self-reported confidence criteria for quality control of sound localization experiments using crowdsourcing. We reported that using these criteria, low-quality participants can be extracted from the evaluation results. In the future, we plan to extend this work to other tasks.

Acknowledgements

Part of this work was supported by the SECOM Science and Technology Foundation.

References

- [1] J. Parson, D. Braga, M. Tjalve and J. Oh, “Evaluating voice quality and speech synthesis using crowdsourcing,” in *Proc. TSD*, Pilsen, Czech Republic, pp. 233–240 (2013).
- [2] G. Kazai, J. Kamps, M. Koolen and N. M.-Frayling, “Crowdsourcing for book search evaluation: Impact of HIT design on comparative system ranking,” in *Proc. ISGIR*, Beijing, China, pp. 205–214 (2011).
- [3] S. Oyama, Y. Baba, Y. Sakurai and H. Kashima, “Accurate integration of crowdsourced labels using workers’ self-reported confidence scores,” in *Proc. Int. J. Conf. Artificial Intelligence*, Beijing, China, pp. 2554–2560 (2013).