Comparison of remote experiments using crowdsourcing and laboratory experiments on speech intelligibility

Ayako Yamamoto[1], Toshio Irino[2], Kenichi Arai[3], Shoko Araki[3], Atsunori Ogawa[3],
Keisuke Kinoshita[3], and Tomohiro Nakatani[3]

[1,2] Faculty of Systems Engineering, Wakayama University,
930 Sakaedani, Wakayama, 640-8510, Japan
[3] NTT Communication Science Laboratories,
2-4 Hikaridai, Sekika-cho, Sorakugun,Kyoto, 619-0237, Japan
[1]`yamamoto.ayako@g.wakayama-u.jp`, [2]`irino@wakayama-u.ac.jp`,
[3]`{kenichi.arai.yw,shoko.araki.pu,atsunori.ogawa.gx,`
`keisuke.kinoshita.mb,tomohiro.nakatani.nu}@hco.ntt.co.jp`

Abstract

Many subjective experiments have been performed to develop objective speech intelligibility measures, but the novel coronavirus outbreak has made it very difficult to conduct experiments in a laboratory. One solution is to perform remote testing using crowdsourcing; however, because we cannot control the listening conditions, it is unclear whether the results are entirely reliable. In this study, we compared speech intelligibility scores obtained in remote and laboratory experiments. The results showed that the mean and standard deviation (SD) of the remote experiments' speech reception threshold (SRT) were higher than those of the laboratory experiments. However, the variance in the SRTs across the speech-enhancement conditions revealed similarities, implying that remote testing results may be as useful as laboratory experiments to develop an objective measure. We also show that the practice session scores correlate with the SRT values. This is a priori information before performing the main tests and would be useful for data screening to reduce the variability of the SRT distribution.

Index Terms: speech intelligibility, remote testing, crowdsourcing, speech reception threshold, speech enhancement

1. Introduction

Subjective speech intelligibility experiments provide fundamental information to develop objective intelligibility measures (e.g.,[1, 2]). They have been usually performed in a sound-proof room with well-controlled equipment in a laboratory. However, the novel coronavirus (COVID-19) outbreak has made it very difficult to conduct such formal experiments. One solution is to perform remote testing with sound presentation and response collection using web pages. Although the participants can perform the experimental tasks at any location, it is almost impossible to control the acoustics and listening conditions, including their hearing levels. Hence, control is usually relinquished. Consequently, it is unclear whether the results are entirely reliable. This situation is a serious problem for any psychoacoustic experiments, and some good practices to overcome the issue were reported in [3].

However, from another point of view, remote testing using crowdsourcing is advantageous in collecting massive amounts of data from various participants when the control problem is not very serious. For example, it would be possible to analyze data categorized by listeners' characteristics if the volume of data is sufficiently large. There have been many reports on using remote testing in speech quality assessments [4, 5, 6, 7, 8].

In practice, remote testing has become popular in the quality assessment of text-to-speech synthesis algorithms. Particularly, it seems to be virtually a de facto standard in Interspeech competitions. To improve reliability, methods for data screening were reported in order to reduce variability and eliminate false answers [5]. Notably, there are relatively few studies about remote testing on speech intelligibility assessment [9, 10, 11, 12]. It has not been common to perform remote testing with crowdsourcing because audio control problems and listeners' hearing levels remain crucial issues.

In this paper, we compared speech intelligibility results obtained from remote and laboratory experiments to verify whether remote testing is usable to develop objective intelligibility measures and to identify important factors toward improving the reliability of remote testing results.

2. Experiments: laboratory and remote

We performed web-based remote testing of speech intelligibility. For precise comparison, the speech sounds for remote testing were basically the same as those used in laboratory experiments carried out to develop a new objective intelligibility model, GEDI [2]. We briefly describe the speech materials that were common to the two experiments (see [2] for details) and explain the differences between them.

2.1. Speech materials

The speech sounds used for the subjective listening experiments were Japanese 4-mora words, spoken by a male speaker (label ID: mis), from a database of familiarity-controlled word lists, FW07 [13]. Note that one mora in Japanese roughly corresponds to a vowel or a consonant-vowel (CV) syllable and is written as a single hiragana character, except for some minor examples[14]. The database comprises several word-familiarity ranks corresponding to the degree of lexical information. Speech sounds were obtained from the sound set with which the participants were the least familiar to prevent listeners from complementing their answers with guesses. The dataset contains 400 words per single familiarity, and the average duration of a 4-mora word is approximately 700 ms.

Babble noise was added to the clean speech to obtain noisy speech sounds, referred to as "unprocessed." The SNR conditions ranged from -6 to $+6$ dB in 3-dB steps, and the duration was adjusted to the original speech sound. Two speech enhancement algorithms were applied to the unprocessed sounds. The first was a simple spectral-subtraction (SS) algorithm [15]. With an over-subtraction factor of 1.0, it is referred to as "$SS^{(1.0)}$."

The second one was a Wiener filter (WF) based algorithm that is commonly used in various systems because of its effectiveness with low computational costs. In particular, the WF based on a pre-trained speech model (PSM) [16] was used in [2]. With Wiener gain parameter values of 0 and 0.2, the WF using the PSM is referred to as "$WF_{PSM}^{(0.0)}$" and "$WF_{PSM}^{(0.2)}$," respectively. All noise addition and speech enhancement processes were performed at a sampling rate of 16 kHz, and the final sounds delivered to the listeners were re-sampled to 48 kHz.

2.2. Laboratory experiments

In the laboratory experiments[2], the sounds were presented diotically via a DA converter (OPPO, HA-1) over headphones (OPPO, PM-1) at a sampling frequency of 48 kHz and a quantization level of 24 bits. Sound presentation was controlled using MATLAB in Mac OS X. The sound pressure level (SPL) of the stimulus sounds was 63 dB in L_{Aeq}. These laboratory experiments are referred to as having a moderate SPL. Listeners were seated in a sound-attenuated room with a background noise level of approximately 26 dB in L_{Aeq}.

Fourteen young NH listeners (eight males and six females, aged between 19 and 24 years) participated in the experiments. The participants had a hearing level of less than 20 dB between 125 and 8,000 Hz, and their native language was Japanese. They participated in the experiments only after providing informed consent. The participants were instructed to write down the words they heard using hiragana during a 4-second silent period until the presentation of the next word. The total number of presented stimuli was 400 words, comprising a combination of four speech-enhancement conditions {"Unprocessed", "$SS^{(1.0)}$", "$WF_{PSM}^{(0.0)}$", and "$WF_{PSM}^{(0.2)}$" } and five SNR conditions with 20 words per condition. The total duration of the listening test was about 1 hour. To keep the listeners' attention within a reasonable range, we restricted the maximum number of words to 400 in order to cover all SNR conditions and enhancement algorithms. Each subject listened to a different word set, which was assigned randomly to avoid bias caused by word difficulty.

We also performed complementary low SPL laboratory experiments, 43 dB in L_{Aeq} (i.e., a -20 dB reduction from the above experiment), to estimate the effect of the SPL on speech intelligibility [17]. Another 14 NH listeners participated in the experiments. The speech materials and procedures were essentially the same, except the SNR conditions ranged from 0 to +12 dB in 3-dB steps, since we assumed a decrease in intelligibility due to the low SPL.

2.3. Remote experiments

The remote experiments were performed using web pages that had been newly developed for speech intelligibility tests[18]. To reduce the experiment duration to within 1 hour, we divided the speech-enhancement conditions into two parts: {"Unprocessed", "$SS^{(1.0)}$"} and {"$WF_{PSM}^{(0.0)}$", "$WF_{PSM}^{(0.2)}$"}. Each set consisted of 200 non-overlapping words, i.e., 10 words x 20 sessions. All participants listened to the same word set, since dynamic assignment was unavailable on the web pages.

The participants were instructed to write down the words they heard using hiragana during a 4-second silent period until the presentation of the next word. The answers were filled in on the provided answer sheets (PDF), which had been printed in advance. After listening to ten words (i.e., one session), the participants were required to type the hand-written words into the answer columns on the web page.

The experimental tasks were outsourced to a crowdsourcing service provided by Lancers Co. Ltd. in Japan [19], where,

it is claimed that 100,000 workers are registered, with their personal information, including skills. We recruited 30 participants per experiment without specifying any conditions regarding age, gender, hearing level, and educational background. The only requirements were to use a personal computer and wired headphones or wired earphones, to avoid Bluetooth devices and loudspeakers. As a result, there was a large variety of participants aged from their mid-20s to 60 years old. The first experiment was finished within 2 days. The second experiment opened a few days after. Any worker can participate in the experimental task on a first-come-first-served basis. Since we wanted the same workers to participate in both experiments, we added some advertising phrases about the second experiment at the end of the first one. Sixteen workers participated in both, to a total of 44 participants.

Initially, on the web pages, the participants were required to read information about the experiments before giving informed consent by clicking the agreement button twice in order to be transferred to the experimental task web page. Google Chrome was specified as a usable browser because it plays wav files with 48-kHz and 16-bit properly in both Windows and Mac. The participants set their devices at an easily listenable level.

To familiarize with the experimental tasks, the participants took a training session in which they performed a very easy task using the same procedure as in the test sessions. The speech sounds were drawn from words in the highest familiarity rank with an SNR above 0 dB. After the analysis, it was found that this practice session may play an important role in data screening, as described in section 3.5.

3. Results

The participants' responses in the remote experiments in section 2.3 were compared with the results of the laboratory experiments [2, 17] in section 2.2.

3.1. Data cleansing

The remote experiment data consisted of lists of 4-mora words typed in by the participants. We also collected scanned versions of the hand-written answer sheets to confirm that the answers had been entered correctly and to discourage the workers from cheating by giving irresponsible answers. All data were stored on the local website.

Data cleansing was performed in two steps. The first step involves checking whether the 4-mora words the participants typed-in were entered correctly. This was done using a program, and any errors were corrected in accordance with the tendency of other participants' answers. The endeavor was not very time consuming. The second step is to compare the typed-in and hand-written words for whole answers. Any errors were also corrected, except when the answers were not interpretive. Few corrections were necessary. We also counted the number of hand-written corrections, which were probably made at the end of the session, since the participants were only allowed 4 seconds to write down the words during each listening period. As a result, we found that one participant had corrected their hand-written answers more than 80 times. We assumed that they did not understand the experimental instructions. We therefore excluded their results in both experiments from the analysis. Consequently, data for 29 participants were analyzed for each experiment. The second step took 20 to 30 minutes per participant. Although this was feasible for 30 participants, it would be difficult to accomplish given the large volume of data. However, we found that the following results were fairly unchanged, even without the second step.
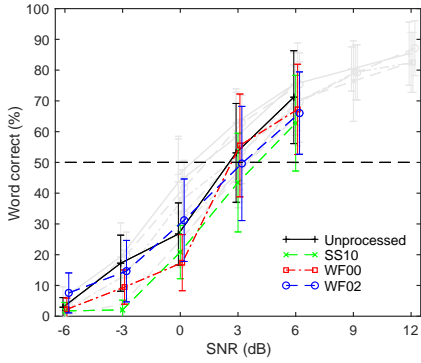
Figure 1: *Crowdsourcing remote experiment. Mean and SD of word correct (%). SS10: "SS$^{(1.0)}$", WF00: "WF$_{\mathrm{PSM}}^{(0.0)}$", WF02: "WF$_{\mathrm{PSM}}^{(0.2)}$".*
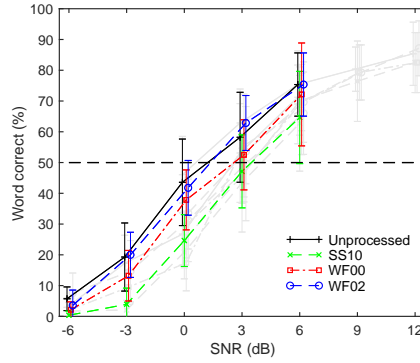

Figure 2: *Laboratory experiment (moderate SPL, 63 dB L_{Aeq})[2]. Mean and SD of word correct (%)*
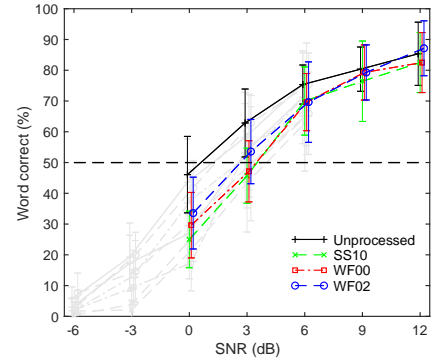

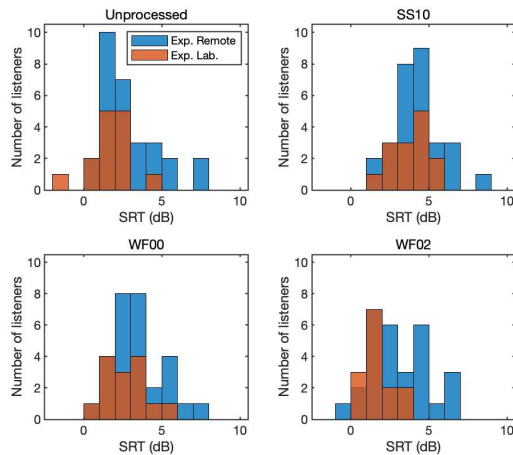Figure 3: *Laboratory experiment (low SPL, 43 dB L_{Aeq})[17]. Mean and SD of word correct (%)*

Figure 4: *Histogram of SRT in dB for four speech-enhancement conditions. Blue: Remote experiment, Red: Moderate SPL laboratory experiment.*

Figure 5: *SRT in dB. Blue: Remote experiment, Red: Moderate SPL laboratory experiment.*: $p < 0.05$, **: $p < 0.01$.

3.2. Psychometric function of speech intelligibility

The psychometric functions of word correct rates were calculated for each speech-enhancement condition as a function of the SNR. Figure 1 shows the results of the remote experiments with 29 participants. The error bar represents the standard deviation (SD) across the participants. Figure 2 shows the results of the moderate SPL laboratory experiments with 14 participants. The faint gray lines behind the colored lines show the results of the other experiments. The lines were lower and the SDs were higher in Fig. 1 than in Fig. 2.

Figure 3 shows the results of the low SPL laboratory experiments with 14 participants. The lines between 0 dB and 6 dB were roughly overlapping, at least within the range of the SD, with those in Fig. 2. As a result, even though the SPL difference was 20 dB, the effect of the SPL on speech intelligibility was not very large. It can be assumed that the listening levels in the remote experiments did not exert a significant effect on the participants' performance.

3.3. Speech reception threshold

Cumulative Gaussian psychometric functions were estimated from the data of the individual participants and the speech-enhancement conditions using a fitting procedure[20]. The speech reception threshold (SRT) is the SNR value where the psychometric function reaches a 50 % word correct rate.

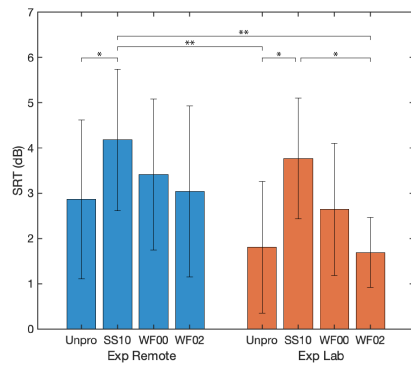Figure 4 shows histograms of the SRT values obtained in the remote experiments in Fig. 1 (blue) and in the laboratory experiments in Fig. 2 (red) for each speech-enhancement condition. The peaks of the histograms were roughly the same, but the SRT distributions were extended to more than 5 dB in the remote experiments. This is consistent with the higher SDs in Fig. 1. The larger variability may correspond to the diversity of the participants, whereas the participants in the laboratory experiments were restricted to young university students.

Figure 5 shows the mean and SD values of the SRTs dependent on the speech-enhancement conditions for the two experiments. Two-way analysis of variance (ANOVA) showed that there were significant main effects in the speech-enhancement conditions and the two experiments, but the interaction was not significant. Multiple comparison analysis showed that there were significant differences ($p < 0.05$) between "Unprocessed" and "SS$^{(1.0)}$," but not between "SS$^{(1.0)}$" and WF$_{\mathrm{PSM}}^{(0.2)}$ in the remote experiments. In contrast, there were significant differences for both cases in the laboratory experiments. This is the main difference between the two experiments. There are two significantly different conditions ($p < 0.01$) across the two experiments, although the meaning of this is not easily interpreted. The important issue is that there were no significant differences between the other combinations.

The variations in the SRTs across the speech-enhancement conditions were similar between the two experiments. When developing and verifying a new objective speech intelligibility model, as in the case of GEDI [2], "Unprocessed" was used as the reference condition to fix the parameter values, and the other speech-enhancement conditions were used to evaluate prediction performance. In this context, the results of the remote
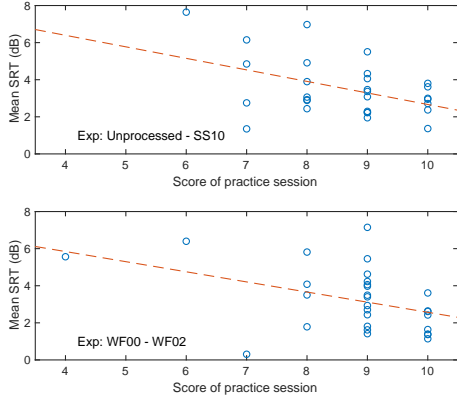
Figure 6: *Relationship between the practice session scores and the mean SRT (dB) with data (circles) and regression lines (dashed lines). Upper panel: First remote experiment; Lower panel: Second remote experiment*

experiments could be usable, as well as the laboratory experiments.

3.4. Prediction of SRT

To survey the factors influencing individual participants' SRT values, we performed stepwise regression analysis using generalized linear models. The target variable was the mean SRT of two speech-enhancement conditions, i.e., "Unprocessed" and "$SS^{(1.0)}$" in the first remote experiment and "$SS^{(1.0)}$" and $WF_{PSM}^{(0.2)}$ in the second remote experiment. We collected nine explanatory variables pertaining to participants' characteristics from the data registered on the crowdsourcing site and the experimental procedure: (1) age, (2) gender, (3) listening device (headphones or earphones), (4) reliability estimated from consistency of ID registration, (5) number of corrections of handwritten words, (6) number of inconsistencies between handwritten and typed-in words, (7) number of corrections due to a different mora count, (8) word correct score in the practice session (where 10 is a perfect score), and (9) duration of experiments.

The stepwise procedure, with a simple linear regression model, yielded simple equations as a function of the practice session scores. The equations for the first and second remote experiments were:

$$\text{SRT(dB)} = 8.88 - 0.63 \times \text{score} \quad (p = 0.015), \quad (1)$$
$$\text{SRT(dB)} = 8.03 - 0.55 \times \text{score} \quad (p = 0.023). \quad (2)$$

The linear models were significantly different from the constant models, and the coefficients were very similar. The other factors were ruled out. The result implies that the practice session score is the only factor influencing the SRT values.

Figure 6 shows the SRT values and the regression lines. There was a clear tendency of negative correlation. The prediction errors were about 1.4 dB and 1.6 dB, which are not very small but are comparable with the SDs shown in Fig. 5.

3.5. Data screening

The practice session score is a priori information derived before performing the main tests. This is particularly important to judge whether the main tests are worth executing to obtain useful information. The score could be useful for data screening. If the practice session score is low, it may be that the participants did not fully understand the experimental procedure or they had difficulty filling in the words during the 4-second intervals of silence. There may be other reasons.
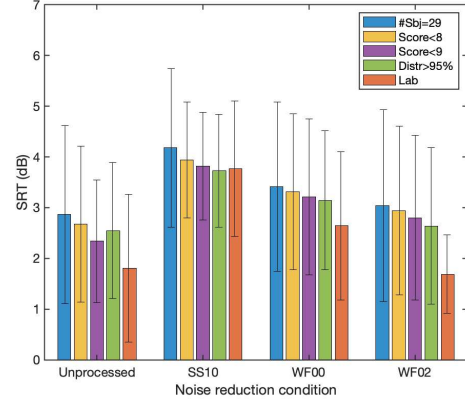

Figure 7: *Effect of data screening on the SRTs. Remote and laboratory experiments shown in Fig. 5 (blue and red); Data elimination by the practice session score less than 8 (yellow) and less than 9 (purple), and by cumulative Gaussian distribution greater than 95% (green).*

We evaluated the effect of data reduction using the practice session scores. Firstly, the data of those participants who scored less than 8 were eliminated. Consequently, participant numbers were reduced to 24 and 26 in the first and second remote experiments, respectively. The result is shown in the yellow bars of Fig. 7. The mean and SD values of the SRTs reduced slightly compared to those in the original data with 29 participants (blue bars). Secondly, the data of those participants who scored less than 9 were eliminated. This reduced participant numbers to 17 and 22 in the first and second experiments, respectively. Again, the mean and SD of the SRTs dropped further but did not reach the level of the laboratory experiments (red bars).

It is also possible to reduce the data after inspecting the response distribution. This seems to be common practice in sound quality assessment [5]. Initially, a Gaussian function was fitted to the SRT values shown in Fig. 4, since a t-distribution with a degree of freedom of 28 is sufficiently close to the Gaussian. The samples above 95% of the cumulative Gaussian distribution were then eliminated. The results are shown by the green bars in Fig. 7. The mean and SD values of the SRTs reduced slightly but, again, did not reach the level of the laboratory experiments (red).

Consequently, it was difficult to select the remote data to be close to the laboratory data, probably because the two populations were different. However, it is worth noting that the a priori information about the practice session score works as well as the posteriori information about the distribution of the results.

4. Summary

In this study, we compared speech intelligibility results obtained in remote and laboratory experiments. Although the mean and SD of the SRT of the remote experiment were higher than those of the laboratory experiments, the variation in the SRTs across the speech-enhancement conditions was very similar between them. This implies that results of remote experiments may be usable to develop objective intelligibility measures, in addition to those of laboratory experiments. We also found that the a priori information about the practice session scores was useful for data screening to reduce the variance of the SRT.

5. References

[1] T. H. Falk, V. Parsa, J. F. Santos, K. Arehart, O. Hazrati, R. Huber, J. M. Kates, and S. Scollie, "Objective quality and intelligibility prediction for users of assistive listening devices: Advantages and limitations of existing tools," *IEEE signal processing magazine*, vol. 32, no. 2, pp. 114–124, 2015.

[2] K. Yamamoto, T. Irino, S. Araki, K. Kinoshita, and T. Nakatani, "Gedi: Gammachirp envelope distortion index for predicting intelligibility of enhanced speech," *Speech Communication*, vol. 123, pp. 43–58, 2020. [Online]. Available: https://doi.org/10.1016/j.specom.2020.06.001

[3] "Remote testing wiki." [Online]. Available: https://www.spatialhearing.org/remotetesting/

[4] S. Buchholz and J. Latorre, "Crowdsourcing preference tests, and how to detect cheating," in *Interspeech 2011, Twelfth annual conference of the international speech communication association*, 2011.

[5] 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 (ICASSP)*. IEEE, 2011, pp. 2416–2419.

[6] B. Naderi, T. Polzehl, I. Wechsung, F. Köster, and S. Möller, "Effect of trapping questions on the reliability of speech quality judgments in a crowdsourcing paradigm," in *Interspeech 2015, Sixteenth annual conference of the international speech communication association*, 2015.

[7] R. Z. Jiménez, L. F. Gallardo, and S. Möller, "Influence of number of stimuli for subjective speech quality assessment in crowdsourcing," in *2018 Tenth international conference on quality of multimedia experience (QoMEX)*. IEEE, 2018, pp. 1–6.

[8] B. Naderi and R. Cutler, "An open source implementation of itu-t recommendation p. 808 with validation," *arXiv preprint arXiv:2005.08138*, 2020.

[9] M. Cooke, J. Barker, M. L. G. Lecumberri, and K. Wasilewski, "Crowdsourcing for word recognition in noise," in *Interspeech 2011, Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[10] M. Cooke, J. Barker, M. G. Lecumberri, and K. Wasilewski, "Crowdsourcing in speech perception," *Crowdsourcing for speech processing: Applications to data collection, transcription and assessment*, vol. 137, p. 172, 2013.

[11] A. Paglialonga, E. M. Polo, M. Zanet, G. Rocco, T. van Waterschoot, and R. Barbieri, "An automated speech-in-noise test for remote testing: Development and preliminary evaluation," *American Journal of Audiology*, vol. 29, no. 3S, pp. 564–576, 2020.

[12] A. Padilla-Ortiz and F. Orduña-Bustamante, "Binaural speech intelligibility tests conducted remotely over the internet compared with tests under controlled laboratory conditions," *Applied Acoustics*, vol. 172, p. 107574, 2021.

[13] K. Kondo, S. Amano, Y. Suzuki, and S. Sakamoto, "Ntt-tohoku university familiarity-controlled word lists 2007 (fw07)," 2007. [Online]. Available: http://research.nii.ac.jp/src/en/FW07.html

[14] "Wiki: Mora_(linguistics)," last: 17 Mar 2021. [Online]. Available: https://en.wikipedia.org/wiki/Mora_(linguistics)

[15] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4. Institute of Electrical and Electronics Engineers, 1979, pp. 208–211. [Online]. Available: http://ieeexplore.ieee.org/articleDetails.jsp?arnumber=1170788

[16] M. Fujimoto, S. Watanabe, and T. Nakatani, "Noise suppression with unsupervised joint speaker adaptation and noise mixture model estimation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 3 2012, pp. 4713–4716. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6288971

[17] R. Iwaki, "The effect of sound level reduction on speech intelligibility (in japanese)," 2019, bachalor thesis, Faculty of Systems Engineering, Wakayama University.

[18] "Speech intelligibility test web page (in japanese)," last: 17 Mar 2021. [Online]. Available: https://web.wakayama-u.ac.jp/~irino/ExpWeb/ExpSpIntel/index.php

[19] "Lancers co. ltd." [Online]. Available: https://www.lancers.jp

[20] F. A. Wichmann and N. J. Hill, "The psychometric function: I. fitting, sampling, and goodness of fit," *Perception & psychophysics*, vol. 63, no. 8, pp. 1293–1313, 2001.