TECHNICAL & APPLIED ARTICLE

OPEN ACCESS

MASAVE: A mobile test system for audio-visual experiments at home

Ewald Strasser[1,*] , Thomas Brand[1], and Jan Rennies[2]

[1] Department für Medizinische Physik und Akustik, Universität Oldenburg and Cluster of Excellence Hearing4all, Oldenburg 26129, Germany

[2] Fraunhofer-Institut für Digitale Medientechnologie IDMT, Hör-, Sprach- und Audiotechnologie and Cluster of Excellence Hearing4all, Oldenburg 26129, Germany

Abstract – We present a mobile apparatus for audio-visual experiments (MASAVE) that is easy to build with a low budget and which can run listening tests, pupillometry, and eye-tracking, e.g., for measuring listening effort and fatigue. The design goal was to keep the MASAVE at affordable costs and to enable shipping the preassembled system to the subjects for self-setup in home environments. Two experiments were conducted to validate the proposed system. In the first experiment we tested the reliability of speech perception data gathered using the MASAVE in a less controlled, rather noisy environment. Speech recognition thresholds (SRTs) were measured in a lobby versus a sound-attenuated boot. Results show that the data from both sites did not differ significantly and SRT measurements were possible even for speech levels as low as 40–45 dB SPL. The second experiment validated the usability of the preassembled system and the use of pupillometry measurements under conditions of darkness, which can be achieved by applying a textile cover over the MASAVE and the subject to block out light. The results suggest that the tested participants had no usability issues with setting up the system, that the temperature under the cover increased by several degrees only when the measurement duration was rather long, and that pupillometry measurements can be made with the proposed setup. Overall, the validations indicate that the MASAVE can serve as an alternative when lab testing is not possible, and to gather more data or to reach subject groups that are otherwise difficult to reach.

Introduction

In the recent years, we saw that a lot of attempts to replicate key studies in psychology failed. This phenomenon is known as the replication crisis, which is, besides other causes (see [1]), to a large part caused by underpowered studies (see [2]). The obvious reason to perform an underpowered study is of course the high effort that comes with testing of big sample sizes in a lab setting.

Researchers have therefore employed alternatives to inviting test subjects to the laboratory and conducting experiments under highly controlled conditions. One alternative are web-based platforms. Web-based testing approaches can lead to more diverse samples and therefore are more representative of the general population [3], and obviously enable larger samples sizes. They also can deliver results quite fast [4], and they are easy to replicate [5]. However, the huge advantages of web-based testing also come with number of drawbacks. One issue is the uncontrollability of web-based studies. Neither the environment nor the hardware to perform the study are under the

control of the researcher. Another problem is that special hardware, like EEG or reliable pupillometric measurement devices, are not readily available when subjects use their own equipment.

A possible solution for this can be to bring the lab to the user. For example, the study reported in [6] presented a mobile apparatus for audio-visual experiments built into a modified caravan. The caravan is sound-insulated and hosts all necessary technical facilities to perform experiments in virtual acoustic environment. The caravan can be driven to schools and retirement homes where there are many individuals available on site, which might otherwise refrain from participating in listening studies, for example due to low technical affinity or organizational effort related to visiting the lab. One downside of this approach is that it is very expensive and time-consuming to build such a caravan.

The approach we want to present here is in-between the very limited hardware-related possibilities and low degree of control of web-based testing and the building of a full-fledged mobile testing facility. Our aim was to build a device that is lightweight, not very expensive and can be built quickly and, if desired, in larger quantities. The need

*Corresponding author: ewald.strasser@uni-oldenburg.de

for a lightweight design was motivated by the possibility to send the preassembled device by mail to our subjects. Cost efficiency and quick assembly time were desired to allow researchers to react to a sudden need of alternative experimental platforms, e.g., at times when lab-testing is impossible and budget is limited. The same requirements hold for parallel testing using many devices in order to increase statistical power.

We designed our device to be flexible for a wide range of listening experiments. The integration of an eye-tracker equips the device with the ability to measure one of the more widely used physiological measurements of listening effort. In our case we intended to use the device to measure the Pupil Unrest Index (PUI) [7] which is a measure of fatigue and has to be performed in darkness.

In this paper we provide all necessary details and instructions to rebuild the device. We also present the results of two validation tests, which focused on the following key aspects that have to be taken into consideration for the intended use: (1) whether listening test results gathered with MASAVE are comparable with results from a listening booth, (2) whether suitable lighting control can be achieved by applying a textile cover for reliable pupillometric testing, (3) whether MASAVE can be set up and used by naïve subjects, and (4) whether the environmental conditions that prevail under the cover of MASAVE are suitable for human subjects.

The first experiment aimed to explore whether the results that are gathered with MASAVE are comparable with results gathered in a listening booth. We simulated different spatial conditions via headphones and compared the effects of different speech processing strategies on speech perception on speech recognition thresholds (SRTs), i.e., signal-to-noise ratio (SNRs) required for 50% speech intelligibility. This experiment was selected because it represents a variety of aspects of typical experiments in hearing research (in which different listening conditions and technical systems are compared) to illustrate that the device can be used for a wide range of listening experiments. SRTs were then compared between optimal laboratory conditions and conditions in which the MASAVE was placed in a less optimal measurement environment. In the second experiment, the feasibility of conducting pupillometric measurements with the proposed system was validated. In addition, changes of environmental conditions (temperature, CO_2 concentration, humidity) resulting from extended use of the textile cover were measured, and aspects related to usability and effort associated with the unsupervised self-setup were investigated by means of questionnaires.

Requirements, system components and design

The main purpose of the MASAVE is to conduct listening experiments. We designed it such that all components are included in the system, and no additional components except power supplies have to be provided externally. A schematic of the MASAVE is shown in Figure 1 and a list of all components is provided in Table 1.
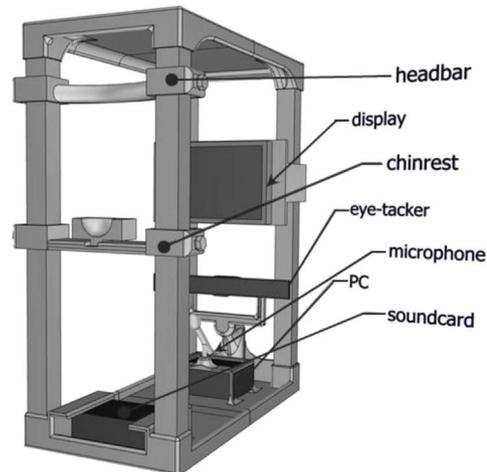

Figure 1. Schematic drawing of the MASAVE.

Table 1. Components that are used for the Mobile apparatus for audio-visual experiments (MASAVE).

	Quantity
PC setup	
Fanless PC IS1000: Intel core i7 8GB ram SSD	1 piece
Keyboard	1 piece
Mouse	1 piece
Display: SunFounder 10.1″ IPS Monitor	1 piece
Mobile hotspot	1 piece
Webcam	1 piece
Eye-tracker	
Eye-tracker: Gazept GP3 HD	1 piece
Setup sound	
Soundcard: Focusrite Scarlett 2i2 3rd Gen	1 piece
Headphones: Vic Firth SIH2	1 piece
Microphone: Rode M5 MP	1 piece
Supporting structure	
Tubular PVC profiles (square 40×40 mm)	~3 m
3D printing PLA Filament	~5 kg
Curtain: Molton (300 g/m^2)	~6 m^2
Bolts and screws M6 (chinrest, display holder)	9 pieces each
Head bolt m10 (microphone holder)	1 piece

Apart from fulfilling the technical requirements for the intended usage, the components were selected based on their price and weight to maximize flexibility in mobile use. Multiple user interface options for collecting responses from the subjects were included to allow for flexible experimental design options. For some of the components, the supporting structure was explicitly designed to ensure a tight fit. These components are listed including their manufacturer and model type. Other components (like a webcam or mouse) were not fitted specifically to the structure of the MASAVE. For the supporting structure, which holds all the other components, we used PVC profiles with an inner diameter of 40 mm by 40 mm and 3D-printed parts [8] made from polylactic acid (PLA). The PVC profiles were

used as support between the bottom and the roof of the device (see Fig. 1).

The entire structure has outer dimension of 30 cm × 58 cm × 70 cm ($W \times L \times H$) and was designed such that all technical components and cables can be placed without extending beyond the outer edges. 3D printing made it possible to rapidly design prototypes and evaluate them on face value. It also allowed the design of a very lightweight structure that was exactly tailored to the needs of our device. In our case we wanted to have a chinrest and a holder for the eye-tracker that keeps the tracker in place even while carrying.

The design of the chinrest was inspired by [9]. It is intended to ensure a stable position of the subject's head, which is particularly relevant for pupillometry measurements. The chinrest as well as the display can be adjusted in their vertical position to account for individual preferences of the participants. This is done by loosening the screws that clamp the parts to the supporting structure, adjusting them as needed and tighten them up again. The 3D design of the MASAVE including the 3D printing templates was published at [8].

The PC included in the MASAVE, which runs the experimental software, is a Fanless IS1000 from TX-Team [10]. A fanless PC was needed because we wanted to avoid audible sound emissions as well as structure-borne sound to be transferred over the structure of our device to the chin rest and finally to the head of the subject. The device uses an Intel Core i7 (3rd generation) and 8 GB ram which qualifies it for a wide range of listening experiments with simultaneous eye-tracking. The PC is also very lightweight (approx. 1200 g) which helps to decrease the overall weight and increase the mobility of the system.

For pupillometry we used a GP3 eye-tracker [11]. It is a low-cost remote eye-tracker and can be interfaced via Transmission Control Protocol/Internet Protocol (TCP/IP). In our case we used Matlab [12] to record the pupillometric data.

As a soundcard we used a Focusrite Scarlett 2i2 [13]. This soundcard is also low cost but can be used with ASIO drivers. It has two analog output and input channels. It has a low weight (approximately 470 g) and a compact design which makes it very suitable for the intended application.

As a display we used a 10.1 inch SunFounder IPS LCD monitor which is interfaced via HDMI and has a maximum resolution of 1280 by 800 pixels. This display was chosen because it can be easily disassembled and put into a customized frame.

As headphones we used the Vic Firth SIH2. These circumaural headphones are low budget, have a relatively high passive damping, and are therefore well suited for non-optimal sound environments.

A computer mouse and keyboard were included as user interface devices. We used a standard wired mouse in the present setup because it is less prone to errors (e.g., empty batteries, lost connection to receiver) and can be stored in the space between PC and soundcard after each experiment. A small standard keyboard (dimensions 26.1 cm × 8.4 cm × 1.6 cm) can also be stored between soundcard and PC. The present design of the MASAVE would make regular typing difficult, but it would be possible to use some keys, for example for reaction time experiments or for occasional typing.

The MASAVE further comprises a Rode M5 microphone, a compact half-inch cardioid condenser microphone, which can be used with the Focusrite Scarlett 2i2 soundcard we chose. The microphone is used for remote communication with subjects and is also used to collect verbal responses of the subjects during testing (which can be evaluated, e.g., via automatic speech recognition, ASR). Furthermore, the microphone could also be used to monitor the noise background of the room by logging changes in sound pressure level that exceed a predefined level. Such noise monitoring could potentially be used to identify and exclude trials or blocks which are corrupted by unexpected noise during the measurement at the listener's environment (e.g., doorbells, dog barking etc.). In the validation experiments of this study, this feature was not yet included.

The MASAVE was also designed with pupillometry in mind. One application is to measure the Pupil Unrest Index (PUI) proposed by [7]. For the PUI, it is necessary to measure the pupil size of the subjects for 11 min in darkness. Since it is difficult to make sure that unsupervised subjects darken their test environment sufficiently, we designed a textile cover that fits the device and also the torso of the subject (see Fig. 2). The cover was made of molton ($300 \ \text{g/m}^2$) that can be used as a stage curtain. The design of the cover is also published at [8]. The effectivity of the cover to block out light was tested with a luxmeter (Sauter SO 200K) and is reported in the results section. We also included a mobile hotspot to guarantee that an internet connection is available, which is necessary for remote control of the experiment.

Overall, the MASAVE had a weight of about 10 kg and cost less than 2800 EUR (excluding VAT) at the time of conducting this validation study.

Experiment I: Speech perception measurements

The goal of the first validation experiment was to test if the proposed MASAVE could be used to measure speech recognition and subjective listening effort with a similar reliability as measurements in controlled laboratory rooms. Conditions were selected to comprise several listening scenarios relevant for speech perception research including effects of energetic and informational masking (see review in [14]), spatial listening as well as possible benefits of speech enhancement algorithms. In particular, the experiment was designed to include a wide range of speech levels, also comprising rather low speech levels which are most likely to be affected by noise in the test environment. It is beyond the scope of this technical paper to discuss in detail why the selected conditions differed with respect to

Figure 2. MASAVE with simulated subject without textile cover (top left), with textile covering the MASAVE structure (top right), and with textile covering the MASAVE and the entire torso to block out light (bottom).

intelligibility and effort. The important question was if the data collected in the two environments were comparable. This was analyzed by means of an equivalence test [15][1].

Methods

Subjects

Twenty subjects (two male, eighteen female) aged between 19 and 29 years participated in this study. All subjects were German native speakers and had self-proclaimed normal hearing. Subjects were paid 10€ per hour for their participation and gave written informed consent before testing.

Stimuli

The stimuli consisted of a target talker and two interfering talkers. All talkers were male. The sentences of the target talker were taken from the German matrix sentences test (Oldenburger Satztest, OlSa) [16]. These sentences have the fixed structure of name word, verb, numeral, adjective, object. For each word, ten alternatives are available which can be randomly combined to produce syntactically correct but semantically unpredictable sentences. The interfering talkers also uttered sentences from the OlSa, but

were recorded with different talkers [17]. The target talker had an average speaking rate of 3.8 syllables per second and the interfering talkers had an average speaking rate of 5.8 syllables per second. The similarity in sentence structure could be expected to produce very challenging listening conditions when subjects tried to focus on the target talker, at least without any strong unmasking cues such as, e.g., spatial separation [18].

The interfering speech of each of the two maskers was generated by concatenating three randomly assigned sentences, and then selecting a random starting point within these sentences. The earliest possible start was at the beginning of the 3-sentence string and the latest possible start was the point at which the rest of the three-sentence string was equal to the length of the current target sentence.

Two different spatial conditions were used. To render the desired spatial allocation of the different talkers we used head-related room impulse responses (HRIRs) of [19]. The target was always presented from the front (0°). The interfering talkers were either presented at azimuthal angles of ±90°, or also from the front (co-located, 0°).

The presentation level of the combined target-interferer signal was fixed at 65 dB SPL. The SNR for all signals was defined as the ratio between the target level and the level of the sum of both interferers after the convolution with the HRIRs, but before any signal enhancement.

Three processing conditions were used in the validation study, i.e., one unprocessed condition, in which stimuli were created as described above and presented to the subjects. In addition, two speech enhancement algorithms were used which aim to improve speech perception. One algorithm was the binaural minimum variance distortionless response beamformer with partial noise estimation (BMVDR-N) [20]. This algorithm employs a binaural beamformer and mixes a portion of the original signal back in the processed signal to preserve some binaural cues. It was shown to reduce SRTs in conditions with maskers spatially separated from the target talker [21]. For co-located maskers, no benefit was expected because there was no spatial benefit. The other algorithm was the so-called ideal binary mask (IBM), which analyzes target and masker signals (using oracle knowledge) in time-frequency bins (128 frequency channels spanning from 80 Hz to 8 kHz and 20 ms time windows with an overlap of 10 ms) and discards all bins in which the target energy is lower than the energy of the combined maskers.

This type of processing was shown to considerably reduce the impact of informational masking because explicit confusions of target and masker words are no longer possible [18, 22].

By combining the two spatial conditions with the three processing conditions, we obtained six types of signals:

- Unprocessed co-located
- Unprocessed separated
- BMVDR co-located
- BMVDR separated
- IBM co-located
- IBM separated

---

[1] In a first approach we analyzed the data with a linear mixed model approach. However, the results indicated a null finding which is more appropriately analyzed by an equivalence test.

Based on data in the literature, we could expect very low SRTs in some of the experimental conditions. This was intended because low SRTs would correspond to low speech levels (for a fixed overall level of target and maskers) and, hence, produce conditions which might be vulnerable to sub-optimal acoustic conditions in the test environment.

The setup was calibrated to 100 dB sound pressure level (SPL) using a Brüel and Kjær (B&K) 4153 artificial ear, a B&K 4947 1/2 inch microphone, a B&K ZC-0032 preamplifier, and a B&K 2250 sound level meter.

Procedures

Subjects performed an SRT test and a listening effort test for each of the six combinations of spatial constellation and signal processing. The intelligibility test and the listening effort test were presented in blocks. Whether the intelligibility test or the listening effort test was done first was predetermined with a permutation plan. The order of the six different conditions was randomized. Each of the two blocks was preceded by two training conditions that made sure that subjects understood the test and to familiarize them with the OlSa procedure and speech material to avoid a strong impact of training effect [16].

The SRT measurement was conducted using an open-source Matlab framework [23]. For each condition, one test list of 20 OlSa sentences was used. Subjects used a GUI displaying the entire matrix of 50 words (5 words × 10 alternatives per word) as pushbuttons to indicate the words they had recognized. Depending on the number of correctly recognized words, the SNR in the next trial was adjusted according to an adaptive procedure [24] to converge to the SRT.

Subjective listening effort was measured with the Adaptive Categorical Listening Effort Scaling (ACALES) [25]. The ACALES method utilizes a rating scale with 13 categories, which ranges from "extreme effort" (13) to "no effort" (1), with an extra category "only noise" for trials in which subjects cannot detect the target talker at all. The task of the subjects was to rate how effortful it was to listen to the previously presented sentence of the target talker.

During this study, no pupillometric data were collected. The validation that the PUI can be measured with MASAVE was part of the second validation experiment described in the next section.

Test facilities

The MASAVE was evaluated by comparing the experimental data (SRT and listening effort ratings) in a standard listening booth and in the lobby of the listening booth (see Fig. 3). Both of the described procedures were repeated at two different days, once at the listening booth (left picture) and once in the lobby (right picture). Whether the booth or the lobby came first was also predetermined with the permutation plan. The second test was done at least one day and at most a week after the first test.
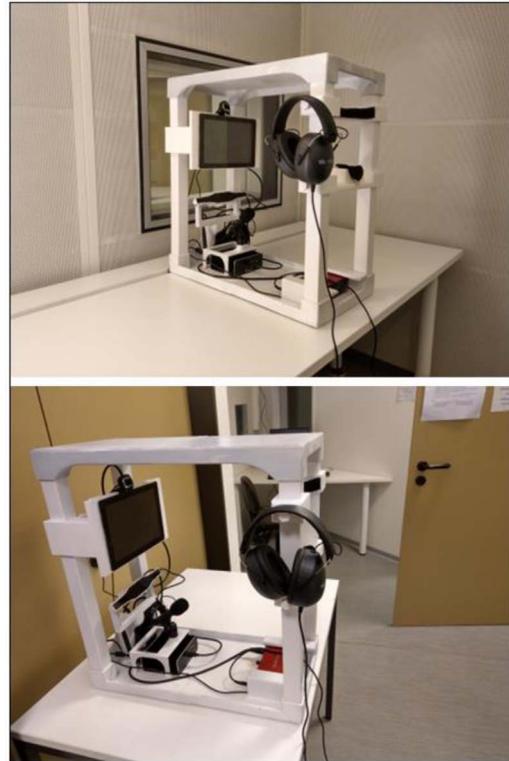

Figure 3. The same experiments were done in the listening booth (top) and in the lobby of the listening booth (bottom).

The lobby of the listening booth had a quiescent level of 38 dBA and the listening booth had a quiescent level of 27 dBA. The A-weighted third-octave spectra of both locations are shown in Figure 4. The increased noise level between about 50 and 1000 Hz in the lobby was mainly due to fan noise, which was clearly audible when not wearing the headphones and still perceptible when wearing the headphones.

It could therefore be expected that the background noise had an impact on measured SRTs, especially in the IBM conditions where speech levels were expected to be low, and a considerable amount of masker energy was removed by the signal processing. This impact should be observable as a bias towards higher SRTs (and possibly higher subjective listening effort) in the lobby compared to the listening booth.

Results

Measured SRTs of the six conditions are shown as box-plots in Figure 5 for the listening booth (light gray) and the lobby (dark gray). A first visual inspection indicated that the experimental conditions fulfilled the expectation of producing considerable SRT differences. Specifically, there seemed to be an advantage due to the spatial separation compared to co-located stimuli in all processing conditions, and the speech enhancement probably also had a beneficial effect on SRTs. Lowest SRTs were measured for IBM processing. There seemed to be also a trend towards a better
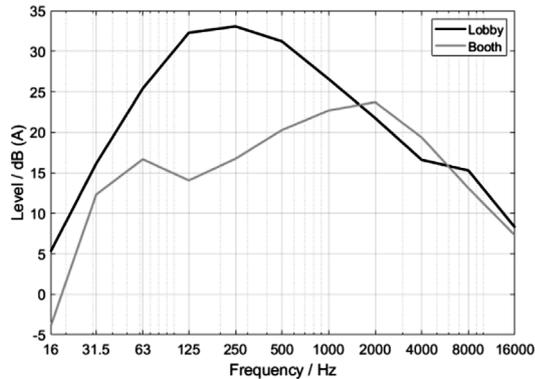
Figure 4. One-third octave band levels of the background noises in lobby (black) and in the listening booth (gray).

Figure 6. Boxplots of the SNR required to produce listening effort ratings of "moderate effort", which is right in the middle of the listening effort scale.
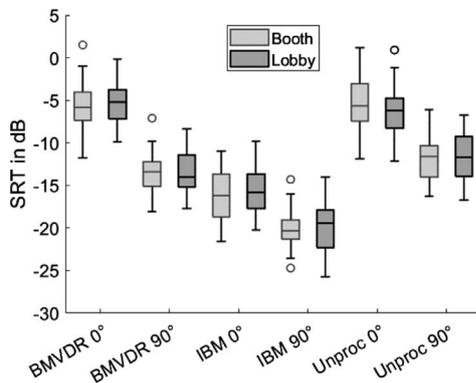

Figure 5. Boxplots of measured SRTs for each condition.

median SRT compared to the unprocessed condition for BMVDR processing in the spatial conditions. There was no obvious difference in median SRTs between booth and lobby.

Figure 6 shows results of the listening effort scaling. SNRs corresponding to "moderate effort", i.e., the middle category (7) of the 13-point scale were interpolated from each individual psychometric function and are illustrated here as boxplots in the same way as the SRT data. The listening effort ratings showed a larger interindividual variability, and revealed the same trend, i.e., highest median effort was observed for co-located maskers and unprocessed stimuli and stimuli processed using BMVDR, and lowest median effort was observed for IBM-processed stimuli. Overall, the differences between conditions were less clear due to the larger interindividual differences, but there seemed to be no obvious difference between the booth and the lobby.

The main question of this validation study was if systematic differences existed between the two test locations. We therefore statistically tested the experimental data using R [26] and the package TOSTER [27]. To test whether the effect of the testing site was negligible or substantial on SRT or subjective listening effort we performed a so-called equivalence test [15]. Conceptually this test can determine whether an observed effect size is truly larger
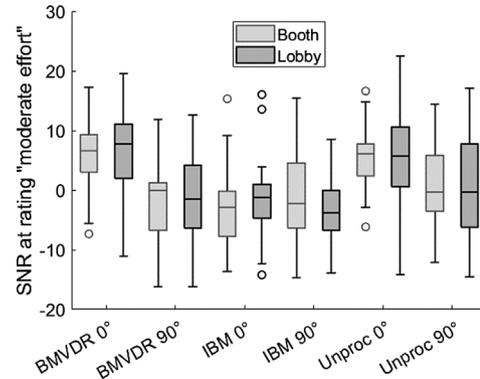
than an upper predefined effect size. The limits can be set symmetrically so that it is equally likely that sample 1 has a bigger mean value than sample 2 or vice versa, or it can be defined asymmetrically. By defining those two expected effect sizes, so called equivalence bounds are established. The next step is to calculate a confidence interval at a predefined certainty level (α) for the observed data. Combining the equivalence bounds with the confidence interval can result in four relevant outcomes: First, the confidence interval can overlap with the zero line and be within the equivalence bounds, which means that the effect is statistically indistinguishable from zero and that the effects on both sides are smaller than the predefined minimum relevant level. In this case the effect is statistically undistinguishable from zero within the predefined certainty level, and the samples can be regarded as equivalent in the sense that the effects are not bigger than the predefined equivalence bounds. Second, the confidence interval can cross an equivalence bound but not zero which means that the samples are not equivalent and statistically different. Third, the confidence interval can neither cross zero nor an equivalence bound which mean that samples can be regarded as statistically different but that they are equivalent since the effect size is smaller than the predefined level. Fourth, the confidence interval can cross an equivalence bound and also a zero which means that statistically the samples are not different but that the observed effect size could potentially be bigger than the predefined level. For more information on this see [15].

For our study we assumed at least an effect size of .5 (Cohens *d*) which is by convention often regarded as a medium sized effect. We used an equivalence test for dependent samples since each participant performed all experimental conditions.

For the analysis of the SRT data we used the mean SRT gathered from the booth (-12.0 dB) and the lobby (-12.1 dB) as well as the standard deviations of both conditions (6.1 dB booth, 5.9 dB lobby) and further the correlation ($r = .93$) between the results of each participant in both conditions in each sub-condition. Analyzing the data resulted in a significant equivalence test result for

dependent samples with a $t(113) = -4.9$ and a $p < .001$, with the bounds of equivalence of -1.1 and 1.1 dB and a defined significance level of $.05$. The confidence interval for the difference between both testing sites was between $-.3$ and $.4$ dB. This means that both testing sites are statistically indistinguishable from each other.

We also analyzed the subjective listening effort with an equivalence test. For the analysis of the subjective listening effort data we used the mean SNR for each step at ACALES scale gathered from the booth (1.4 dB) and the lobby (1.4 dB) as well as the standard deviations of SNR of both conditions (11.7 dB booth, 12.0 dB lobby) and further the correlation ($r = .92$) between each participants results in both conditions at each categorization. Again, the equivalence test became significant with a $t(1481) = -18.747$ and a $p < .001$ with the bounds of equivalence of -2.4 and 2.4 dB and a defined significance level of $.05$. The confidence interval for the difference between both testing sites is between $-.1$ and $.3$ dB and can therefore be assumed as statistically equivalent.

Individual SRT data for the two test sites are directly compared in Figure 7. Each data point corresponds to one condition and one subject. The gray dashed line represents the intersection, i.e., the perfect agreement between booth and lobby. The solid gray lines indicate a corridor of ±2 dB around this intersection, which corresponds to the 95% confidence interval of the SRT measurement [16]. It could be observed that most data points (68%) fell within this corridor and that there was, overall, a very high linear correlation of individual SRTs between the two test sites ($R = 0.87$), although individual outliers were observed. This was also reflected in the bias measure (-0.1 dB), which was calculated as the y-intercept of a linear regression function with unity slope.

The corresponding data of the listening effort scaling are shown in Figure 8.

For this representation, the individual SNRs required to obtain listening effort ratings of each rating category (1–13) were derived, i.e., there are 13 data points for each condition and subject. This scatterplot is noisier than the SRT data (44% of the data points were within ±2 dB of the intersection), but again a very high linear correlation ($R^2 = 0.85$) and a negligible bias (0.1 dB) were observed, indicating no systematic offset between the two test sites.

Testing the cover for light dimming

The version of MASAVE that is presented here was designed with the requirements of the PUI procedure [7] in mind. Therefore, we needed to ensure that the cover can effectively block out light.

Procedure

To test whether the cover can block out light, we measured the light intensity, in lux, at the position of the eyes, with a lux meter (Sauter SO 200K). At first, we established
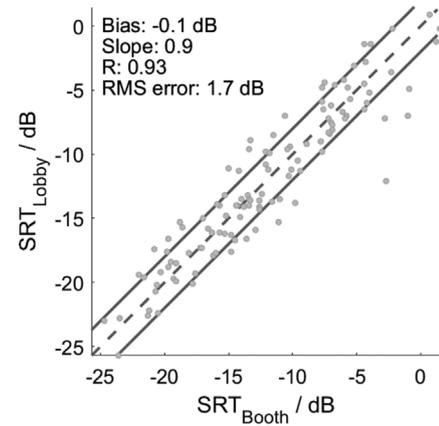

Figure 7. Comparison of SRTs measured in the lobby and the booth for each condition (unprocessed, IBM, BMVDR-N) and each subject. Solid lines represent a corridor of ±2 dB around the intersection.
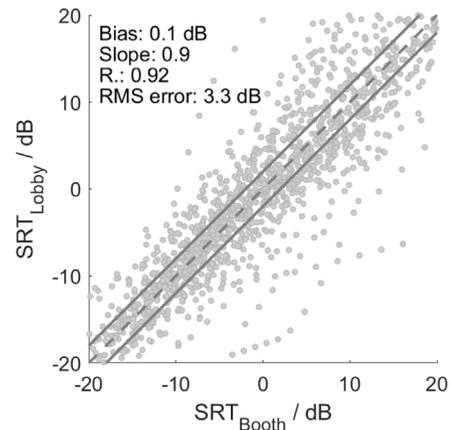

Figure 8. Comparison of the estimated SNR for each step of the ACALES scale (1–13, effortless – extreme effort) between the measurements in the lobby and the booth for each condition (unprocessed, IBM, BMVDR-N) and each subject. Solid lines represent a corridor of ±2 dB around the intersection

a baseline measurement without the cover. We positioned the MASAVE at a table near a big window. The ceiling lights were also turned on and the measurement was done at a cloudy day in the afternoon. The situation was considered to be close to a worst-case scenario: assuming that future participants will be willing to follow instructions to employ light dimming and avoid light from windows, it is unlikely that they will choose light conditions such as the ones tested here. The light sensor was positioned between the chinrest and the head-bar where the right and the left eye would be. Chinrest and head-bar were adjusted for a test situation with a subject of 170 cm size, sitting at a table with a height of 75 cm at a chair that was adjusted for a height of 48 cm. The chinrest was adjusted at a height of 35 cm relative to the table and the head bar was adjusted at a height of 50 cm relative to the table. The measurement with the cover placed over the MASAVE was taken at the

same location and measurement point, right after the baseline measurement.

Results

Two measurements, one with and one without the cover, were taken and compared. The baseline measurement without the cover gave a value of 520 Lux, which is a normal value for office rooms according to [28]. The same measurement, at same daytime, was done with the cover on, and this gave .08 Lux. This is less than .01 footcandle (=.11 Lux) suggested by [29] for PUI measurements.

Experiment II: Usability, PUI measurements, and environmental conditions

The goal of the second experiment was to validate if the MASAVE device can be set up by naïve users, if valid pupillometric measurements can be done with the device, and how the environmental conditions under the textile cover change when using it over an extended period of time. To this end, we set up a study that was typical for the type of study we had in mind when we designed the device. Participants were instructed to imagine that the preassembled MASAVE arrived with a manual and that they had to set up the device on their own. After this was done correctly the experimenter could take over remotely and start the experiment. In practice, the remote access of the experimenter could be realized with, for example, Windows Remote Desktop. For some experiments, the intervention could be as little as entering the subject ID or, when MASAVE would be sent to one person at a time, MASAVE could be preconfigured entirely so that the experiment starts automatically after setting it up and connecting the power plug. The experience of the subject was operationalized with two NASA-TLX questionnaires [30]. To test whether valid pupillometry can be done with the device we measured a PUI before and after the experimental condition. The textile cover was therefore applied over the entire duration of the measurement. We also measured temperature, CO_2 concentrations and relative H_2O concentrations of the air under the cover at multiple time points over the session.

Methods

Subjects

Nineteen student subjects (five male, thirteen female, one diverse) aged between 20 and 32 years were invited to the study and eighteen subjects completed the study. One subject aborted the study early because the subject perceived the experimental hearing test as too hard. All subjects were German native speakers and had self-proclaimed normal hearing. Subjects were paid 12€ per hour for their participation and gave written informed consent before testing.

Procedure: Usability testing

For the usability part of the study, subjects had to interact with the MASAVE according to a manual that was available to them. The manual can be found in the Supplemental material. The intention was to simulate the situation of sending the MASAVE to naïve participants and letting them set up the device without help. This situation was split into two parts. The first part was the physical setup that included: checking whether the intended space was big enough, lifting the device from the floor to the dedicated area, connecting the power line cable and switching on the PC. The second part included adjusting the chinrest and headrest, pulling on the textile cover, and calibrating the eye-tracker with the assistance of the remotely connected experimenter.

The experimenters observed the subjects without intervening, wrote down whether the subjects were able to perform the tasks and handed out the NASA-TLX questionnaire [31] after the subjects had completed each of the two parts. The NASA-TLX questionnaire consists of six items that represent the subscales "Mental", "Physical", and "Temporal" demands, "Frustration", "Effort" and "Performance". We used the German translation of the questionnaire from [32]. Each item is rated on a scale ranging from low to high, except "Performance" which is rated from good to poor. The scale is segmented into 20 unlabeled areas. For analysis we used the so called Raw TLX where the scores are transformed from the unlabeled scale to a scale that ranges from 0 to 100 and are then averaged over the six subscales. Since the NASA-TLX can be used in a variety of areas, it is difficult to provide benchmarks for comparison. However, [33] performed a meta-analysis and provided empirical data for different task types. Our experiment can be categorized best as a "mechanical task" which encompasses assembly tasks, crane operation, and mechanical maintenance. We will use these data to evaluate our own findings.

Procedure: Pupillometry

After this initial part of the study, subjects had to do a psychoacoustical experiment that had previously been performed under lab conditions and was performed here in a regular office. The shutters of the windows were put down and the door of the office was closed during the experiment which reflects the recommendations we would suggest also for home testing. The psychoacoustical experiment was set up as a dual-task experiment with two task types. The first task was a memory task. Participants had to listen to a matrix sentence and memorize the first, third, and fifth word. The sentence was presented in quiet (no masker) at a clearly audible level. This level was determined by the subjects themselves to ensure that the performance at this task was not dependent on the sensual performance of the subjects. The second task was to recognize another matrix sentence played after the first sentence. This second sentence was masked by competing talkers also uttering matrix sentences. These stimuli were the same as used in

the study before, but we only used the unprocessed collocated condition with different fixed SNRs (10, 5, 0, -5, -10 and -15 dB) to limit this experiment to a single experimental session. We chose the unprocessed collocated condition because it was the most difficult condition in the experiment described above and, presumably, would produce fatigue to occur which we intended to measure using pupillometry here. The SNRs were chosen to ensure that the task was challenging for the subjects but still doable (expected performance mainly above 50% recognition rate based on experiment I).

These SNR conditions were organized in blocks of 20 items and the blocks were presented in randomized order. After each block subjects had to rate their listening effort on a categorical scale that was taken from the ACALES [25]. At the very beginning of the experiment, subjects had to perform a Multi-Dimensional Fatigue Inventory (MFI) [34] and a measure of the Pupil Unrest Index (PUI) [7]. These measures were taken as a baseline for subjective fatigue and physiological fatigue. After the experiment subjects had to perform those measurements again. The difference compared to the baseline measurements was used to estimate fatigue induced by the measurement session. The MFI was also measured in the middle of the experiment to get hold of a potential nonlinear development of fatigue over time.

The calculation of the PUI consists of an artifact rejection step and the actual calculation of the PUI. Before performing those steps, we did an initial quality check of the data which was basically a rule-based version of a visual inspection. First, it was calculated whether a dataset had more than 15% missing data and, second, we checked whether a dataset had more than 5 s periods of missing or identical values. None of our datasets failed these criteria. However, these steps are not necessary for the PUI and the steps provided by [7] are sufficient for the calculation of the PUI.

For the artefact rejection of [7], for each .4 s segment the mean was calculated and the difference for each of those values from the mean was used as an indicator for an artifact. [7] suggests that deviations greater than .1 mm should be excluded from the analysis. This artifact reduction was done such that the window progressed one value at a time and each deviation greater than .1 mm was set as missing value. For the calculation of the PUI, first the data were reduced by calculating the moving average of 39 consecutive values which results in a simple low pass filtering of the data. The absolute values of the difference between each of the consecutive values, of the low pass filtered data, is further summarized for segments of 82 s. This sum is then normalized to 1 min and resembles the PUI for the current time segment. The overall PUI is then the average of all those segments. By abiding to the protocol and the algorithm of [7] we would be able to compare our results to results from other studies, like for example [35] which provided standard values for a large population of adult men and women. This psychoacoustical experiment was performed for about 45 min because, in a pilot study under lab conditions, we found that this procedure seems to cause fatigue that can be measured with the PUI.

Technical details of the pupillometry

The pupil dilation was measured with the above mentioned Gazept GP3 HD eye-tracker which was used with a sampling rate of 60 Hz. The Gazept GP3 HD can be accessed via server software that is installed on the host PC. Via internet protocol the server can be configured and the data from the eye-tracker can be received from the server. The handling of all those tasks as well as the overall experiment was programed in MATLAB [12]. The audio stimuli were presented with SoundMexPro [36] which is a plugin for MATLAB. The processing of the raw data was also done in MATLAB but more sophisticated statistical analysis were done in R [26].

Procedure: Environmental conditions under the MASAVE cover

To test how the environmental conditions under the cover of MASAVE changed over the course of the measurement session, temperature, CO_2 concentrations, and relative H_2O concentrations of the air were measured. Therefore, we logged data with a CO_2-monitor (Air Co2ntrol 5000 [37]) that was placed right behind the display (see Fig. 1). The logging started when participants performed the first PUI and ended after they had performed the second PUI at the end of the session. Participants were instructed to open the cover whenever they felt uncomfortable to perform the task with the cover on. Hence, participants performed under the cover for a maximum of around one hour (11 min PUI two times + ~45 min experiment).

Results

Results: Usability

The analysis of the NASA-TLX data resulted in a median rating of the first part of the evaluation of 5.8 with a standard deviation of 11.1 and in a median rating of the second part of 11.7 with a standard deviation of 15.8 (see Fig. 9). As already mentioned we use empirical data of [33] for so-called "mechanical tasks" as a comparison. In [33] this category comprised 22 studies. The averaged minima were at 20.1 score points and a median rating of 28 score points.

Results: Pupillometry

Figure 10 shows the change in PUI from the first to the second measurement. Almost all subjects had a higher PUI value (i.e., a positive difference) after they had performed the experiment which means that they were more fatigued after the experiment than before. This impression was confirmed by a one-sample t-test which resulted in highly significant results ($p < .001$, $t = 5.24$) with a confidence interval ranging from .09 to .2 mm/min which means that sleepiness was higher after the test than before. Hence, we conclude that the measurement of the PUI is feasible with the MASAVE device.
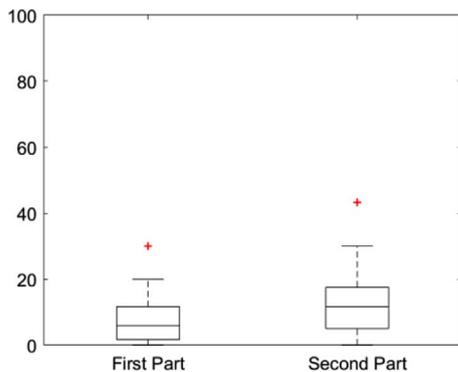
Figure 9. NASA-TLX scores obtained for the first and the second part of the evaluation.
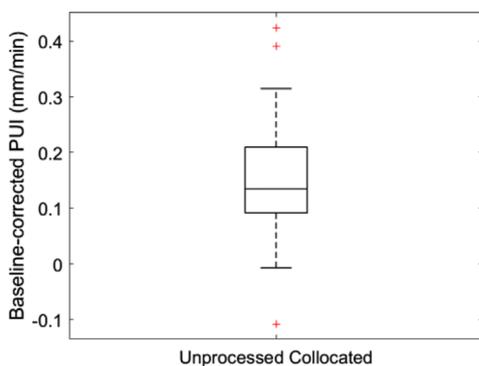

Figure 10. Baseline-corrected PUI measurement.

Figure 11. Mean CO_2 levels (top), relative humidity (middle); and temperature (bottom) logged over the course of the session. Error bars represent two standard deviations.

Results: Environmental conditions

One key observation was that all our participants completed the entire study under the cover although they were explicitly instructed that they could lift the cover and complete the study that way if they felt the need to do so. Figure 11 summarizes the logged CO_2 concentration (top), humidity (middle), and temperature (bottom) collected every 6 min during the session. Error bars represent two standard deviations. As expected, all three parameters seemed to increase continuously over time. CO_2 started with an average level of about 500 ppm and increased to about 3000 ppm at the end of the session. None of our subjects reached the maximum exposure level for a repeated, long-term exposure for an eight hour workday of 5000 ppm as specified in [38].

The increase in relative humidity seemed to be less pronounced (from about 31% to about 38%).

The temperature was about 21 °C (Celsius) at the beginning of the measurements. The average increase was about 6 °C. However, for some subjects it rose to slightly above 30 °C.

Discussion

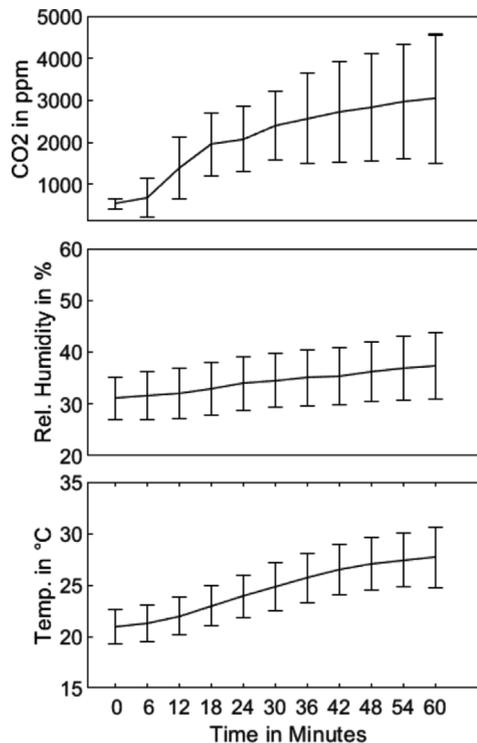The validation experiments of the proposed MASAVE system aimed to test the functional feasibility of conducting listening tests and pupillometry as well as aspects of usability and handling. Apart from functional aspects, our design goals required the MASAVE to be lightweight, not very expensive and it should be possible to build it quickly. These aspects, as well as remaining limitations, are discussed in the following.

The first study investigated whether the performance in listening tests conducted with the MASAVE in less controlled environments is comparable to the performance in a standard listening booth. The experimental scenario we chose was considered to approximate a worst-case scenario in terms of steady noise for the intended use case of our device (clearly audible fan noise). In actual experiments, we would expect subjects to set up the MASAVE in a significantly quieter environment than the lobby tested in this study if they are instructed to perform the test in a quiet room.

Regarding the validity of data gathered with the MASAVE under suboptimal conditions compared to data gathered in a standard listening booth, we had expected a small difference in SRT and listening effort for the conditions with a high noise portion in the experimental signal (unprocessed, BMVDR) and a bigger difference for conditions with very low SRTs and masker levels (IBM). The underlying reasoning was that low levels of noise in the signal should enable listeners to achieve very low SRTs respectively to rate the signals as easier to listen to. In such situations the external noise should have a proportionally stronger effect and differences between booth and lobby were expected to be strongest.

However, we found no statistical differences between the two tested locations. Nevertheless, we would not suggest testing in such an environment. When we compare the 38 dBA we observed with the values obtained in [39] we see that this room falls between a bedroom and living room. Most participants will have a room that is more silent than this and therefore it should be possible to gather reliable data.

Of course, our approach is not an alternative for every type of listening experiment. Experiments that use very low levels of noise will probably experience a greater impact of uncontrollable, external noise. Such experiments could be, for example, SRT measurements without noise or audiometric measures. To get more control over the test situation, especially in such critical studies, it would be desirable to monitor the external noise, e.g., to ensure that predefined thresholds are not exceeded during the runtime of the experiment and/or to repeat trials that are compromised. The auditory environment can be monitored with the built-in microphone of the proposed system. However, special consideration has to be given to privacy issues (e.g., noise level analysis would have to be made on the fly and not based on recordings, unless the subjects agree to audio recordings or other suitable measures to protect the subject's data privacy are implemented). Obviously, this also means that studies that are not very sensitive to occasional breaches of a threshold for external noise are more suitable to be performed with MASAVE than others.

The second study was concerned with the question whether MASAVE can be used and set up by naïve subjects. All subjects were able to setup the device and in general rated the usability of the device as very good. Although it is difficult to assess usability on an objective and comparable level, besides the simple metric of success or failure, the data obtained with the NASA-TLX seems to be promising and handling MASAVE can be classified as very easy compared to other mechanical tasks [33], at least for relatively young subjects as tested here. The related question of the environmental conditions under the textile cover was also addressed in this study. It was found that all measured environmental conditions under the textile cover were far from being harmful over the period tested here, but that the temperature under the cover might be an aspect that would need to be considered when the cover is applied for longer than approximately 30 min, or when the room temperature is already high.

The textile cover is an essential prerequisite of MASAVE to control the lighting conditions for pupillometry. Since we intended to measure the PUI [7] with MASAVE, the cover's effectivity to block out light was measured. Again, we employed conditions that are worse (much brighter) than can be reasonably expected in home studies if subjects are instructed to darken the room. We found that the textile cover provided sufficient darkness to fulfill the requirements proposed in the literature and we could measure the PUI under realistic conditions.

In a later iteration of the MASAVE we intend to implement an Arduino-powered light sensor at the head-bar to monitor possible light intrusions. The Arduino can be interfaced with Matlab and therefore we could monitor and log the lighting conditions, make an estimate of the validity of the measures and/or give automated feedback to participants and the supervisor.

For other pupillometric parameters, like for example mean pupil dilation, it could be necessary to fully control the light conditions, instead of blocking out all light, to enable the full dynamic range of the pupil. This would then also be possible by including LED lamps that are controlled via the Arduino and the light sensor to dynamically adapt the lighting conditions.

Our approach to cover MASAVE and the participant under a cover to control lighting conditions, has the drawback of preventing regular air exchange. This has an impact on environmental conditions under the cover. CO_2 and H_2O levels stayed within reasonable limits during the 1 h test period. For some participants, however, temperature reached about 30 °C which could affect the outcome of the experiment [40] and potentially could make subjects uncomfortable. According to [40], performing office tasks at a temperature of about 30 °C can result in a relative performance drop of up to 10% compared to the optimal temperature of around 21 °C. But most of our participants did not exceed 30 °C even after the whole test period.

Anyway, using MASAVE with the cover in an already very hot room, seems not advisable. However, if an experiment can be divided into blocks of about 30 min and take pauses for ventilation, we assume that there will be no consequences for the outcome of an experiment. It is also worth mentioning that none of the participants quit the study because of the environmental conditions but instead we had one subject that quit because the experimental hearing task was too difficult. Furthermore, the study design of, e.g., PUI measurements could also be organized such that the cover is not applied for the entire duration of the measurement session (as in this study), but rather only when required to conduct pupillometry. In the second study we also measured the PUI of our subjects. We found that it is possible to employ this procedure with MASAVE and that it provides valid data.

Finally, we also want to comment on our goals to build a lightweight, relatively cheap, and quickly to build device. The MASAVE weighs about 10 kg, which is light enough to be sent by mail or to transport the MASAVE to any mobile testing site. However, the present design might be a bit too bulky and still too heavy for some subject groups like for example feeble people, older people, children, and people with certain disabilities, which might require support when receiving and setting up the MASAVE.

At the time of conducting our study, the MASAVE cost about 2800€ (excluding VAT) including all hardware and components. This is not expensive in relation to a full lab space. Considering that there is no real need for a dedicated room to employ the MASAVE, it might be a good and affordable opportunity to scale up the measurement capabilities to run listening tests.

We further wanted to be able to build the MASAVE within a short time. We estimate that the MASAVE as proposed here can be replicated within a week. A fused

filament fabrication (fff) printer with a bed size of at least 330 × 330 mm would considerably reduce the effort and time of printing since parts like the chin-bar or the head-bar could be printed in one piece. Therefore, if a certain ability to work with plastic is available and no big changes need to be done on the present design, it can be assumed that the MASAVE can be a solution in situations that require a quick alternative to classic lab testing.

In summary, we conclude that the MASAVE is a cheap, fast to build, and lightweight alternative to testing in listening booths. Results that are gathered with the device are comparable to the results gathered in a standard sound-attenuating listening booth and the environmental conditions under the textile cover will probably not influence study outcomes. This makes the MASAVE an alternative or complementary solution for lab-testing and may offer new possibilities for testing large groups of subjects to avoid underpowered studies.

We hope we contributed a testing approach for other researchers that provides a convenient and cost-effective way of scaling up measurement capabilities.

Disclosure statement

Supplementary material

The supplementary material of this article is available at https://acta-acustica.edpsciences.org/10.1051/aacus/2022049/olm.

References

1. M.J. Brandt, H. IJzerman, A. Dijksterhuis, F.J. Farach, J. Geller, R. Giner-Sorolla, J.A. Grange, M. Perugini, J.R. Spies, A. Van't Veer: The replication recipe: What makes for a convincing replication? Journal of Experimental Social Psychology 50 (2014) 217–224.
2. M. Brysbaert: How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables, Journal of Cognition 2 (2019) 1–38.
3. C. Senior, M. Smith: The internet. A possible research tool? The Psychologist 12 (1999) 442–444.
4. M.J. Crump, J.V. McDonnell, T.M. Gureckis: Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. PloS One 8 (2013) e57410.
5. H. van Steenbergen, B.R. Bocanegra: Promises and pitfalls of Web-based experimentation in the advance of replicable psychological science: A reply to plant (2015). Behavior Research Methods 48 (2016) 1713–1717.
6. F. Pausch, J. Fels: MobiLab – A mobile laboratory for on-site listening experiments in virtual acoustic environments. Acta Acustica united with Acustica 105 (2019) 875–887.
7. H. Lüdtke, B. Wilhelm, M. Adler, F. Schaeffel, H. Wilhelm: Mathematical procedures in data recording and processing of pupillary fatigue waves. Vision Research 38 (1998) 2889–2896.
8. E. Strasser: MASAVE: A mobile test system for audio-visual experiments at home. https://www.thingiverse.com/thing:5172334. Accessed 31.12.2021.
9. A. Murphy: RestEasy – An open source chin rest for human psychophysics experiments. https://github.com/nimh-nif/SCNI_Toolbar/wiki/RestEasy:-An-open-source-chin-rest-for-human-psychophysics-experiments. Accessed: 27.08.2021.
10. TX-Team: Fanless IS1000. https://shop.tx-team.de/Box-IPC/Fanless-IS1000::2.html. Accessed: 27.08.2021.
11. gazepoint: GP3 HD Eye-Tracker. https://www.gazept.com. Accessed: 30.08.2021.
12. T. M. Inc., MATLAB, The MathWorks Inc., Massachusetts, 2020.
13. F. Plc: scarlett-2i2. https://focusrite.com/de/audio-interface/scarlett/scarlett-2i2. Accessed: 30.12.2021.
14. G. Kidd, H.S. Colburn: Informational masking in speech recognition. In: John C. Middlebrooks, Jonathan Z. Simon, Arthur N. Popper, Richard R. Fay, Eds. The Auditory System at the Cocktail Party. Springer Handbook of Auditory Research, Springer, 2017, pp. 75–109.
15. D. Lakens: Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. Social psychological and personality science 8 (2017) 355–362.
16. K. Wagener, T. Brand, B. Kollmeier: Entwicklung und Evaluation eines Satztests in deutscher Sprache Teil III: Evaluation des Oldenburger Satztests (in German). Development and evaluation of a German sentence test–Part III: Evaluation of the Oldenburg sentence test. Zeitschrift für Audiologie 38 (1999) 86–95.
17. S. Hochmuth, B. Kollmeier, B. Shinn-Cunningham: The relation between acoustic-phonetic properties and speech intelligibility in noise across languages and talkers, in Proceedings of German Annual Conference on Acoustics, 2018, pp. 628–629.
18. J. Rennies, V. Best, E. Roverud, G. Kidd Jr.: Energetic and informational components of speech-on-speech masking in binaural speech intelligibility and perceived listening effort. Trends in Hearing 23 (2019) 2331216519854597.
19. H. Kayser, S.D. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, B. Kollmeier: Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses. EURASIP Journal on Advances in Signal Processing (2009) 1–10.
20. D. Marquardt, S. Doclo: Interaural coherence preservation for binaural noise reduction using partial noise estimation and spectral postfiltering. IEEE/ACM Transactions on Audio, Speech, and Language Processing 26 (2018) 1261–1274.
21. C.F. Hauth, N. Gößling, T. Brand: Performance prediction of the binauralMVDR beamformer with partial noise estimation using a binaural speech intelligibility model, in Speech Communication; 13th ITG-Symposium, 2018, pp. 1–5.
22. D.S. Brungart, P.S. Chang, B.D. Simpson, D. Wang: Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation. The Journal of the Acoustical Society of America 120 (2006) 4007–4018.
23. S.D. Ewert: AFC – A modular framework for running psychoacoustic experiments and computational perception models, in Proceedings of the International Conference on Acoustics AIA-DAGA, 2013, pp. 1326–1329.
24. T. Brand, V. Hohmann: An adaptive procedure for categorical loudness scaling. The Journal of the Acoustical Society of America 112 (2002) 1597–1604.

25. M. Krueger, S. Michael, H. Inga: Entwicklung einer adaptiven Skalierungsmethode zur Ermittlung der subjektiven Höranstrengung, in Proceedings of German Annual Conference on Acoustics, 2015.

26. R.C. Team: A language and environment for statistical computing. MSOR Connections (2014).

27. D. Lakens: Introduction to equivalence testing with TOSTER. Advances in Methods and Practices in Psychological Science 1 (2018) 259–269.

28. A.f. Arbeitsstätten: Technische Regeln für Arbeitsstätten Beleuchtung (ASR A3.4). Issued by Bund, BAuA, 2011.

29. J.W. McLaren, J.C. Erie, R.F. Brubaker: Computerized analysis of pupillograms in studies of alertness. Investigative Ophthalmology & Visual Science 33 (1992) 671–676.

30. S.G. Hart: NASA-task load index (NASA-TLX); 20 years later, in Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 2006, pp. 904–908.

31. S.G. Hart, L.E. Staveland: Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In: Advances in psychology, Elsevier, 1988, pp. 139–183.

32. M. Sepehr: Assessment of subjective mental workload using NASA-task load index, in Proceedings of the 7th European Annual Conference On Human Decision Making And Manual Control, 1988, 69–75.

33. R.A. Grier: How high is high? A meta-analysis of NASA-TLX global workload scores, in Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 2015, pp. 1727–1731.

34. E. Smets, B. Garssen, B. Bonke, J.C. De Haes: The multidimensional fatigue inventory (MFI) psychometric qualities of an instrument to assess fatigue. Journal of Psychosomatic Research 39 (1995) 315–325.

35. B. Wilhelm, A. Körner, K. Heldmaier, K. Moll, H. Wilhelm, H. Lüdtke: Normwerte des pupillographischen Schläfrigkeitstests für Frauen und Männer zwischen 20 und 60 Jahren. Normal values of the pupillographic sleepiness test in male and female subjects aged 20 to 60 years. Somnologie 5 (2001) 115–120.

36. H. Oldenburg (Ed.), SoundMexPro, Hörzentrum Oldenburg, Oldenburg, 2020.

37. DOSTMANN-electronic-GmbH: Air Co2ntrol 5000 https://dostmann-electronic.de/product/aircontrol-5000-co2-istrument-with-data-logger.html?cid=6. Accessed: 18.05.

38. Deutsche Forschungsgemeinschaft: MAK-und BAT-Werte-Liste 2021. Senatskommission zur Prüfung gesundheitsschädlicher Arbeitsstoffe, Mitteilung 57, 2021.

39. A.S.o. America: ANSI/ASA S12.2-2019: Criteria For Evaluating Room Noise, 2019.

40. O. Seppanen, W.J. Fisk, Q. Lei: Effect of temperature on task performance in office environment, in 5th International Conference on Cold Climate Heating, Ventilating and Air Conditioning, 2006.