# Crowdsourcing listening tests using an audio polling station

**Olivier Robin*[1,2], Abdelghani Benghanem†[1,2], Philippe Gournay‡[3], Olivier Valentin§[1,2], Philippe-Aubert Gauthier¶[4,2], Roch Lefebvre‖[3], and Alain Berry**[1,2]**

[1]Groupe d'Acoustique de l'Université de Sherbrooke, Université de Sherbrooke, Sherbrooke, Québec, Canada;
[2]Centre for Interdisciplinary Research in Music, Media, and Technology, McGill University, Montréal, Québec, Canada;
[3]Groupe de Recherche sur la Parole et l'Audio, Université de Sherbrooke, Université, Sherbrooke, Québec, Canada;
[4]Ecole des arts visuels et médiatiques, Université du Québec à Montréal, Montréal, Québec, Canada.

## Résumé

Afin de proposer une alternative aux tests d'écoute contrôlés en laboratoire, une approche basée sur la production participative a été testée par le biais d'une station de vote audio installée dans le cadre d'une exposition sur le son. Les personnes visitant l'exposition étaient invités à donner leur avis sur la qualité sonore des véhicules et sur la qualité sonore des algorithmes de codage audio. Pour les mêmes échantillons sonores, une série de résultats a été préalablement obtenue dans des conditions de laboratoire avec un nombre limité de participants. Dans des conditions partiellement non contrôlées et avec un plus grand nombre de participants, les résultats obtenus à l'aide de la station de vote sont conformes aux résultats obtenus en laboratoire.

**Mots clefs:** Tests d'écoute, production participative, qualité sonore

## Abstract

As an alternative to laboratory controlled listening tests, a crowd sourced approach was tested using an audio polling station installed in a sound-related exhibition. Visitors were asked to provide their opinion concerning sound quality of vehicles and sound quality of audio coding algorithms. Using the same audio samples, another series of evaluation results were obtained in laboratory conditions with a limited number of participants. With partially uncontrolled conditions and a larger number of participants, the results obtained using the audio polling station are in line with laboratory results.

**Keywords:** Listening tests, crowd sourcing, sound quality

## 1 Introduction

Formal listening tests are used in a large variety of sound-related research areas, from music perception and digital sound encoding to sound quality research [1], and generally regarded as the most reliable method for audio quality evaluation. In practical terms, such tests generally require complex and controlled protocols that involve consequent manpower and demanding preparation. These constraints often result in a reduced number of participants, and the same trained or practiced listeners might be recurrently enrolled to simplify the setup of tests but also warrant results consistency [2]. Indeed, some standards even explicitly call for experienced listeners, like the International Telecommunication Union ITU-R BS.1534 [3] which specifies the use of at least 20 expert participants. Expert listeners are usually preferred to non-expert ones for such qualifying tests because assessors should be experienced in detecting small impairments in audio signals. While tests conducted using a relatively small group of experts are expected to provide a better and quicker indication of the likely results in the long term, such expert testing might also lead to excessively refined results that are not fully representative of the target customers [4]. Indeed, products or applications will be used by a much more significant number of people with a reduced sensitivity concerning audio quality [5].

Crowdsourcing has been established as a powerful tool to collect and gather human subjective data [6, 7]. The main advantages of using crowdsourcing generally include reduced costs, improved speed and flexibility, together with the acquiring of large data sets. These aspects provide leverage for reducing the complexity of listening tests, and several solutions based on this sourcing model have been proposed including web-based tests, mobile laboratory units and smartphone applications. The latter have for example been used for the production of noise maps through a participatory approach [8, 9]. Mobile laboratory units were initially proposed for on-site hearing screening in the 1960s [10] and recently extended for on-site listening experiments including virtual acoustic environments [11]. However, the great majority of the crowdsourced approaches for listening tests have been web-based [12–15]. Web-based listening tests can be theoretically performed by everyone who has access to a computer with a compatible web browser and an Internet connection. Using web-based listening tests, the data collection pro-

*. Olivier.Robin@usherbrooke.ca
†. Abdelghani.Benghanem@usherbrooke.ca
‡. Philippe.Gournay@usherbrooke.ca
§. m.olivier.valentin@gmail.com
¶. gauthier.philippe-aubert@uqam.ca
‖. Roch.Lefebvre@usherbrooke.ca
**. Alain.Berry@usherbrooke.ca

cess can be largely sped up with reduced time and location constraints. In Cartwright et al. [13], data were collected from 530 participants in only 8.2 hours. Different quality control mechanisms have been proposed to ensure reliable results and limit potential bias sources, and the use of headphones instead of loudspeakers is especially recommended [7]. Several studies [7, 13] have shown that if the listening experiment was properly designed including quality control and limitations of possible stimuli and scenarios, minor differences exist between listening tests carried out in a laboratory environment and those carried out over the Internet.

Classically, listening tests are performed using dedicated equipment in a controlled environment where listeners have to go - whereas most of the cited crowd sourced approaches rely on tests made in uncontrolled environment using variable equipment (generally owned by listeners). The proposal of mobile laboratory units is surely unique in the sense that the laboratory environment goes to the listener. An idea that has not yet been explored in the case of listening tests is making available to the general public a given equipment in a public environment.

The main contribution of this work is thus a proof of concept for the use of an audio polling station for crowd sourcing listening tests, *i.e.* to collect the general public opinion in partially controlled conditions. Such an approach implies that no recruiting and handling of subjects is required while large group of subjects can be tested, and a specific cross-section of the general population can be reached by the placement of the polling station.

In the present work, an audio polling station was installed in an exhibition concerning sound in a science museum. Visitors were asked to provide their opinion, on a voluntary basis, concerning audio samples following a paired-comparison paradigm. The polling station design is detailed in Section 2. The considered test cases are presented in Section 3. Results obtained using a controlled laboratory experiments and the polling station are compared in Section 4, followed by concluding remarks in Section 5.

## 2 The polling station design and installation

Born out of a collaboration between the Sherbrooke Museum of Nature and Science and Université de Sherbrooke, the exhibition, 'Sound, only sound !', is a touring interactive exhibition that includes fourteen zones covering four different dimensions of sound (seeing, touching, feeling and hearing sound) [16]. One of the interactive zones includes a polling station for crowdsourcing listening tests which is composed of a desktop computer, a touchscreen monitor (Elo 7200), an external sound card (Audient id4) and closed headphones (dbi pro-705), see Figure 1(a). The sound pressure levels were calibrated using a binaural manikin (GRAS 45BB KEMAR Head & Torso, equipped with large ears and GRAS 40AD 1/2" microphones), and it was verified that they did not exceed the WHO recommendations (equivalent sound pressure level $L_{eq}$ over 8 hours not exceeding 75 dB(A) [17]). The same headphones were used for all the participants, which ensured consistent audio level and quality. The background noise in the room was not measured during the museum opening hours, but its effect was estimated to be limited since exhibition areas were quiet ones. Listening using closed headphones was thus considered to provide adequate shielding against this low background noise and a sufficient signal-to-noise ratio. The user interface was developed using the AB test page of Web Audio Evaluation Tool [12], and adapted in terms of content and subject presentation, see Figure 1(b).

Every visitor of the exhibition could access the polling station. Participants were informed that these tests were part of a research project, and that adult consent was required for persons under the age of 18. The only collected information were the answers to the listening tests that followed AB tests, i.e. following a paired-comparison paradigm [18]. Therefore, no ethics approval was required by the *Comité d'éthique pour la recherche*, the internal review board at *Université de Sherbrooke* because (1) this research involved only the observation of individuals in public places, (2) did not involve planned or direct interventions by researchers with participants, and (3) the research subjects did not have a reasonable expectation of privacy and the dissemination of research results does not identify specific individuals [19].

A case-independent approach was followed, *i.e.* listening tests concerning various topics (sound quality of vehicles, music and speech audio compression quality, audio compression effect on voice, sound of backup alarms) were combined. For each test, individuals (or subjects) were asked to listen to two audio samples (A and B), and make a preference choice between A or B. The series of tests was continuously and randomly presented. Examples of icons and corresponding questions are provided in Figure 1(c). The order of presentation of the A/B pairs was also randomized. Concerning sound quality of vehicles and music and speech audio compression quality, results were also previously obtained in controlled laboratory conditions and here compared to the results obtained using the polling station during the first presentation of the exhibition at Sherbrooke (see Sections 3 and 4).

More than 4 000 opinions on the whole series of topics and tests were collected in six weeks at Sherbrooke, Québec, Canada during Summer 2019. After this first iteration, the exhibition also toured in two other locations (The Exploration Place, Prince George, British Columbia, Fall 2019 and Resurgo Place, Moncton, New Brunswick, Winter 2020), but was unfortunately stopped by the COVID-19 pandemic.

## 3 Considered test cases

### 3.1 Sound quality of vehicles

Like many attributes related to the perceived quality of a product, sound has become an important factor that influences the consumers' perception of the quality of a vehicle. Consequently, automotive manufacturers have undertaken efforts to design the acoustic signature to match the vehicle image in the customers' mind, as a way to optimize the "desire-to-buy" of their products. Unsupervised perceptive evaluations of the sound signature might be a solution to overcome the diffi-

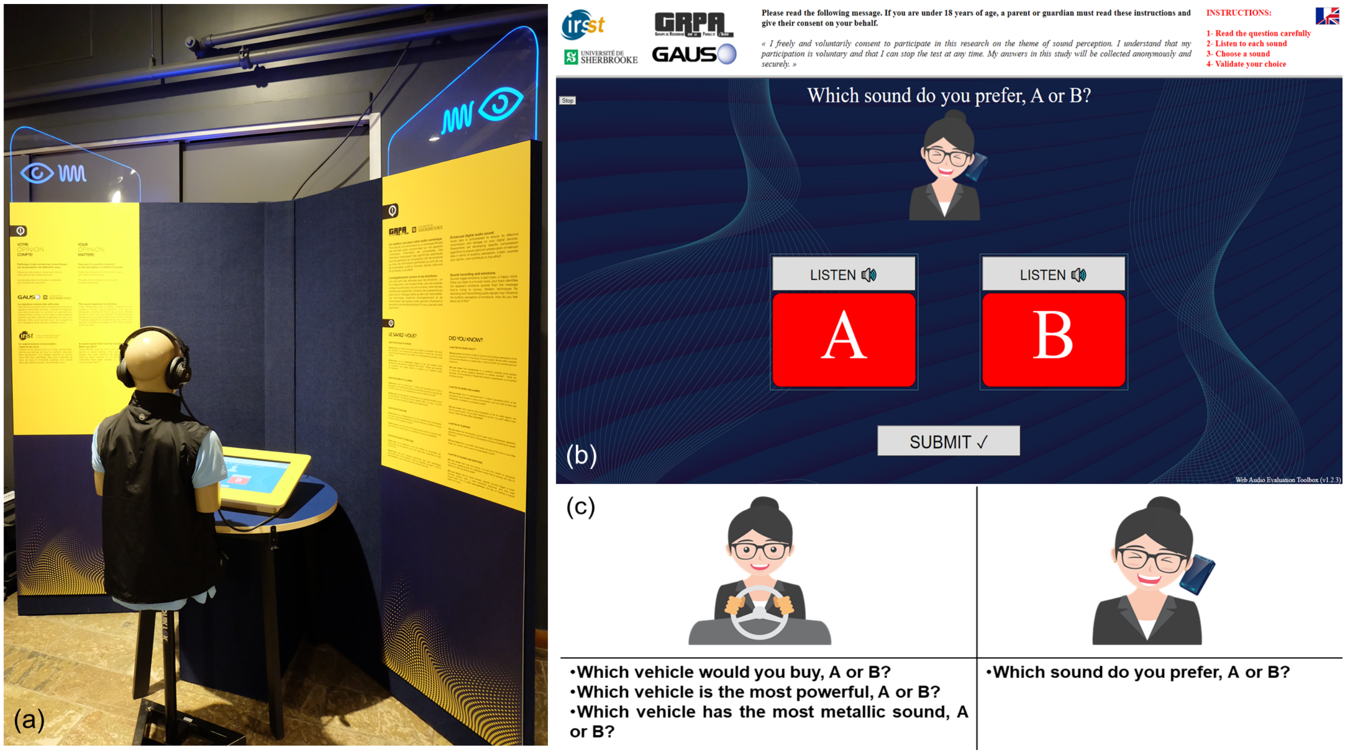


**Figure 1:** (a) The polling station as presented at Sherbrooke Nature and Sciences Museum - the acoustic manikin illustrates the typical positioning of a participant ; (b) Close-up view of the user interface ; (c) The presented icons and corresponding questions for sound quality of vehicles and sound quality for audio coding algorithms, respectively.

culties linked to classical perceptive measurements performed in laboratory environments (recruiting and handling subjects, providing tools and environment for performing tests), and could expand the possibilities for gathering consumers opinion. The interior sounds of seven side-by-side recreational vehicles (SSV) were recorded on the passenger side while the vehicles were rapidly accelerating on an asphalt road, from 0 to 60 km/h in a few seconds (Wide-open throttle condition). Recordings were performed using a binaural mannequin (GRAS 45BB KEMAR Head & Torso, equipped with large ears and GRAS 40AD 1/2" microphones), with a sampling rate of 48 kHz and a 24-bit resolution. Audio samples presented to the participants had a duration of 5 seconds, and short fade-ins and fade-outs were applied in the beginning and at the end of each sound sample so that they could be repeated without audible artifacts. Audio samples were also equalized to the same global loudness value and were presented to participants using Sennheiser headphones (chosen among models HD600, HD555, HD598, or HD579) [20, 21]. Each headphone was frequency-equalized by filtering each sound sample with appropriate frequency response amplitude-only for the left and right channels of each headset, using 2048-order zero-phase finite impulse response filters. This creates the same output signals from the headphones as those measured with the binaural mannequin. A validation of sound reconstruction was performed : the binaural microphones were installed on the KEMAR mannequin equipped with large artificial ears and the spectrum of the sounds recorded on the mannequin was compared with the spectrum of the original sounds measured in the interior of the vehicle cabin in operating condition.

Twenty-one pairs of sounds to be evaluated were included in the polling station. With seven recordings, all recordings were compared against all others. The A/B testing procedure was used to evaluate each pair of sounds regarding the "desire-to-buy" and two perceptual attributes ("powerful" and "metallic"), chosen among the outcomes of a rapid sensory analysis performed with a pool of consumers of recreational vehiclest [20, 21].

### 3.2 Sound quality for audio coding algorithms

The objective of this experiment was to compare two audio coding technologies : the xHE-AAC profile of the MPEG-D USAC standard [22], and Layer III of the MPEG-1/2 audio compression standards (MP3). The xHE-AAC codec is the high-quality codec used in the USAC verification test [23]. The MP3 codec is the LAME high quality codec version 3.99.3 operated at a constant bit rate [24].

The xHE-AAC technology is more recent (2012) and therefore in principle more efficient than the MP3 technology which was approved in 1992. Three content categories (speech, music, and speech-over-music) were considered, with two representative audio samples per category. The subjects could listen to and compare two coded versions, one with xHE-AAC and the other with MP3. The xHE-AAC encoder was forced in its linear predictive coding mode and used at a fixed bit rate (24 kbps) while MP3 was used at one of four possible bit rates (from 32 up to 96 kbps). Comparisons thus included twenty-four pairs of stereo audio samples (six different audio samples times four MP3 bit rates). Participants had then to express their preference towards one or the other. Neither the identity of the coders nor the MP3 bit rate were disclosed to the subjects. The same experiment was carried out in laboratory conditions with eight expert listeners.

# 4 Results

## 4.1 Sound quality of vehicles

Figure 2 presents the results of the listening tests obtained using the polling station, compared to the ones obtained using classical supervised listening tests performed with a panel of 17 SSV users. These supervised listening tests did not involve A/B comparisons but rather ratings of individual sounds on a 0-100 scale with respect to the various attributes. The results for the polling station correspond to the ratio between the number of times a sound was chosen and the total presentation number of that sound, while SSV users panel results are thus expressed as the median of the given ratings. To allow for direct comparison, the scores of the two panels were centered and reduced (mean was subtracted and results were divided by standard deviation).

Results from Fig. 2 indicate that unsupervised polling stations provide similar evaluations than the ones performed in controlled conditions by the users' panel. Indeed, whether it is for the "powerful" attribute, the "metallic" attribute, or the "desire-to-buy", results are consistent regardless of the evaluation method. Note that for the supervised tests, the "powerful" and "metallic" attributes were suggested by the same panel of users as a result of a preliminary sensory analysis.

Also the subjects involved in the polling-station experiment were a different cross-section of the total population, whereas the lab subjects were all SSV drivers. They may respond differently to these sounds purely due their familiarity with such vehicles in daily life, but the results obtained show that the subjects similarly react to audio samples.

Indeed, coefficients of determination ($R^2$), computed using Matlab R2020a, support that results obtained using the polling station are correlated with those obtained with the users' panel ("powerful" attribute : $R^2 = 0.64$ ; "metallic" attribute : $R^2 = 0.76$ ; "desire-to-buy" : $R^2 = 0.91$). The largest correlation is obtained for the "desire-to-buy" evaluation case, which is attributed to the fact that the "desire to buy" is likely less abstract than the two other perceptual attributes. $R^2$-related observations are confirmed by two-tailed sign test procedures which reject the null hypothesis at a 5 % significance level, corroborating that unsupervised polling stations seem to be appropriate candidates to perform sound quality assessments as they provide similar results than those obtained using users' panel ("powerful" attribute : $p = 0.0243$ ; "metallic" attribute : $p = 0.0087$ ; "desire-to-buy" : $p = 0.0009$).

## 4.2 Sound quality for audio coding algorithms

The results obtained show that the percentage of preference towards xHE-ACC at 24 kbps decreases when the MP3 bit rate increases (Figure 3). This confirms the statement made in section 3.2 that MP3 is less efficient than xHE-AAC and requires higher bit rate to achieve the same level of subjective quality. The results also show that the percentage of preference towards xHE-AAC is on average slightly larger for speech than for music. This is deemed normal given that the xHE-AAC encoder was forced to operate in its linear predictive coding mode, which is particularly efficient on speech signals [23].

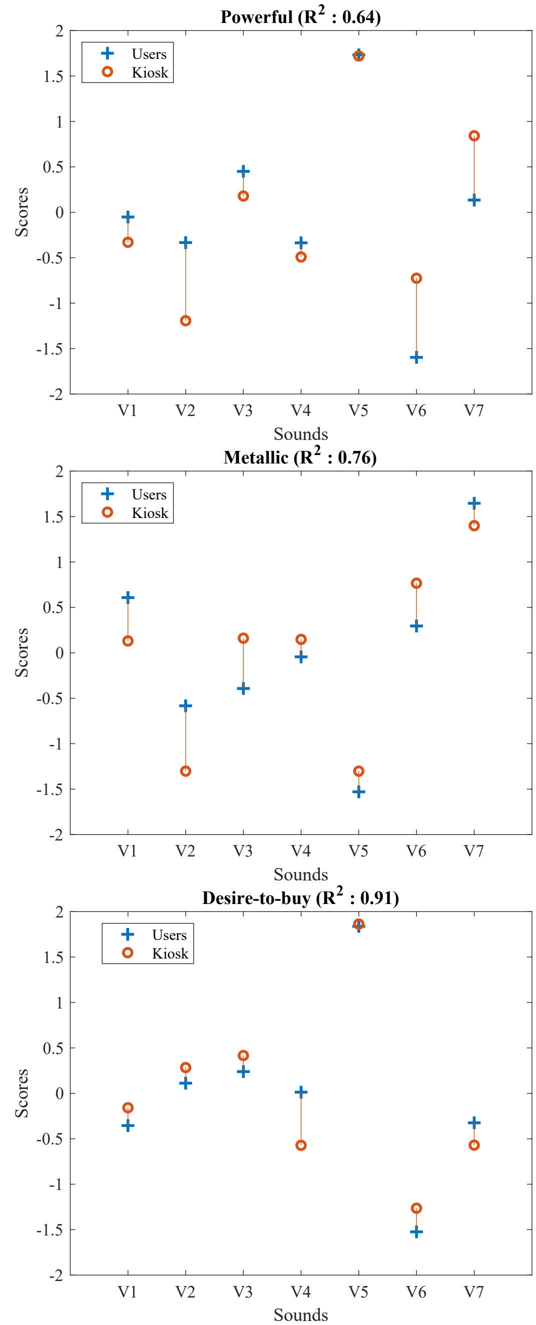


**Figure 2:** Perceptual evaluations obtained for the "powerful" attribute (top left), the "metallic" attribute (top right) and the "desire-to-buy" (bottom), using the polling station (○ markers, labelled as "kiosk") and a panel of users under controlled conditions (+ markers, labelled as "users"). Each evaluated vehicle is numbered from V1 to V7.

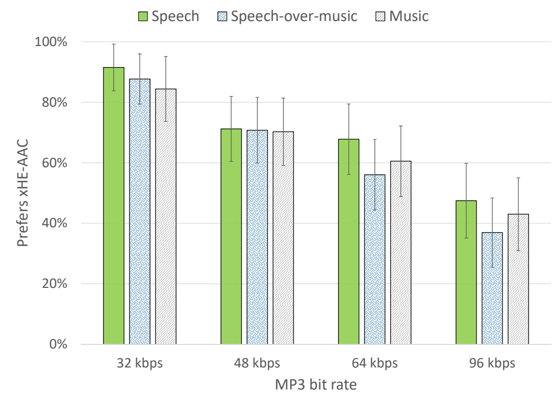


**Figure 3:** Percentage of preference in favor of the xHE-AAC codec as a function of the MP3 bitrate, per content type, with 95% confidence intervals. Results obtained with the polling station.

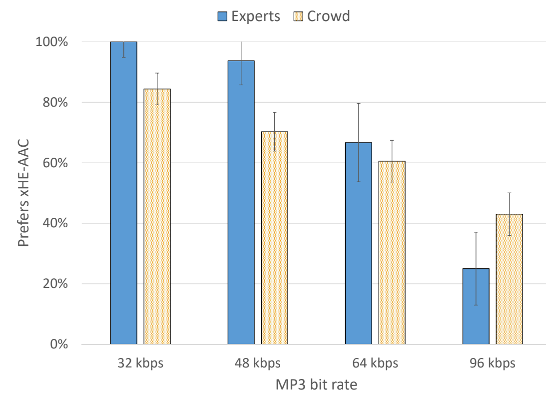


**Figure 4:** Percentage of preference in favor of the xHE-AAC codec as a function of the MP3 bitrate for all content types, with 95% confidence intervals. Comparison of results with experts or crowd.

Figure 4 compares the results obtained in the laboratory ("Expert" series) to those obtained with the polling station ("Crowd" series) for all content types. Both series exhibit the same downward trend, which is however more pronounced with expert listeners than with the crowd. There are two main reasons for expert listeners being more discriminatory. First, expert listeners have extensive experience in this kind of subjective assessment and are recruited for this given objective. Then, they operated in a controlled, thus presumably quieter and distraction-free, environment. Note that there are exactly 48 expressed preferences per data point for the "Experts" series compared to around 190 for the "Crowd" series in Figure 4. Overall, this experiment shows that the polling station makes it possible to reach the same conclusions as the expert auditors, at least when the differences in audio quality between conditions are relatively large.

## 5 Discussion

According to the results obtained, using a polling station makes it possible to reach similar conclusions as experts or trained auditors and confirms to be a possible alternative to listening tests in laboratory conditions. The trends observed in experts vs crowd evaluations are comparable, but usually less discriminatory in the case of the crowd sourced evaluation which is in line with previous works [2,4]. The number of opinions collected in six weeks together with the results obtained shows that a polling station allows quick and easy access to sound evaluations from a large panel of participants, that would have otherwise required highly time-consuming laboratory tests.

Also and compared with web-based or laboratory-based approaches, physical polling stations could be possibly installed in non usual locations (companies, stores, town halls) to attract particular group or type of subjects and to collect their opinions on various sound-related topics (sound quality of products, soundscape, environmental noise, among others). The next step of this work is to continue the analysis of gathered data, including (1) the other presented topics like perception of backup alarms and (2) to verify the geographical consistency of obtained results (*i.e.* when the exhibition toured at two other locations, did the observed trends show possible dependency on sociological and/or demographic factors ?). Indeed, the setup of the polling station after the pandemic will include simple sanitary measures. Disposable and sanitary headphone covers should be used on ear pads. The hard parts of the headphones like the headband and the touchscreen should be cleaned with adequate wipes and diluted cleaning or disinfectant solutions.

## Acknowledgments


The authors would like to thank all the subjects who kindly participated in this study. The support of Sherbrooke Nature and Science Museum (MNS2) for the setup and development of the 'Sound, only sound !' exhibition has been instrumental and here acknowledged. The support from Web Audio Evaluation Tool is also here acknowledged. The authors wish to acknowledge the financial support received from the Faculty of Engineering, Université de Sherbrooke, to setup the exhibition and from the Centre for Interdisciplinary Research in Music Media and Technology (CIRMMT), Québec, Canada, to support the analysis of data.


## References


[1] S. Bech and N. Zacharov, *Perceptual Audio Evaluation - Theory, Method and Application* (Wiley, 462 pages, 2006).

[2] W. Howie, D. Martin, S. Kim, T. Kamekawa, and R. King, "Effect of skill level on listener performance in 3D evaluation, " J. Audio Eng. Soc. 68 (9), 628-637 (2020).

[3] Recommendation ITU-R BS.1534-3 (10/2015), Method for the subjective assessment of intermediate quality level of audio systems

[4] N. Schinkel-Bielefeld, N. Lotze, F. Nagel, "Audio quality evaluation by experienced and inexperienced listeners, " Proc. Mtgs. Acoust. 19, 060016 (2013)

[5] M. Schoeffler and J. Herre, "The relationship between basic audio quality and overall listening experience, " J. Acoust. Soc. Am. 140, 2101 (2016).

[6] D.C. Brabham, "Crowdsourcing as a Model for Problem Solving : An Introduction and Cases," Convergence, 14(1), 75–90 (2008).

[7] T. Hoßfeld, C. Keimel, M. Hirth, B. Gardlo, J.Habigt K. Diepold and P. Tran-Gia, "Best Practices for QoE Crowdtesting : QoE Assessment With Crowdsourcing, " IEEE Trans. Multimedia, vol. 16, no. 2, pp. 541-558 (2014).

[8] J. Picaut, N. Fortin, E. Bocher, G. Petit, P. Aumond, G. Guillaume, “An open-science crowdsourcing approach for producing community noise maps using smartphones, ” Building and Environment, **148**, 20-33 (2019).

[9] A. Radicchi, D. Henckel and M. Memmel, “Citizens as smart, active sensors for a quiet and just city. The case of the ‘open source soundscapes’ approach to identify, assess and plan ‘everyday quiet areas’ in cities, ” Noise Mapping 5, 1–20 (2018).

[10] W. Taylor, W. Burns and A. Mair, “A mobile unit for the assessment of hearing, ” Ann. Occup. Hyg. 7, 343–352 (1964).

[11] F. Pausch and J. Fels, “MobiLab – A Mobile Laboratory for On-Site Listening Experiments in Virtual Acoustic Environments,” Acta Acust. united Ac., 105 (5), 875–887 (2019).

[12] N. Jillings, B. De Man, D. Moffat, and J. D. Reiss, “Web audio evaluation tool : A framework for subjective assessment of audio, ” Proc. of 2nd Web Audio Conference, April 4-6, Atlanta, USA (2016). http://hdl.handle.net/1853/54595, accessed 10 November 2020

[13] M. Cartwright, B. Pardo, G. J. Mysore and M. Hoffman, “Fast and easy crowdsourced perceptual audio evaluation, ” 2016 IEEE ICASSP, Shanghai, 2016, pp. 619-623.

[14] B. Pedersen, “Hi-precision audio in listening tests - also in the browser ?, ” Proc. of 3rd Web Audio Conference, August 21-23, London, UK (2017). https://qmro.qmul.ac.uk/xmlui/handle/123456789/26144, accessed 10 November 2020.

[15] M. Schoeffler, S. Bartoschek, F.-R. Stöter, M. Roess, S. Westphal, B. Edler and J. Herre, “webMUSHRA — A Comprehensive Framework for Web-based Listening Tests, ” Journal of Open Research Software, 6(1), 8 (2018).

[16] “SOUND, ONLY SOUND !” (2020) [Exhibition]. Musée Nature Sciences Sherbrooke (MNS2). https://mns2.ca/en/travelling/travel-with-the-sound/, accessed 22 May 2021.

[17] WHO (1999). Guidelines for community noise. Geneva : World Health Organization (http://apps.who.int/iris/handle/10665/66217?show=full, accessed 14 June 2020).

[18] J.L. Punch, B. Rakerd, and A.M. Amlani, “Paired-comparison hearing aid preferences : Evaluation of an unforced-choice paradigm, ” J. Am. Acad. Audiol., 12 : 190–201 (2001).

[19] Tri-Council Policy Statement : Ethical Conduct for Research Involving Humans – TCPS 2 (2018), Article 2.3 (https://ethics.gc.ca/eng/policy-politique_tcps2-eptc2_2018.html, Accessed 23 August 2021).

[20] A. Benghanem, O. Valentin, P.-A. Gauthier and A. Berry, “Sound quality evaluation of vehicles : effect of global loudness equalization on sensory profile,” Proc. of ICSV26 conference, Montréal (2019), https://www.iiav.org/content/archives_icsv_last/2019_icsv26/index.html, accessed 10 November 2020.

[21] A. Benghanem, O. Valentin, P.-A. Gauthier and A. Berry, “Sound quality of side-by-side vehicles : Investigation of multidimensional sensory profiles and loudness equalization in an industrial context,” Acta Acustica, 5, 1–19 (2021).

[22] M. Neuendorf et al., “MPEG Unified Speech and Audio Coding - The ISO/MPEG Standard for High-Efficiency Audio Coding of all Content Types,” J. Audio Eng. Soc., 61 (12), 956–977 (2013).

[23] S. Quackenbush and R. Lefebvre, “Performance of MPEG Unified Speech and Audio Coding,” 131th AES Convention, New York, USA, October 20-23 (2011).

[24] LAME high quality MPEG Audio Layer III (MP3) encoder, version 3.99.3, The LAME Project, http://lame.sourceforge.io/, accessed 10 November 2020.


CONTROL
NOISE
LOWER PROJECT COSTS
IMPROVE SPEECH PRIVACY
BOOST COMFORT
& WELLNESS
QUICK ROI
INCREASE PRODUCTIVITY
FACILITY
FLEXIBILITY
ENHANCE WORKPLACE CULTURE
SUPPORT FOCUS
LogiSon ACOUSTIC NETWORK
SOUND.
THAT WORKS.™
Sound masking is more than a product. It's a service provided by professional technicians who know the effect isn't achieved from the moment they power the system, but by tuning the sound to an independently-proven curve. Designed right, tuned right—that's our motto. And the result is more consistent, comfortable and effective sound masking.
www.logison.com
© 2021 KR MOELLER ASSOCIATES LTD. LOGISON IS A REGISTERED TRADEMARK OF 777388 ONTARIO LIMITED. PHOTO BY VINCENT LIONS.