Challenges on level calibration of online listening test: a proposed subjective method

Léopold Kritly[1, 2], Vincent Basecq[2], Christ Glorieux[3], Monika Rychtáriková[1, 4]

[1] KU Leuven, Faculty of Architecture, Hoogstraat 51, 9000 Ghent / Paleizenstraat 65, 1030 Brussel, Belgium
{Leopold.Kritly; Monika.Rychtarikova}@kuleuven.be

[2] EPF-Graduate School of Engineering, 3 bis rue Lakanal, F-92330 Sceaux, France
Vincent.Basecq@epf.fr

[3] KU Leuven, Department of Physics and Astronomy, Laboratory of Acoustics - Soft Matter and Biophysics,
Celestijnenlaan 200D, 3001, Heverlee, Belgium
Christ.Glorieux@kuleuven.be

[4] STU Bratislava, Faculty of Civil Engineering, Department. Of Architecture, Radlinského 11, 81005, Bratislava,
Slovak Republic

Abstract

In addition to many disruptive consequences in society, the COVID-19 pandemic has also posed challenges on experimental research. The resulting limitations on gatherings of people impeded the attendance or participation of human subjects in experiments. In the context of subjective assessment of sound stimuli by people, listening tests in a laboratory could in principle be replaced by online listening tests, which are moreover more easy to organize for larger amounts of subjects. However, in case of online presentation of sounds, the test environment is not controlled and different apparatuses can introduce a bias in the results. For listening tasks involving sound source localization, compared to loudspeakers, the use of headphones and auralization of sounds taking into account the Head-Related Transfer Function (HRTF) are beneficial.

Some psychometric listening tests require a particular excitation level in order to guarantee the consistency of the results over different test people and conditions. When a listening test is offered online, then the listening people typically do not have measurement tools around for reliable quantitative level calibration. The question is whether a subjective calibration method could be developed, which is based on the possible ability of a listening person to equalize a given stimulus to a defined level, based on his or her acoustic memory.

In this work, an unsupervised subjective method for level calibration of online presented sound has been investigated on 17 test persons, using pre-recorded speech of a female speaker as a reference signal. The subjectively iterated level was then determined by making use of calibrated reference headphones. The accuracy of the proposed method relies on the classification of the participant practice in terms of speech loudness using a survey prior to the test procedure. The described procedure, which is easy to implement and requires only a few minutes, was found to yield a prediction accuracy of ± 3.8dB.

Keywords: Psychoacoustic, perception, listening test, level calibration, online experiment.

1 Introduction

The COVID-19 pandemic has impacted our society at multiple levels [1], [2] including experimental and behavioural research. In the field of perceptional acoustics research, the conduction of listening tests has often been suspended to limit the propagation of this pandemic. As a result, listening procedures were needed to be performed differently, often by remote participation to ensure contact-free experiments. Online platforms [3] have been a popular option [4]–[7] in this research field, offering a safe approach to gather subjective data, as well as an alternative to in-laboratory experiments. A number of platforms allow the development, hosting and sometimes even recruitment of behavioural experiments including Pavlovia [8], [9], Amazon’s Mechanical Turk [10] especially through the open-source framework psiTurk [11], WebExp [12], Gorilla[13], jsPsych [14], Lab.Js [15], and Worldlikeness [16]. Such test environments also provide a way to disseminate a listening procedure to a larger [17] and more diverse audience, as they are no longer limited to the people living around the research facilities. Everyone could potentially attend, including participants from other countries [18].

Recent studies tackle the concern regarding the quality of datasets acquired through online experiments provided that a careful and suitable design of the experimental task is implemented [19], [20]. However, a controlled environment cannot be guaranteed within online listening experiments, thus not systematically assuring a defined background noise, level of the excitation stimuli, type and placement of headphones or speakers used.

Several types of listening tests require a particular level of presentation of stimuli, due to the level-dependence of the auditory perception. These experiments are often, but not exclusively related to the discrimination of spectral coloration, i.e. the changes in frequency distribution between stimuli, which, due to the non-linear perceived frequency response of the human auditory mechanism [21], [22], is level dependent, as demonstrated by the equal loudness curves [23].

In this context, in online experiments, where test persons are not equipped with pre-calibrated listening devices or calibrators, it is challenging to accomplish presentation of stimuli at a certain level of excitation. Typically, reports on research that made use of online tests mention this complication, but they do not mention how the issue was tackled. In some speech-related studies, participants were requested to calibrate the volume of the experiment by matching a defined speech to either a comfortable level [24]–[26] or to their own convenience [27]. The latter case consists of adjusting the level of speech stimuli to sound natural, i.e. to commonly encountered vocal levels, without providing specific conditions to help visualizing a reference level.

In this work, we have developed a subjective method of level calibration of listening tests using a pre-recorded speech and assess its consistency over a defined test population. In the following, first the method is presented. Next, the experimental conditions are depicted. The performance of the method in terms of targeting a certain sound pressure level of presented stimuli proposed method is discussed, and a perspective is given on potential refinements.

2 Proposed calibration method

The concept of the proposed calibration method is to use a sound known and employed on a daily basis by the potential participant to a listening test. The stimulus should have a sound pressure level relatively constant under given conditions to be used as a reference. Speech satisfies this condition in case of a low background noise exposure [28]. In this study, the average sound pressure level of speech was assumed to be 55dBA. This level lies within the optimum level range of speech to minimize the listening difficulty in quiet environments [29]. The sound pressure level of speech in case of noise exposure is influenced by the Lombard effect, inducing a rise of the voice level and pitch related to the level of the exposed noise [30].

Other studies have shown a similar vocal level when exposed to a soft background noise (around 50dBA) [31], [32] varying from 57dBA to 62dBA.

Speech fragments are characterized by relatively constant sound pressure levels, and are good candidates to be used as calibration signals of which the level can be recalled from people's auditory memory. The process of remembering the level of a typical speech introduces some inaccuracy in the calibration level, but is expected to remain within an acceptable range.

3 Experimental conditions

3.1 Apparatus and participants

The assessment of the proposed calibration method has been performed under two distinct experimental conditions.

The first test session was conducted in a 125m^3 semi-anechoic chamber. The participants were seated at a desk disposed in a corner of the room. The stimuli were digitally broadcasted from a desktop computer located outside of the anechoic room using a Scarlett 6i6 (Focusrite®) patched to a listening test unit HPS IV (Head Acoustics®) using SPDIF protocol. The stimuli were generated by open-back headphones HA II.1 (Head Acoustics®) connected to the listening test unit.

The second test session was performed in a classroom without any significant noise exposure. To avoid noise interfering with the test, the experimental sessions were scheduled outside the school breaks. The listening equipment was composed of a Scarlett 8i6 3rd Gen (Focusrite®) and high-end open-back headphones HD650 (Sennheiser®).

Two different setups have been used for practical reasons. They had similar performance and were calibrated with pink noise using a dummy head HMS III (Head Acoustics®) in a semi-anechoic room. Headphones were used in both cases as the related listening tasks involved some source localisation for whom these devices are beneficial in combination with the use of the adequate Head-Related Transfer Function (HRTF).

Both devices were operated at a sampling frequency of 48kHz with 16bits depth. A computer monitor was used to display the graphical interface under the two conditions. The participant interacted with the interface by means of a standard computer keyboard and a silent mouse (Logitech® M220).

The listening procedure was conducted on 17 sighted people, among which there were 9 women and 8 men from 20 to 59 years old. All participants volunteered to conduct this experiment and prior to the test had given an oral informal consent to use their contribution within a research context. No compensation was given to any participant. The hearing performance of the test people was not assessed. However, they were asked if they suffered from any hearing impairment or were often exposed to loud noise/sounds. The experiment also kept track of participants speaking at a particularly loud level. In the case where the participants had any hearing impairment or experience with noise exposure, they were labelled under a “special” category. The participations were anonymized and classified with an arbitrary number.

Six participants (P[1: 4, 6: 7]) used the first apparatus and 11 people (P[5, 8: 17]) used the second one. Some participants performed multiple tests on different days to assess the consistency of this test within a subject as well as to increase the statistical samples of this study. The participant labelled as P1 corresponds to the designer of this experiment. Participants P[1: 7] had some experience performing listening tests. The others were novice participants.

Participants P[13: 17] have been included in the special group, corresponding to the people with potential underestimation of speech level due to either a hearing related issue or an abnormal voice level.

These participants were included in this group for the following reasons:

- P13: Participant speaking loudly,
- P14: Aged person with known mild hearing loss (< 40dBHL),
- P15: Person with known hearing loss, having a history of attending loud events,
- P16: Participant working with loud equipment, often listening to loud music and sounds with closed-back headphones,
- P17: Person wearing earbuds before the test with significantly audible music, suggesting a listening experience at high sound pressure level.

3.2 Stimulus

The stimulus intended to be used as a calibration signal was a recorded speech fragment spoken by a Belgian female speaker who was fluent in English. This speech was recorded with a lapel microphone (Sennheiser® MKE-2P) placed 5 cm in front of the speaker's lips and mounted on a fitness headset support (Samson® Qe). The speaker was seated at a table of a quiet office. The recited text fragment was taken from the abstract of a biography of Katherine Johnson [33] in English. The speech level was not equalized between sentences, in order to keep organic features of the recording, therefore sounding natural to the listener. The complete recording was 66 seconds long and was split into 2 parts to avoid loss of attention span or even irritation of the participant by listening to the same sound for a repeated number of times, required to assess the validity of such method.

The duration of the pauses between sentences was kept as original and varied from 0.4 to 0.8s.

Most of the energy of the speech was located between 200Hz and 2.8kHz (-26dBFS, i.e. 95% of the sound pressure) as seen in **Figure 1**. The spectrum contained several peaks in the range between 200 and 700Hz, which could be assimilated as main formants of this speech and are respectively located at 200Hz, 400Hz, 575Hz and 700Hz.

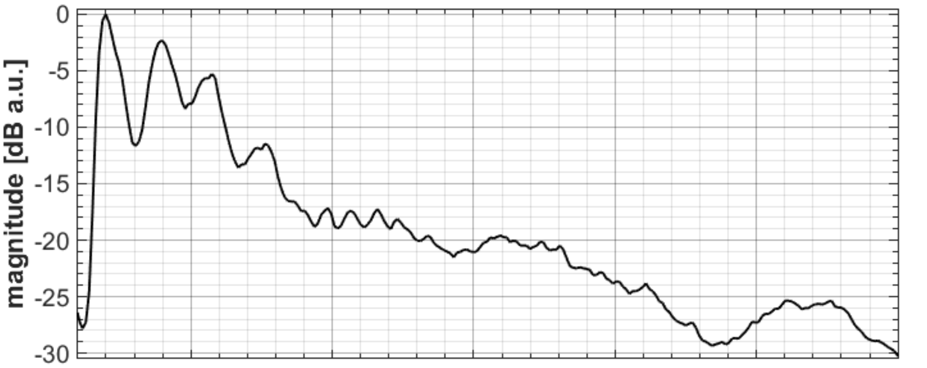

Figure 1. Spectrum of recorded speech fragment recited by a female speaker. The spectrum has been computed by averaging a sequential section of 4096 samples (i.e. 85ms) windowed by a Hanning frame with a 50% overlap ratio by means of a ssfft. The spectrum is displayed from 100 Hz to 3kHz where 95% (-26dBFS) of the sound pressure is located. The low frequencies (<100Hz) have not been displayed, mostly consisting of undesired noise induced by tiny microphone movements and frictions with its stand.

3.3 Test procedure

The assessment of the calibration level procedure was made by requesting the participants to equalize the sound pressure level of the recorded speech to correspond to a typical conversation level in a quiet environment. We asked them to picture themselves talking to someone at around 1.5m, as if the two speakers were seated on opposite sides of a table during any social event. It was mentioned to the participant that human beings tend to speak louder wearing a face mask. The participants were advised to perform the equalization level with a reference baseline corresponding to a conversation without a facemask. The equalization was performed by tuning the volume of the operating system as well as applying a correction factor to the amplitude of the signal. These changes in volume were made in real-time through a custom graphical interface computed in Matlab®. The participants were asked to first perform their equalization by changing the volume of the operating system and, if necessary, fine-tune the amplitude of the speech with the correction factor. This task was performed 20 times on either part of the recorded speech whose amplitude was attenuated by a randomly defined factor. This factor was computed at each iteration of the calibration performed by a participant within the test interface, thus generating a random sequence of the presentation amplitude of the speech. This factor could only take values between [-20, 0]dB. The interface reset the volume of the operating system as well as the correction factor to its default values to avoid exposing the participant to potentially loud speeches which could cause a temporary rise of the hearing threshold of the subject. The program kept track of the emission levels of each test iteration as well as the volume parameters defined by the participant. The completion of this test took between 7 and 14 minutes across participants.

4 Results

As discussed in section 3.1, the participants were split into two groups, one with known hearing problems or excessive noise exposure labelled as “special” and the other labelled as “normal”.

For both groups, the participant having repeated the experiment in separated sessions had consistent results, their average estimated value being similar between sessions as seen in **Figure 2** (P1, P2, P4), except for P13 (cf. **Figure 3**) whose average estimation differs by 4dB. This participant had reported having calibrated the speech to a quieter level on the second session.

The participants were consistent within a test session, their averaged standard deviation was about 1.9dB for the normal group and 2.6dB for the special group.

The estimated level of the speech considerably varies between participants of the normal group mostly due to the overestimation of the speech level for P5, P6 and P12. The average calibration level of the recorded speech of this group was 52.3dBA (std = 3.8dBA). The variance between participants is audible, but nevertheless small enough for online tests in which the level is not extremely critical.

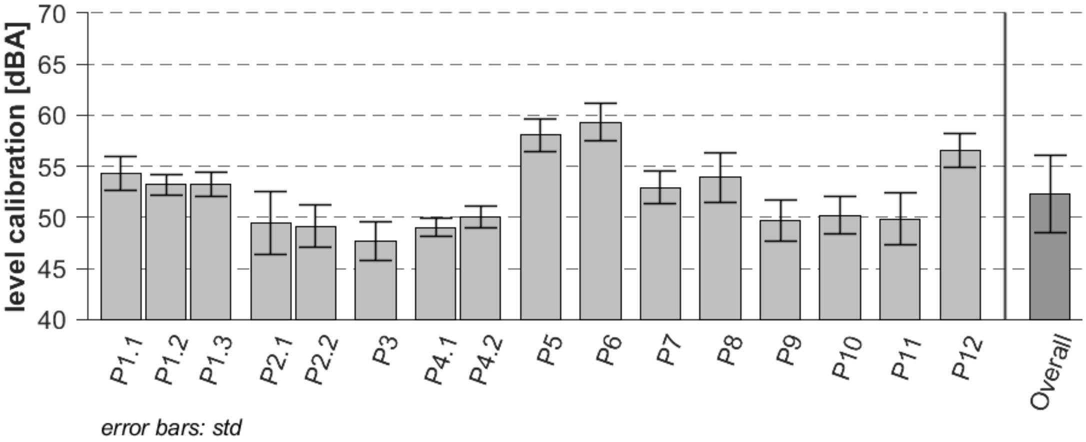

Figure 2. **Average level calibration of recorded speech over the normal group of participants.** Different sessions of the same participant are labelled by adding an identifier after the name participant (eg. P1.1, P1.2). The average level calibration over all participants and test sessions for the normal group is about 52.3dBA (std = 3.8dBA).

For the special group, the variance of estimation within a listening session was slightly larger than for the normal group, and remain too large to make statistically significant conclusions. The average level calibration of the recorded speech of this group was 60.9dBA (std = 3.4dBA).

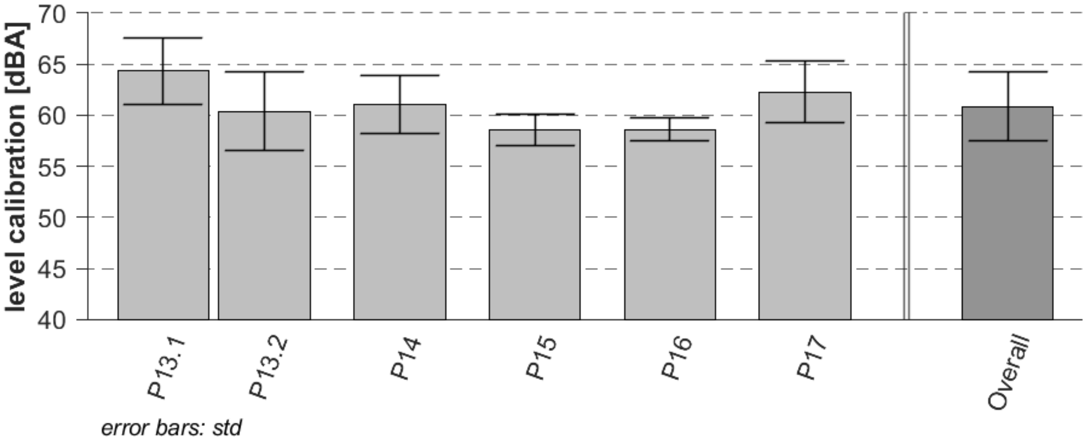

Figure 3. Average level calibration of a recorded speech over the special group of participants. Different sessions of the same participant are labelled by adding an identifier after the name participant (eg. P1.1, P1.2). The average level calibration over all participants and test sessions for the special group is about 60.9dBA (std = 3.4dBA).

5 Discussion

Calibrating an experiment using a subjective method with speech proved to be an approach worth considering for experiments that are not too critically sensitive to level differences. The proposed calibration procedure was designed to investigate audibility features in human echolocation, in which case the performance is only weakly sensitive to small changes in loudness [34], [35]. The differentiation of the participants based on their hearing performance and noise experience seems to provide a more accurate calibration for these participants. The results on the special group, i.e. people with potential hearing or voice-related problems would suggest considering a higher sound pressure level as a reference for the speech applied to this group. However, by splitting participants into groups based on an oral survey, i.e. subjectively acquired data, the participant affiliation to one of the groups could be exposed to misdiagnoses. Indeed, people with mild hearing loss might not be aware of their impairment if it did not significantly disturb their communication or perception of their surroundings. Moreover, assessing whether the level of speaking is loud is highly subjective, as it would suggest the existence of a standard vocal effort known by the subject. A participant might not be aware of his/her abnormal or high vocal level, especially if this fact was not openly shared with him/her prior to the test. Lastly, people listening to loud music, movies or video games are often not aware that the level of their entertainment is louder than it should to preserve the integrity of their hearing.

In order to remove subjective acquired data with possible bias, the survey could be extended with an objective assessment of the hearing threshold [36] of the participants, as well as the level of their voice. The research could further be extended by requesting the test persons to level-match a given music fragment to its usual exposure. However, this would require the researcher to have physical access to the participant, which would defeat the purpose of a subjective calibration, i.e. calibrate an unknown listening system to perform an online or remote listening test. Considering the uncertainty of the calibrated level with the proposed method, we would suggest using an objective calibration method with a dummy head, artificial ear or other calibrators for listening tests carried out within research facilities.

Furthermore, level matching a pre-recorded speech might be challenging as the participants did not encounter the reciting speaker in real life. It could be even more complex in case the participant does not fluently speak the language of the heard speech. To tackle the latter issue, an alternative method would be to ask the participants to record discussion or speech at about 1 to 1.5m with their computer or phone and use this to calibrate the experiment. This approach would add time to the test procedure and would require a dynamic calibration tool built-in the test platform, but seems worth to consider.

It is worth mentioning that this experiment assumed the use of headphones or earphones and could be coupled to an additional procedure to ensure it by means of Huggins pitch [37] or anti-phased sines [38].

Lastly, the proposed subjective method does not account for the frequency response of the listening apparatus and would suggest the use of a system with a flat response over the hearing frequency range to accurately calibrate the experiment. This issue could be partially dealt with by compensating for the frequency response of the headphones using a database of measured headphones or earphones. Headphones tend to colour more significantly the signal, i.e. altering the weighing of its frequency curve, compared to a typical DAC and amplifier combo. Moreover, this process would rely on the use of headphones whose frequency response does not significantly differ from the measured one. Indeed, the frequency response of headphones tends to be altered when ageing. Lastly, flattering the frequency response can be quite challenging in the low-frequency range in the case of closed-back headphones whose design typically enhances this frequency range.

Acknowledgements

L.K. is grateful to the EPF - Graduate School of Engineering and the Faculty of Architecture of KU Leuven for the support given to his thesis research.

References

[1] WHO, “Coronavirus disease (COVID-19) pandemic.” [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019. [Accessed: 07-Jan-2021].

[2] European Centre for Disease Prevention and Control, “COVID-19 pandemic.” [Online]. Available: https://www.ecdc.europa.eu/en/covid-19-pandemic. [Accessed: 08-Jan-2021].

[3] T. Grootswagers, “A primer on running human behavioural experiments online,” *Behav. Res. Methods*, vol. 52, no. 6, pp. 2283–2286, 2020.

[4] T. H. Pedersen, S. Antunes, and B. Rasmussen, “Online listening tests on sound insulation of walls: A feasibility study,” in *Proceedings of EURONOISE 2012*, 2012, pp. 1219–1224.

[5] T. X. F. Seow and T. U. Hauser, “Reliability of web-based affective auditory stimulus presentation,” *Behav. Res. Methods*, no. 0, pp. 10–12, 2021.

[6] S. Zhao *et al.*, “Rapid ocular responses are modulated by bottom-up-driven auditory salience,” *J. Neurosci.*, vol. 39, no. 39, pp. 7703–7714, 2019.

[7] T. Pankovski, “Screening For Dichotic Acoustic Context And Headphones In Online Crowdsourced Hearing Studies,” *Can. Acoust.*, vol. 49, no. 2 SE-Article-Psychological Acoustics, Jul. 2021.

[8] “Pavlovia.” [Online]. Available: https://pavlovia.org/. [Accessed: 19-Mar-2021].

[9] J. Peirce and M. MacAskill, *Building experiments in PsychoPy*. Sage, 2018.

[10] K. Crowston, “Amazon Mechanical Turk: A Research Tool for Organizations and Information Systems Scholars,” in *Shaping the Future of ICT Research. Methods and Approaches*, 2012, pp. 210–221.

[11] T. M. Gureckis *et al.*, “psiTurk: An open-source framework for conducting replicable behavioral experiments online,” *Behav. Res. Methods*, vol. 48, no. 3, pp. 829–842, 2016.

[12] F. Keller, S. Gunasekharan, N. Mayo, and M. Corley, “Timing accuracy of web experiments: A case study using the WebExp software package,” *Behav. Res. Methods*, vol. 41, no. 1, pp. 1–12, 2009.

[13] A. Anwyl-Irvine, J. Massonnié, A. Flitton, N. Kirkham, and J. Evershed, “Gorilla in our MIDST: An online behavioral experiment builder,” *bioRxiv*, no. April 2019, pp. 388–407, 2018.

[14] J. R. de Leeuw, “jsPsych: A JavaScript library for creating behavioral experiments in a Web browser,” *Behav. Res. Methods*, vol. 47, no. 1, pp. 1–12, 2015.

[15] F. Henninger, Y. Shevchenko, U. K. Mertens, P. Kieslich, and B. Hilbig, “lab.js: A free, open, online study builder,” vol. 6, 2019.

[16] T. Y. Chen and J. Myers, “Worldlikeness: A Web crowdsourcing platform for typological psycholinguistics,” *Linguist. Vanguard*, vol. 7, no. s1, pp. 1–11, 2021.

[17] I. Adjerid and K. Kelley, “Big data in psychology: A framework for research advancement,” *Am. Psychol.*, vol. 73, no. 7, pp. 899–917, 2018.

[18] N. Stewart, J. Chandler, and G. Paolacci, “Crowdsourcing Samples in Cognitive Science,” *Trends*

Cogn. Sci., vol. 21, no. 10, pp. 736–748, 2017.

[19] S. Clifford and J. Jerit, “Is There a Cost to Convenience? An Experimental Comparison of Data Quality in Laboratory and Online Studies,” *J. Exp. Polit. Sci.*, vol. 1, no. 2, pp. 120–131, 2014.

[20] J. Rodd, “How to maintain data quality when you can’t see your participants,” *APS Obs.*, vol. 32, no. 3, 2019.

[21] H. Z. Hugo Fast, *PsychoAcoustics - Facts and Models*. 2007.

[22] B. C. J. Moore, *An introduction to the psychology of hearing*. Brill, 2012.

[23] ISO/TC 43 Acoustics, *Acoustics — Normal equal-loudness-level contours*. ISO 226:2003.

[24] B. Naderi, R. Zequeira Jiménez, M. Hirth, S. Möller, F. Metzger, and T. Hoßfeld, “Towards speech quality assessment using a crowdsourcing approach: evaluation of standardized methods,” *Qual. User Exp.*, vol. 6, no. 1, pp. 1–21, 2020.

[25] J. Pauwels, S. Dixon, and J. D. Reiss, “A Front End for Adaptive Online Listening Tests,” in *Web Audio Conference*, 2021.

[26] A. Yamamoto *et al.*, “Comparison of remote experiments using crowdsourcing and laboratory experiments on speech intelligibility,” 2021.

[27] W. Ziegler *et al.*, “Crowdsourcing as a tool in the clinical assessment of intelligibility in dysarthria: How to deal with excessive variation,” *J. Commun. Disord.*, vol. 93, no. November 2020, pp. 1–16, 2021.

[28] K. . Pearsons, R. . Bennett, and S. Fidell, “Speech levels in various environments,” *Bolt Beranek Newman*, vol. Report No., no. May, p. Canoga Park, CA, 1976.

[29] M. Kobayashi, M. Morimoto, H. Sato, and H. Sato, “Optimum speech level to minimize listening difficulty in public spaces,” *J. Acoust. Soc. Am.*, vol. 121, no. 1, pp. 251–256, 2007.

[30] L. M. Stowe and E. J. Golob, “Evidence that the Lombard effect is frequency-specific in humans,” *J. Acoust. Soc. Am.*, vol. 134, no. 1, pp. 640–647, 2013.

[31] A. Weisser and J. M. Buchholz, “Conversational speech levels and signal-to-noise ratios in realistic acoustic conditions,” *J. Acoust. Soc. Am.*, vol. 145, no. 1, pp. 349–360, 2019.

[32] O. W. O., “Average Speech Levels and Spectra in Various Speaking/Listening Conditions,” *Am. J. Audiol.*, vol. 7, no. 2, pp. 21–25, Oct. 1998.

[33] Wikipedia, “Katherine Johnson.” [Online]. Available:
https://en.wikipedia.org/wiki/Katherine_Johnson. [Accessed: 13-Dec-2020].

[34] B. N. Schenkman and M. E. Nilsson, “Human echolocation: Pitch versus loudness information,” *Perception*, vol. 40, no. 7, pp. 840–852, 2011.

[35] L. J. Norman and L. Thaler, “Stimulus uncertainty affects perception in human echolocation: Timing, level, and spectrum.,” *J. Exp. Psychol. Gen.*, vol. 149, no. 12, pp. 2314–2331, 2020.

[36] ISO/TC 43 Acoustics, *Acoustics — Audiometric test methods — Part 1: Pure-tone air and bone conduction audiometry*. ISO 8253-1:2010.

[37] A. E. Milne, R. Bianco, K. C. Poole, S. Zhao, A. J. Billig, and M. Chait, “An online headphone screening test based on dichotic pitch,” *bioRxiv*, no. 2017, 2020.

[38] K. J. P. Woods, M. H. Siegel, J. Traer, and J. H. McDermott, “Headphone screening to facilitate web-based auditory experiments,” *Attention, Perception, Psychophys.*, vol. 79, no. 7, pp. 2064–2072, 2017.