Enhancing Screen Reader Intelligibility in Noisy Environments

Dragan Ahmetovic, Gabriele Galimberti, Federico Avanzini, Cristian Bernareggi,
Luca Andrea Ludovico, Giorgio Presti, Gianluca Vasco, Sergio Mascetti
Department of Computer Science, Università degli Studi di Milano, Italy

Abstract—People with blindness or severe low-vision access mobile devices using screen readers. However, noisy environments can impair screen reader intelligibility. During mobility, this could disorient or even endanger the user. To address this issue, we propose three screen reader speech compensation techniques based on environmental noise: speech rate slowing, adaptive volume increase, and adaptive equalization.

Through a study with twelve participants in three simulated noise scenarios, we evaluate screen reader intelligibility and the perceived distraction from the soundscape, with and without compensations. Four of the proposed compensations, in particular those that pair speech rate reduction with volume or equalization adaptation, significantly improve screen reader's speech intelligibility in all considered scenarios, and the compensations do not have a significant impact on the distraction from the soundscape.

Index Terms—Speech compensation, visual impairments.

I. INTRODUCTION

People with Blindness or Severe Low-Vision (BSLV) access mobile devices through screen readers [1]. These accessibility services augment touchscreen interactions with verbal descriptions of the explored screen content, enabling non-visual access to Graphical User Interface (GUI) elements [2]. However, during mobility, the intelligibility of the screen reader feedback can be made difficult by the presence of environmental sounds, such as traffic noise or people's voices [3].

Some users address this problem by manually increasing the screen reader volume when needed [4]. This solution only partially mitigates the problem because, in case of sudden noise, the user may not have time to adapt the volume. Also, changing the volume can distract the user from the current activity. Keeping the screen reader at its maximum volume is not a solution as well, because the screen reader feedback could be perceived as intrusive, it may spotlight the user's disability [5], or it could cover important sound cues from the environment, making mobility more difficult or even endangering the user [3]. Another possible solution is to reduce the screen reader speech rate to improve speech understanding in presence of noise [6]. However, this approach also reduces the information throughput of the screen reader which might not be desired.

To address these issues, our proposal is to dynamically adapt the screen reader output based on environmental noise. The goal is to improve the screen reader intelligibility, without increasing the distraction from the surrounding soundscape and without permanently reducing the screen reader speech rate.

To this end, we designed three compensation techniques:

- **Rate** – Applies a flat speech rate reduction in presence of environmental noise;
- **Vol** – Adaptively increases or decreases speech volume based on the intensity of environmental noise;
- **Eq** – Adaptively increases or decreases speech volume only for frequencies impacted by environmental noise.

A preliminary evaluation, conducted with 4 participants with BSLV, assessed the effect of the proposed techniques on the screen reader speech intelligibility in presence of background noise. The participants commented that making the speech both louder and slower would make it easier to comprehend. Thus, in a second preliminary evaluation, conducted with other 6 representative participants, we also included conditions obtained as combinations of speech slowing with the other two compensations: **Rate+Vol** and **Rate+Eq**. A final evaluation, conducted with 12 participants, also evaluated the distraction caused by the screen reader speech with respect to the background soundscape.

Specifically, in the main study, we evaluated the compensations in three typical noise scenarios [4]: **Crowd**, **Traffic**, and **Subway**. The scenarios were realistically simulated in a silent chamber with a quadraphonic audio setup, playing 4-channel real-world recordings at a sound pressure level consistent with the real situation. For the experiment, we used a corpus of text sentences [7], read by the screen reader and conveyed through bone conduction headphones, with and without compensation. We measured speech intelligibility as the percentage of the correctly understood words, and the distraction caused by the screen reader speech with respect to the soundscape by asking the participants to pinpoint the direction of a contextual sound played concurrently with the speech feedback.

The results confirm that noisy environments severely impact screen reader speech intelligibility but compensation approaches can mitigate this effect significantly. In particular, the combined compensations (**Rate+Vol** and **Rate+Eq**) provide consistent improvement in all soundscape scenarios. The perceived distraction from the surrounding soundscape, caused by the screen reader speech, does not significantly change when compensations are used. This indicates that the proposed compensations can improve speech intelligibility without impacting the user's ability to pay attention to the surrounding environment. Therefore, our approach is a practical solution to improve the understanding of screen reader speech feedback in noisy environments and during mobility.

II. RELATED WORK

Mobile devices are convenient for accessing information ubiquitously, and, for this reason, they are frequently used by many people with disabilities [8]. They can also be used to provide assistive capabilities, including access to visual cues detected with computer vision techniques [9], support in social networking [10], and in particular mobility and navigation assistance on-the-go [11], [12], [13]. Due to these functionalities, almost the entirety of people with BSLV in developed countries use mobile devices, and over 80% of them use smartphones for mobility assistance [14].

A. Challenges in Mobility Screen Reader Use

People with BSLV commonly use mobile devices through screen readers [2], which can be employed to access most native functionalities [8] as well as assistive capabilities provided by third-party developers. However, screen reader usage during mobility presents three main issues. First, many people with BSLV feel that the speech feedback is intrusive to people around them, or even perceived negatively by others [5]. This effect is even more apparent in people with invisible disabilities, like mild low-vision [15].

Second, the environmental noise may mask screen reader speech [16], preventing people from hearing it. This problem is especially present in very loud scenarios, such as the subway, crowded places, and while walking in the traffic [4]. Because of this, people with BSLV frequently use headphones while listening to screen reader during mobility [17].

Third, the soundscape may also be partially covered by the screen reader speech, in particular when using headphones [18]. Thus, the speech could draw away the user's attention from the environmental sound cues which may be useful for orientation (*e.g.*, acoustic traffic signals) or for avoiding dangerous situations (*e.g.*, an approaching car). Most users partially offset such issues by using a single headphone, while only a minority adopt pass-through or bone conduction headphones due to high prices, unfamiliarity with these solutions, and concerns with their output sound quality level [17].

B. Audio Adaptation in Presence of Noise

In the context of accessibility and acoustic communication for people with BSLV, prior works have noted the impact of noise on screen reader speech intelligibility [6], in particular in urban scenarios such as traffic, crowded environments, and travel hubs such as subway stations [4]. In such cases, the manual adaptations of speech rate [6] or volume [4] were reported as possible coping mechanisms. However, as noted above, the users may not have the time to manually change the setting in case of sudden noise and the action can distract them from their current activity.

Approaches for the automatic adaptation of sound based on ambient noise, especially in mobile contexts, have been proposed using several different techniques. These techniques can be grouped into two broad clusters: active noise cancellation and noise-based processing of the useful (*i.e.* information carrying) audio signal. In the first cluster, headphones or headsets based on active noise-cancellation use adaptive audio processing for reducing the background noise [19]. However, noise-canceling solutions are not suited for the purposes of this work since important sound cues from the environment may be rendered inaudible, making mobility more difficult or even endangering the user [3].

The second cluster comprises methods that automatically adjust volume, dynamic range, or other sound features based on ambient noise, which is in turn monitored and analyzed through built-in microphones on mobile devices. One example is the adaptive control of the smartphone volume based on user activity and ambient noise, with the purpose of improving the perception and recognition of alert and notification sounds in noisy environments [20]. Another example is the adaptive control of the dynamic range (compression) of the audio being played, depending on the level of the environmental noise around the listener, which is measured using the microphone on the mobile device [21]. None of the previous works address the challenge of improving screen reader intelligibility. Instead, in this paper, we propose and evaluate audio adaptation techniques specifically designed for this purpose.

III. COMPENSATION TECHNIQUES

We designed three base compensation techniques (**Rate**, **Vol** and **Eq**) and two techniques combining **Rate** with the remaining compensations (**Rate+Vol** and **Rate+Eq**). In the following, for brevity, we indicate the absence of a compensation technique with the term **None** (meaning that no compensation technique is used).

The compensation techniques have been designed to be deployable on mobile devices, without requiring proprietary hardware: they only need an audio output device (*e.g.*, headphones), a microphone (possibly the one integrated in the headphones), and limited computational power.

A. Adaptive Volume

The key idea of dynamic volume compensation (**Vol**) is to adjust the speech volume in an adaptive way, depending on the environmental noise that surrounds the device. The goal is to keep the *Signal to Noise Ratio* (SNR) between the speech level and the soundscape level constant but, at the same time, to limit the intervention to an acceptable range (in order to avoid lowering the volume too much in quiet scenarios or saturating the device in loud scenarios).

To achieve this, a gain factor $k(t)$ is computed as the ratio between the target SNR, which can be tuned based on user's hearing and preferences, and a running measure of the actual SNR, denoted in Eq. (1) by SNR_t and $SNR_r(t)$ respectively.

$$k(t) = \frac{SNR_t}{SNR_r(t)} = \frac{SNR_t}{\frac{RMS_{\text{signal}}(t)}{RMS_{\text{noise}}(t)}} = SNR_t \cdot \frac{RMS_{\text{noise}}(t)}{RMS_{\text{signal}}(t)},$$
(1)

In Eq. (1) *RMS* stands for *Root Mean Square*, more specifically, the *running* RMS, *i.e.* the RMS evaluated on a moving temporal window. The larger the window, the smoother the intervention, at the cost of slower responsiveness. This processing completely occurs in the time domain. **Vol** compensation
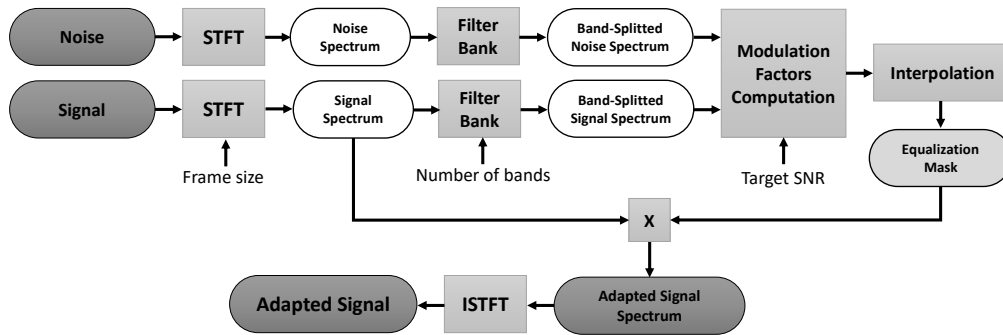
Fig. 1. Adaptive sonification based on spectral equalization.

can be fine-tuned by setting two main parameters: SNR_t, and the running RMS window size. Moreover, in order to avoid excessive volume compensation (low k values would render the sound almost inaudible in quiet environments, while high values would result in audible signal distortion and may cause pain or even hearing damage), we added the possibility of limiting $k(t)$ to a range $R_k = [k_{min}, k_{max}]$.

We chose the value $SNR_t = +1.2$ as it was found to be the preferred value for speech signals in a previous study on acceptable noise levels and preferred SNRs for speech and music [22]. Values for $R_k = [-3, +20]$ dB were chosen in order to avoid signal clipping as well as to minimize attenuation. For the estimation of the running RMS, we used a temporal window with duration 46ms and an overlap of 23ms between successive windows, in order to minimize artifacts (glitches) resulting from an exceedingly fast adaptation.

B. Adaptive Equalization

The goal of this technique is to use spectral modifications, rather than volume adjustment, to reduce the masking of the verbal message by the background noise. In fact, the concept of masking does not merely depend on volume; rather, it consists in the alteration of the perception of the single frequency components of the message if the noise occupies the same spectrum region with sufficiently high energy. Thus, adaptive Equalization compensation (**Eq**) accounts for the proper *auditory masking* effect by realizing a multi-band version of **Vol**. First, the signal is split into a number of frequency bands, then a different modulation factor is computed for each band, so as to reduce masking effects only where needed.

Fig. 1 shows the process to compute the adapted signal. The first step is to take the *Short Time Fourier Transform* (STFT), to extract the spectra of both the signal and the noise. Then, these spectra are processed through an ad-hoc filter bank to divide them into a given number of bands. Ideally, the band subdivision should approximate what in psychoacoustics is called a *critical band*, that is the band of audio frequencies within which a second tone will interfere with the perception of the first tone by auditory masking, since they activate the same area of the basilar membrane [23]. Equivalently, this filter bank should serve the purpose of simulating the human auditory system by modeling those band-pass filters (whose

width follows the ERB scale [24]). Nevertheless, when dealing with speech signals, such a fine subdivision in frequency bands followed by a per-band gain modulation may result in a complete cancellation of vocal formants, which are the spectral characteristics of vocal signal conveying information. To avoid this loss of information, we opted for a reduced number of bands. Finally, a different gain factor is calculated for each band (whereas for **Vol** a single modulation factor is applied to the whole signal). The per-band gain factors are computed with the same approach summarized in Eq. (1), the only difference being that values are now computed for each band.

To apply the computed gain changes, instead of multiplying each band for the corresponding factor and summing the bands together (which would introduce some phase issues), we interpolate the factors so to obtain an equalization curve with the same resolution of the STFT frames. In particular, we use a Piecewise Cubic Hermite Interpolating Polynomial, which has the property of preserving the maximum and minimum points of the initial function, thus avoiding equalization overshooting or undershooting. Once the desired equalization curve is computed, multiplying it by the spectrum of the signal returns the spectrum of the adaptive signal. The latter is brought back into the time domain through the Inverse Short Time Fourier Transform (ISTFT), thus obtaining the adaptive audio signal.

For all the parameters in common with the **Vol** compensation, we chose the same values. In addition to those, **Eq** compensation requires one additional parameter, which is the number of bands. We used 5 bands, which resulted in good preservation of the vocal formants, as discussed above.

C. Adaptive Speech Rate

People with BSLV have higher listening rates than sighted people [6] and usually keep the screen reader at the highest possible speed allowing them to understand the speech in a quiet environment, as also confirmed by our participants. However, it is yet unclear how noise impacts screen reader intelligibility by people with BSLV [6]. Additionally, as reported in prior works, expert screen reader users often change their screen reader speech rate based on context [4], and this approach is also considered a possible adaptation to improve screen reader intelligibility in presence of noise [6].

Speech rate compensation (**Rate**) applies a flat reduction of the screen reader speech rate when noise is detected. We

implemented this as a 15% speed reduction with respect to the *base speech rate* (*i.e.*, participant's preferred speech rate). This parameter was empirically selected after interacting with the preliminary study participants: when we asked for their preferred speech rates, six out of ten answered with a range of rates and in four cases the range size was 15% (in the other two cases the range was 5%). To avoid the speech rate to change within a sentence or a single word, the solution does not adapt continuously based on the changes in the ambient noise. Investigating the design space of the continuous screen reader speech rate adaptation based on the changes in environmental noise is a possible future extension of our work.

D. Combined Compensations

Based on the comments from the participants in the preliminary studies, we have also explored the combinations of the proposed compensation techniques. The three base compensation techniques change the way speech is reproduced (**Vol**, **Eq**), or the way it is generated (**Rate**). Combining **Vol** and **Eq** is not meaningful, as it would merely reduce to an **Eq** technique with a different SNR. Instead, we designed two additional techniques that combine **Rate** with either **Vol** or **Eq**, called **Rate+Vol** and **Rate+Eq**, respectively. In these compensations, speech is generated at the speed defined by **Rate** and reproduced after applying **Vol** or **Eq**.

IV. EVALUATION

We conducted a set of user studies with participants to assess the effect of the proposed compensation techniques on speech intelligibility in the presence of noise, and on the distraction caused by the screen reader speech with respect to the environment soundscape. The research was approved by the Ethics Committee of our University (approval 49/22). As part of the iterative design process, we first evaluated our techniques with sighted participants [25], then we conducted three studies with representative participants with BSLV. The first two of these are preliminary studies, which are described in Section IV-A. The last one is the main study, which is described in detail in the rest of this section.

A. Preliminary Studies

Two preliminary studies were conducted with 4 and 6 participants with BSLV, respectively, as a part of our design and parameter tuning process. In the first study, we assessed the three base compensation techniques (**Rate**, **Vol**, **Eq**). For this, we simulated a noisy soundscape and reproduced speech feedback (possibly compensated) that the participants were asked to listen and repeat (see Section IV-C). Motivated by the participants' comments, which highlighted that compensating both sound intensity and speech rate could be useful, in the second study we introduced the combined compensations: **Rate+Vol** and **Rate+Eq**. We also corrected the compensation parameters, in particular, to avoid the crackling noise that would appear with specific sound frequencies on bone conduction headphones, as mentioned in Section IV-C.

Considering the evaluation methodology, we introduced two main changes between the preliminary studies and the main study. First, in the preliminary studies, we assessed distraction as a subjective measure reported by the participants. Instead, in the main study, we introduced an objective measure of the distraction caused by the speech feedback, by assessing the participants' ability to pinpoint the direction of an environmental sound, reproduced concurrently with the speech. The second difference regards the speech rate. In the preliminary studies, all participants used the same speech rates. Instead, in the main study, we configured the screen reader speech rates based on the preferred settings for each participant.

B. Audio Stimuli

In order to quantitatively assess both the speech intelligibility and distraction, we prepared a set of **audio stimuli**, each combining a **soundscape**, a **speech signal** that the participants are asked to listen and repeat despite the background soundscape, and a **contextual sound**, the direction of which the participants are asked to pinpoint while the speech feedback is played. Each **audio stimulus** is characterized by three properties: the soundscape, the speech signal compensation (if any), and the direction of the contextual sound. More specifically, audio stimuli are audio tracks with six channels. Four channels reproduce the soundscape with its contextual sound from one of three directions (*i.e.*, left, right, front) and two channels reproduce the speech signal. The speech signal is always reproduced at the same time offset with respect to the soundscape in order to have consistent listening conditions. Instead, each contextual sound is played in a random instant while the speech signal is played.

To generate the **speech signals**, we followed the approach proposed in [7]. Each sentence (in Italian) is constructed from a word table with 5 columns and 10 rows, by randomly picking one word from each column. The combination of the extracted words forms a sentence that is grammatically correct but semantically unpredictable. The resulting combinations have the advantage of focusing the participant's attention on the actual comprehensibility of the sentences. The sentences were reproduced using the *VoiceOver*[1] screen reader (female voice). The screen reader allows the user to specify a preferred speech rate, indicated as a percentage of the maximum speed. Thus, for each participant, we generated personalized **speech signals** by specifying the speech rate that the participant is used to.

Four different **soundscapes** were selected, three with a high and one with a low noise level. The soundscape with a low noise level (**Suburban**) represents a baseline condition in which the speech compensation is not needed. It consists of a recording of the noise in a suburban residential area with low traffic. The noisy soundscapes are a subway station during train arrival (**Subway**), a crowded local market (**Crowd**), and a trafficked city street (**Traffic**) as they were reported to be particularly challenging scenarios [4]. Each soundscape is trimmed in order to have a duration of 15 seconds.

For each soundscape, we also prepared a **contextual sound**: the sound of a closing gate in **Suburban**, the sound of a cash receipt being printed in **Crowd**, the sound of opening bus doors in **Traffic**, and the intermittent signal of opened doors

---

[1]https://www.apple.com/accessibility/vision/

of a subway in **Subway**. Each contextual sound exists in three variations: one is played from the participants' left, one from the right, and one in front of the participant. The recordings of the soundscapes and contextual sounds are available online[2].

C. Apparatus

The experimental set-up mimics a real-world scenario in which ambient noise reaches the listener's ears through a purely acoustical path, while the speech signal is delivered through headphones. Consequently, the audio stimuli described in the previous section were reproduced through two distinct audio streams: one for the soundscapes and the contextual sound, the other for the speech signal.

Regarding the soundscapes and the contextual sound, it was first necessary to recreate a realistic simulation, in order to provide participants with the impression of being immersed in an everyday acoustic environment. Such a goal was addressed through a quadraphonic reproduction system: four audio channels were routed to the corresponding speakers arranged into a square, with the listener in the center. The speakers are positioned on the front-left, front-right, back-left, and back-right with respect to the listener. With respect to stereophony, quadraphony allows a better spatialization of the sound and, consequently, the possibility for the listeners to better locate the sound events presented to them. The speakers were installed in a silent chamber (where the tests took place), a soundproof room acoustically treated to dampen sound reflections, thus enabling to simulate wide sound scenarios, despite its limited size. Contextual sounds are played by two speakers, depending on the direction. For example, when the contextual sound is played on the left of the participant, the two left speakers (*i.e.*, front-left, back-left) are used.

Speech signals were conveyed to the participant using Z8 Docooler bone conduction headphones. This choice had a twofold motivation: first, they ensure the maximum possible transparency with respect to the external soundscape, as they do not occlude the ear canal; second, they guarantee a good response in the characteristic speech frequency range (*i.e.* $200\,\mathrm{Hz}$ to $4\,\mathrm{kHz}$). Nevertheless, we experimentally verified that these specific headphones introduce an unpleasant vibration around $230\,\mathrm{Hz}$, therefore we processed all the headphone signals with a bell equalizer centered on the aforesaid frequency, with a filter gain of $-4.5\,\mathrm{dB}$, and an overall gain of $+2.4\,\mathrm{dB}$ in order to linearize the device response. We calibrated the system to produce an uncompensated speech signal with average $SNR = 0$ with respect to a baseline outdoor soundscape without strong sources of noise (*i.e.*, **Suburban**). This means that the uncompensated speech signal is tuned to be intelligible in such a scenario, without being intrusive to others. Experiments show that this is indeed the case (see Section V).

D. Evaluation Protocol

The evaluation was organized into five phases: initial questionnaire, instructions, calibration, listening tasks, and a final open-ended questionnaire. For each participant, the experiment

[2]https://noise-soundscapes.netlify.app/

lasted for about one hour. The initial questionnaire collects the participants' information (see Section IV-F). In the instruction phase, the participant is invited to seat on a chair in the silent chamber and to wear the bone conduction headphones. Then, the supervisor explains how the experiment works and in particular that the participant has to repeat the sentence in the speech signal as they understand it and indicate with their finger the direction from which the contextual sound comes (left, right, front). While explaining the test, the supervisor reproduces each soundscape, an example of a speech signal, and the contextual sounds.

During calibration, the base speech rate to be used during the test is assessed. For this, the **Suburban** soundscape is used, as it represents a situation with little environmental noise. The supervisor plays an audio stimulus with the preferred speech rate reported by the participant. If the participant can correctly understand the sentence, this speech rate is selected. Otherwise, the process iterates with a speech rate slower by 5%, until the participant can correctly understand the sentence.

The listening task phase consists of a set of 61 tasks. During each task, the supervisor plays an audio stimulus and the participant repeats the sentence and indicates the direction of the contextual sound. The supervisor takes note of the answers provided by the participant. Four initial tasks are used for training, to ensure that the participant correctly understood what to do. The results of these tasks are not recorded. The following 57 tasks include three audio stimuli with the **Suburban** soundscape and 54 with the noisy soundscapes (*i.e.*, **Crowd**, **Traffic**, **Subway**). The three audio stimuli with the **Suburban** soundscape serve as a baseline and use an uncompensated speech signal. Each of them has a contextual sound played from a different direction. For each noisy soundscape there are 18 audio stimuli, 6 for each direction of the contextual sound: one without compensated speech signal, others with a different compensation each. The 57 tasks were organized into 3 sessions, separated by a short break of about 2 minutes. To minimize ordering effects, compensations and directions of contextual sound were applied through a Latin-square design.

Finally, we asked a series of open-ended questions (see Table II), investigating participants' opinions on the presented compensations and the experiment in general. Additionally, we collected their comments and suggestion for improvements. This part of the experiment was organized as a semi-structured interview and, starting from the questions, the supervisor invited the participant to discuss and report comments.

E. Metrics and Data Analysis

We define the soundscape and compensation technique (if any) as independent variables. The dependent variables are:

- **Speech Intelligibility**, defined as the percentage of the correctly understood words for each listening task;
- **Perceived Direction**, defined as the percentage of the correctly identified contextual sound directions.

To account for the small number of participants, common in accessibility research [26], a repeated measure design is adopted [26], [27]. Score differences were analyzed for statistical significance using Friedman Chi Square test [28],

with Dunn post-hoc test for pairwise comparisons [29], corrected for multiple testing using Benjamini-Hochberg FDR method [30]. We also verified, for each scenario and compensation condition, that there was no significant learning effect across the tasks, using the Mann Kendall Trend Test [31], [32], with Benjamini-Hochberg FDR correction [30].

During the evaluation, for each participant, there are exactly three tasks with the same pair of soundscape and compensation. We compute the average for speech intelligibility and perceived direction metrics among these three tasks. Each average value is then used as a data point in the analysis.

F. Participants

We recruited 12 participants with BSLV (3 female) through local associations, social networks, and word-of-mouth. The recruiting criteria required that (a) the participant was not involved in the preliminary studies, (b) the participant is blind, according to the World Health Organization Classification [33], and (c) the participant does not have a hearing impairment. As reported in Table I, participants had an average age of 47 ($SD = 14.63$) and they were all blind since birth with the exception of **P3**, **P4**, and **P11**.

TABLE I
PARTICIPANTS' DEMOGRAPHIC INFORMATION.

				Expertise		Screen reader		Route frequency	
ID	Age	Sex	Onset	music	mobile	speed	voice	Familiar	Unfam.
P1	21	M	birth	3	2	60	F	daily	rarely
P2	40	F	birth	1	4	80	F	daily	daily
P3	59	M	18	1	2	66	F	daily	rarely
P4	64	M	<18	3	4	55	F	daily	rarely
P5	23	M	birth	3	5	90	F	daily	monthly
P6	45	M	birth	3	5	67	F	daily	rarely
P7	40	F	birth	5	4	75	F	every 48h	rarely
P8	35	M	birth	5	4	60	F	daily	weekly
P9	60	M	>18	4	4	60	F	daily	rarely
P10	54	M	birth	4	4	80	M	daily	monthly
P11	61	M	birth	1	5	55	F	daily	rarely
P12	62	F	birth	3	2	60	F	daily	daily

We collected participants' self-reported expertise with mobile devices and music, as a Likert-like scale item ranging from 1 (no expertise) to 5 (high expertise). Average mobile device expertise was 3.73 ($SD = 1.14$), while musical expertise scores averaged 3 ($SD = 1.41$). All participants use *VoiceOver* screen reader on iOS devices, except **P1** who uses *TalkBack* on an Android device. All participants except for **P10** use the screen reader with a female voice. Preferential screen reader speech rates reported by the participants ranged between 55% and 90% with an average of 66% ($SD = 10.57\%$). Following the calibration phase, seven participants (**P1**, **P3**, **P4**, **P6**, **P8**, **P9**, **P11**) used their reported speech rate. Instead, five participants (**P2**, **P5**, **P7**, **P10**, **P12**) used a speech rate that was lower than what they initially reported (70%, 80%, 70%, 70%, 55%, respectively).

All participants travel known routes frequently but only two participant travel unfamiliar routes daily (**P2**, **P12**). Others traverse unknown routes only when needed (**P6**, **P10**, **P11**), and possibly avoid them (**P3**). Except for **P1**, **P3** and **P10**, participants interact with their mobile device through screen reader during mobility on daily basis. However, some (**P1**, **P3**, **P9**, **P10**) avoid listening the screen reader while they need to focus on the environmental sound (*e.g.*, while walking), unless the screen reader feedback is highly relevant (*e.g.*, navigation instructions in an unknown environment).

We also asked the participants how they usually listen to the screen reader while not moving (*e.g.*, home, office) and while in mobility (*e.g.*, walking, waiting for the train at the station). In general, participants use different audio output devices in different situations. While not in mobility, all participants except for **P5** and **P11** listen to the screen reader through the mobile device speaker but some (**P5**, **P7**, **P9**) alternate this with the use of earphones, headphones, and mono earphone. One participant reported using a Bluetooth speaker (**P11**). During mobility, the preferred listening hardware also changes depending on the situation. The mobile device speaker is used by all participants except **P2** and **P8**, but some (**P5**, **P6**) report that it is sometimes necessary to approach the speaker to the ear in noisy environments. **P9** prefers the mobile device speaker as it allows him to listen to the environmental sound. Many participants (**P2**, **P5**, **P7**, **P8**, **P11**) reported using single earphone (either wired or wireless) as it allows them to listen to environmental sound. Many participants (**P3**, **P6**, **P7**, **P9**, **P11**) reported using stereo earphones when they do not need to focus on the environmental sound (*e.g.*, while on a train). One participant (**P9**) uses *Apple AirPods* as they allow to listen to the environmental sound (we assume the participant uses the "transparency mode" available for this device), and one participant (**P11**) reported using bone conduction headphones.

Half of the participants (**P1**, **P5**, **P6**, **P7**, **P10**, **P12**) cope with noisy environments by approaching the mobile device to their ear to listen to the screen reader. Also, half of the participants (**P2**, **P3**, **P4**, **P6**, **P8**, **P10**) reported that they increase the screen reader volume, and one participant (**P9**) adjusts the volume to balance between the volume of the screen reader and the volume of the environmental sound. Three participants (**P4**, **P7**, **P5**) use stereo earphones when they do not need to concentrate on the environmental sound. Three participants report slowing the *VoiceOver* speech rate in noisy environments (**P3**, **P4**, **P7**). In particular **P3** and **P4** use a *VoiceOver* option to read the text word-by-word or character-by-character. Other coping mechanisms are to find a quieter place to listen to the screen reader (**P10**) or to listen during quieter moments (**P11**), for example during subway train stops.

Finally, we asked which assistive technologies participants use in mobility. Most reported using general-purpose navigation tools (*e.g.*, *Google maps*) combined with assistive technologies for people with BSLV. Only **P9** did not use any navigation application. Popular applications are those for checking public transport timetables and for providing accessible directions, like *BlindSquare* [34], *Lazarillo* [35], *NearBy Explorer* [36], *Ariadne GPS* [37], *ViaOpta Nav* [38], *TomTom navigator* [39], *Microsoft Soundscape* [40], and *SeeingAI* [41].

V. EXPERIMENTAL RESULTS

A. Quantitative Analysis

1) Speech Intelligibility: As expected, in the control condition (*i.e.*, **suburban** soundscape) the participants were able
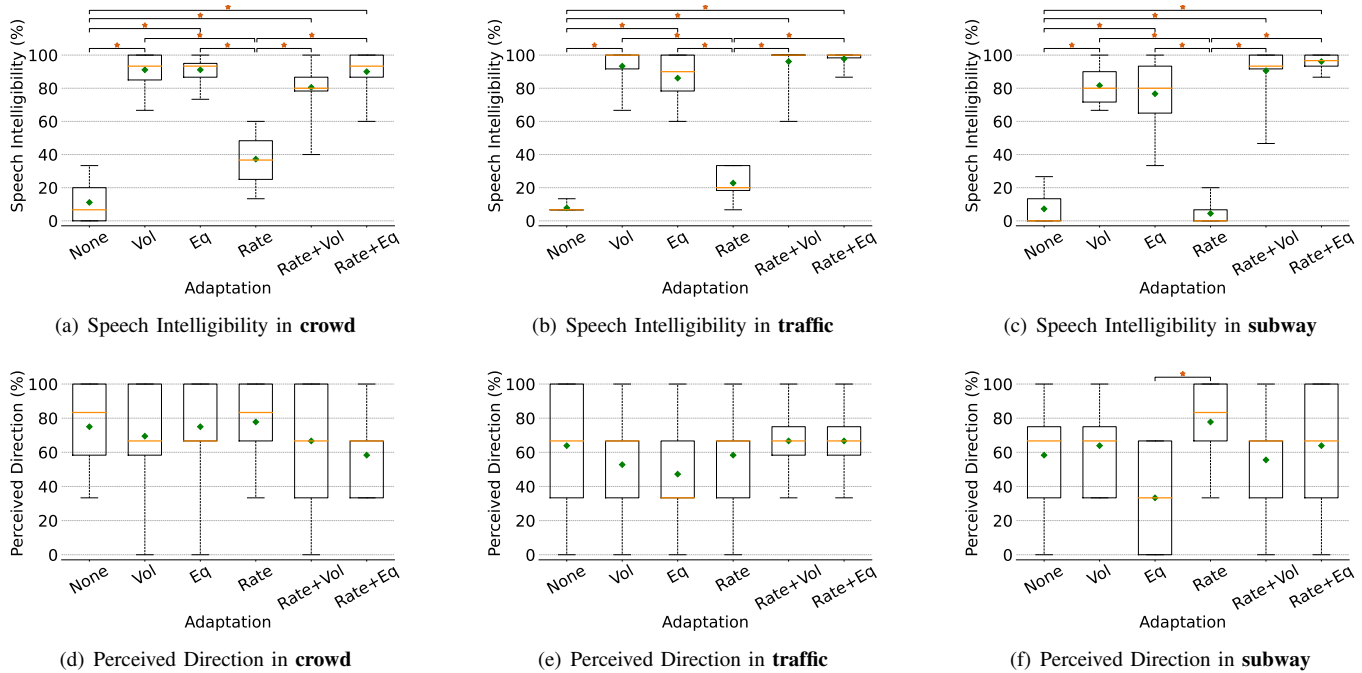
Fig. 2. Speech Intelligibility and Perceived Direction results (◆ mean; ⋆ significant).

to recognize nearly all the words ($M = 99.44\%$, $SD = 1.84\%$). Instead, in presence of noise (see Fig. 2), without any compensation, the percentage of correctly recognized words drops to $11.11 \pm 12.57\%$, $7.78 \pm 2.48\%$ and $7.22 \pm 9.98\%$, for **Crowd**, **Traffic** and **Subway** soundscapes, respectively. This result confirms that, without any compensation, synthetic speech intelligibility is severely impaired in presence of typical environment noise.

All the proposed compensation techniques increase the speech intelligibility average score in all soundscapes. The only exception is the **Rate** compensation in the **Subway** soundscape, for which the intelligibility is lower than without compensation. Specifically, for all the soundscapes, the effect of the sonification technique was found to be significant (**Crowd** $H_{(5)} = 47.17$, $p < .001$; **Traffic** $H_{(5)} = 52.98$, $p < .001$; **Subway** $H_{(5)} = 50.49$, $p < .001$). Pairwise comparisons reveal that **Vol**, **Eq**, **Rate+Vol** and **Rate+Eq** significantly improve the speech intelligibility with respect to both **None** and **Rate**, considering all soundscapes. In particular, **Rate+Eq** achieves the highest intelligibility score for the **Traffic** ($97.78 \pm 4.16\%$) and **Subway** ($96.11 \pm 4.27\%$) soundscapes. Instead, in the **Crowd** soundscape, **Vol** ($91.11 \pm 10.30\%$) and **Eq** ($91.11 \pm 7.37\%$) show the highest improvement over the uncompensated speech.

2) Perceived Direction: In the **Suburban** scenario participants can recognize the correct direction of the contextual sound in almost all cases ($M = 92\%$, $SD = 14\%$). In the other soundscapes, recognizing the correct direction of the contextual sound is much harder. Indeed, the perceived direction score falls to $75 \pm 28\%$ in the **Crowd** soundscape, $64 \pm 32\%$ for **Traffic** and $58 \pm 31\%$ for the **Subway** scenario (see Fig. 2(f)). Contrary to our expectations, the screen reader speech compensations do not seem to worsen the perceived

direction score over the uncompensated condition. Indeed, no significant differences emerge among compensations. The only significant group difference ($H_{(5)} = 15.27$, $p < .001$) is detected in the **Subway** soundscape, where pairwise comparisons reveal that **Rate** is significantly better than **Eq** compensation.

B. Answers to the Final Open-Ended Questions

We summarize the participants' answers to the final questions, which are reported in Table II. All participants agreed that the soundscapes realistically reproduce real-world situations in which it could be hard to listen to the screen reader (Q1). Only **P5** stated that he would not have problems in noisy scenarios, as he would set the volume to the maximum. However, he also admitted that he would have problems in the **Subway**, scenario, as it is particularly noisy.

Besides **P3**, who had more difficulties in distinguishing sounds in the **Traffic** scenario, all other participants considered **Subway** as the soundscape in which it was harder to understand the screen reader (Q2). They noted that this is the noisiest soundscape, it masks the speech output and hence it is more cognitively demanding. Nine participants (**P1**, **P2**, **P3**, **P5**, **P6**, **P7**, **P8**, **P10**, **P12**) considered **Suburban** as the simplest soundscape, as it is less noisy and the volume of the screen reader is higher and separable from the soundscape. Instead, two participants (**P4**, **P11**) considered **Crowd** as the simplest soundscape. In particular, **P4** noted that sounds and voices are more distinct. Finally, **P9** reported that **Traffic** is the simplest soundscape because environmental noise and speech were clearly distinguished in this soundscape.

Considering the contextual sounds (Q3), six participants (**P3**, **P5**, **P6**, **P8**, **P9**, **P10**) regarded them as realistic. However, **P2** considered the **Subway** contextual sound too short and **P11** believed it was not loud enough. **P7** reported that the

TABLE II
FINAL OPEN-ENDED QUESTIONS

Q1	Do the soundscapes realistically reproduce situations in which it can be hard to listen to the screen reader?
Q2	In which soundscape was it harder to understand the screen reader? Which one was easiest?
Q3	Were contextual sounds [the term was introduced to the participant before] realistic? Which contextual sound was harder to perceive?
Q4	Some sentences were read aloud with a slower rate. Did you perceive this?
Q5	Some sentences were read aloud with a higher volume. Did you perceive this?
Q6	Some sentences were read aloud with a higher volume on some frequencies only. Did you perceive this?
Q7	Some sentences were read aloud with a combination of slower rate and higher volume. Did you perceive this?
Q8	What do you think of the bone-conduction headphones that you used during the experiments?

Suburban contextual sound was hard to hear and **P12** stated that it was different than expected. **P2** perceived the sound of cash receipt in **Crowd** as too loud, while **P7** and **P12** considered it unrealistic. Finally, the contextual sound in **Traffic** was perceived to be different than expected by three participants (**P3**, **P5**, **P7**) and hard to separate from the soundscape (**P11**).

All participants were able to perceive the **Vol** compensation and noted that it improves intelligibility (Q5). Eight participants (**P1**, **P3**, **P4**, **P7**, **P9**, **P10**, **P11**, **P12**) were able to perceive the presence of the **Rate** compensation (Q4), and all of them regarded it as useful. Only five participants (**P3**, **P4**, **P7**, **P9**, **P11**) perceived **Eq** (Q6). Since **Vol** and **Eq** are similar, we suspect that most participants were not able to distinguish between the two. Similarly, only five participants (**P4**, **P5**, **P7**, **P9**, **P10**) perceived the combined compensations (Q7).

Considering the use of the bone conduction headphones (Q8), 9 participants (**P3**, **P4**, **P6**, **P7**, **P8**, **P9**, **P10**, **P11**, **P12**) agreed that this is an acceptable solution, mainly because they do not prevent the hearing of environmental sound. However, some participants (**P1**, **P3**) reported that they are uncomfortable and heavy. In particular, **P2** noted that the sound is "external" and hence hard to understand.

C. Participants' Comments

Participants also provided spontaneous comments. We report the key topics that emerged from their analysis.

1) Approach Usefulness: Participants found the proposed approach useful for screen reader usage in noisy environments (**P1**, **P4**, **P7**, **P8**). In particular, **P7**, who already used manual adjustment of volume and speech rate, requested for it to be implemented in the same way as the automated adaptation of screen illumination to the ambient light on mobile devices[3]:

> "It would be great if they [compensations] were made similar to the automatic screen illumination."

2) Combined Compensations: Participants considered the combined compensations useful (**P5**, **P7**, **P9**, **P12**), confirming the findings from the preliminary studies. **P12** noted that **Rate** alone does not improve intelligibility, but it can be useful when paired with **Vol**, which is consistent with quantitative results. Similarly, **P7** noted that raising the volume and lowering the speech rate helps in noisy contexts. In particular, **P5** noted:

> "I sometimes wondered how I was able to understand sentences that I did not previously understand. These [combined compensations] make life easier."

[3]All quotes have been translated from Italian.

3) Trade-off between intelligibility and distraction: Participants also commented on the inherent trade-off between the ability to correctly understand the screen reader speech and to pay attention to the environment (**P1**, **P2**, **P3**, **P4**, **P6**, **P9**, **P10**, **P11**). While most of the comments related to the soundscape and speech volume, some referred to the speech rate (**P10**, **P11**). **P11** in particular felt that slowing the speech rate lowers the volume of some words, while **P10** observed:

> "It is not always useful because slowing too much dilates the time needed [to listen]."

4) Modifications to the Approach: Participants also suggested possible improvements. **P10** mentioned that some screen reader voices mask specific consonant sounds, which makes them harder to understand, and therefore we should consider also the screen reader voice used. **P5** also suggests another possible compensation, which is to delay the screen reader speech when strong background noise is detected:

> "I would've liked to hear the message when the [subway] sound fades out. Instead, the message was played when the sound was on the rise."

VI. DISCUSSION

A. Main Results

The proposed compensation techniques were found to be effective in improving synthetic speech intelligibility across different noisy soundscapes. Specifically, four of the five proposed compensation techniques significantly improved the speech intelligibility in all three noisy soundscapes. The only compensation that did not significantly improve intelligibility is **Rate**. Indeed, some of the participants initially felt that reducing speech rate would not improve intelligibility in presence of noise. For example, one participant in the instruction phase of the preliminary evaluation argued:

> "I don't think that diminishing the speed would impact the speech understanding."

However, after the listening tasks, most participants confirmed that combining rate with the other compensations was useful for louder soundscapes. For example, the same participant retracted the prior opinion:

> "Both raising volume and reducing speed are useful. In general, I think lower speed is more important"

In agreement with this opinion, results show that in two scenarios (**Traffic** and **Subway**), the combined compensations improve on average over the other compensations. While this difference is not significant, possibly due to the limited sample size, it confirms the intuition of the participants to the

preliminary study that combining the compensation techniques can further improve speech intelligibility in noisy scenarios.

We initially hypothesized that compensations would distract the participants from the environment soundscape. However, no significant difference was observed in the **Perceived Direction** score across different conditions. Such a result can be explained by a comment provided during preliminary experiments: when the speech is hard to understand (either too quiet or too fast) it requires additional concentration, hence distracting from the environmental sound. This means that in some cases compensations might actually *reduce* distraction.

B. Experimental Design Limitations

In order to ensure both participants' safety and experimental repeatability, we conducted the evaluation in a controlled environment. This has some implications, including the fact that the soundscapes had to be simulated. To ensure highly realistic simulations, in our experimental settings we used real-world recordings of the target soundscape scenarios, replayed in a silent chamber with a quadraphonic speaker setup and settings replicating real-world listening conditions. Indeed, all study participants agreed that the experimental setting was realistic.

Another consequence is that the participants were in a safe environment, rather than in the real-world, so they did not actually need to focus on the soundscape to prevent dangers. This may have impacted their level of attention with respect to the background sounds. Additionally, distraction was measured as the ability to pinpoint the direction of a single sound, while in general continuous attention should be devoted to the ambient soundscape. To account for this, the contextual sound was played at a random time, and therefore the participants were stimulated to pay attention to the soundscape continuously.

In order to ensure uniformity among the tests we always used the same three soundscapes, reproducing the speech signal always at the same time. While this guarantees that all sentences are read with consistent background noise, participants reported that they got used to the soundscape and were less distracted by it after a while. The experimental design could be improved by using longer soundscapes and by changing the starting point of the soundscape playback across repetitions. However, this would require collecting a larger set of data-points, to compensate for the differences in the listening conditions.

One more consequence of placing sentences in the middle of the soundscapes is that, after the sentence reading ends, there are still a few seconds before the soundscape ends. The participants would wait until the end of the soundscape before repeating what they heard. This required an additional memorization effort and in some cases resulted in participants forgetting what they heard. For this reason, the participants who reported this issue were asked to repeat the sentence immediately, without waiting for the soundscape to end.

To measure speech intelligibility without an influence of an external semantic context we used a random sentence generation approach [7]. One limitation of this approach is that, while the exact words are unpredictable, the phrase structure is always the same (*e.g.*, the first word is always a name). Thus the participants could try to infer a word even if it is not clearly heard. However, this is not unrealistic: in many real-world applications, the screen reader often provides structurally similar messages with a small set of possible words (*e.g.*, "Turn right/left" in a navigation system).

Finally, our experiments involved solely Italian-speaking participants. While we expect the results to generalize to other languages, such an assumption would need to be evaluated with participants speaking different languages. Similarly, while the technique should be applicable to different listening hardware such as the speaker or in-ear headphones, we conducted the experiments solely with bone-conduction headphones. Thus, additional experiments would be needed to verify that the approach generalizes to other listening hardware.

C. Technique Limitations

In designing the **Rate** compensation technique, we decided to apply a flat speed reduction. A different solution could consist in reducing the speech rate proportionally with the noise level. However, for this, we would need to study the correlation between speech intelligibility at different speech rates and environmental noise. This would require a process of parameter tuning that is out of the scope of this paper.

Correlating the environmental noise level with the amount of compensation is also a problem for **Vol** and **Eq**. However, for these compensations, it was possible to tune the settings based on prior works studying the impact of the difference between speech and noise volume on speech intelligibility [42], [43]. Still, fine-tuning these parameters to the specific application domain could help to achieve even better results.

While the speech signals were pre-recorded for convenience, the proposed compensations can be computed in real-time. Indeed, assuming that the compensation techniques are implemented as a part of the speech synthesizer, they would require monitoring the environmental noise and changing the speech generation parameters. In order to rapidly adapt to the environmental noise, the proposed technique employs temporal windows of $46\,\mathrm{ms}$. Since changing the speech generation parameters does not produce any additional delay, the overall delay remains within $50\,\mathrm{ms}$, which can be considered real-time. In order to compute the compensations in real-time, the device microphone could be used to collect samples of the environmental noise. If the microphone is expected to be covered (*e.g.*, if the device is in the user's pocket), an external microphone could be used (*e.g.*, the headset microphone).

VII. Conclusions and Future Work

Considering the importance of mobile devices for the orientation and mobility of people with BSLV, hard-to-understand instructions can result in potentially hazardous situations. This paper proposes compensation techniques to mitigate this problem and provides experimental evidence that they can effectively improve screen reader intelligibility in noisy environments, without negative impacts on the distraction from the soundscape. Consequently, the proposed compensations are a practical solution to this problem. They can also be

easily implemented in existing mobile screen readers without proprietary hardware requirements.

This work paves the way for various research directions. First, it is possible to investigate other compensation techniques, by altering different speech properties (*e.g.*, pitch). Second, we intend to explore how different environmental noise characteristics affect the compensation techniques; an important element of this research would be to find the correlations between environmental noise characteristics and speech intelligibility at different levels of compensation (*e.g.*, speech rates). Additional factors should also be taken into account, including the language and the listening hardware (headphones, speakers). A third research direction is to investigate compensation techniques for sonification instructions, which have been proposed for navigation assistance [44].

ACKNOWLEDGMENTS

REFERENCES

[1] WebAIM, "Screen reader user survey," 2021. [Online]. Available: https://webaim.org/projects/screenreadersurvey9/#disabilitytypes

[2] S. K. Kane, J. P. Bigham, and J. O. Wobbrock, "Slide rule: making mobile touch screens accessible to blind people using multi-touch interaction techniques," in *Conf. on comp. and access.* ACM, 2008.

[3] H. Ye, M. Malu, U. Oh, and L. Findlater, "Current and future mobile and wearable device use by people with visual impairments," in *Conference on Human Factors in Computing Systems.* ACM, 2014.

[4] A. Abdolrahmani, R. Kuber, and A. Hurst, "An empirical investigation of the situationally-induced impairments experienced by blind mobile device users," in *International Web for All Conference.* ACM, 2016.

[5] K. Shinohara and J. O. Wobbrock, "In the shadow of misperception: assistive technology use and social interactions," in *SIGCHI conference on human factors in computing systems.* ACM, 2011.

[6] D. Bragg, C. Bennett, K. Reinecke, and R. Ladner, "A large inclusive study of human listening rates," in *Conference on Human Factors in Computing Systems.* ACM, 2018.

[7] G. E. Puglisi, A. Warzybok, S. Hochmuth, C. Visentin, A. Astolfi, N. Prodi, and B. Kollmeier, "An italian matrix sentence test for the evaluation of speech intelligibility in noise," *Int. J. of Audiology*, 2015.

[8] S. K. Kane, C. Jayant, J. O. Wobbrock, and R. E. Ladner, "Freedom to roam: a study of mobile device adoption and accessibility for people with visual and motor disabilities," in *Conference on Computers and accessibility.* ACM, 2009.

[9] D. Ahmetovic, D. Sato, U. Oh, T. Ishihara, K. Kitani, and C. Asakawa, "Recog: Supporting blind people in recognizing personal objects," in *Conference on Human Factors in Computing Systems.* ACM, 2020.

[10] S. Azenkot and N. B. Lee, "Exploring the use of speech input by blind people on mobile devices," in *International conference on computers and accessibility.* ACM, 2013.

[11] H. Kacorri, S. Mascetti, A. Gerino, D. Ahmetovic, V. Alampi, H. Takagi, and C. Asakawa, "Insights on assistive orientation and mobility of people with visual impairment based on large-scale longitudinal data," *Transactions on Accessible Computing*, 2018.

[12] D. Sato, U. Oh, J. a. Guerreiro, D. Ahmetovic, K. Naito, H. Takagi, K. M. Kitani, and C. Asakawa, "Navcog3: Large-scale blind indoor navigation assistant with semantic features in the wild," *Transactions on Accessible Computing*, 2019.

[13] G. Presti, D. Ahmetovic, M. Ducci, C. Bernareggi, L. A. Ludovico, A. Baratè, F. Avanzini, and S. Mascetti, "Watchout: Obstacle sonification for people with visual impairment or blindness," in *Conference on Computers and Accessibility.* ACM, 2019.

[14] N. Martiniello, W. Eisenbarth, C. Lehane, A. Johnson, and W. Wittich, "Exploring the use of smartphones and tablets among people with visual impairments: Are mainstream devices replacing the use of traditional visual aids?" *Assistive Technology*, 2019.

[15] H. A. Faucett, K. E. Ringland, A. L. Cullen, and G. R. Hayes, "(in) visibility in disability and assistive technology," *Transactions on Accessible Computing*, 2017.

[16] R. Kuber, A. Hastings, M. Tretter, and D. Fitzpatrick, "Determining the accessibility of mobile screen readers for blind users," in *IASTED Conference on Human-Computer Interaction.* ACTA, 2012.

[17] R. Wang, C. Yu, X.-D. Yang, W. He, and Y. Shi, "Eartouch: facilitating smartphone use for visually impaired people in mobile and public scenarios," in *Conf. on human factors in comp. systems.* ACM, 2019.

[18] R. J. P. Damaceno, J. C. Braga, and J. P. Mena-Chalco, "Mobile device accessibility for the visually impaired: problems mapping and recommendations," *Universal Access in the Information Society*, 2018.

[19] B. R. Molesworth, M. Burgess, and D. Kwon, "The use of noise cancelling headphones to improve concurrent task performance in a noisy environment," *Applied Acoustics*, vol. 74, no. 1, 2013.

[20] V. Ambeth Kumar, S. Malathi, A. Kumar, K. C. Veluvolu *et al.*, "Active volume control in smart phones based on user activity and ambient noise," *Sensors*, vol. 20, no. 15, p. 4117, 2020.

[21] A. Mason, N. Jillings, Z. Ma, J. D. Reiss, and F. Melchior, "Adaptive audio reproduction using personalized compression," in *Proc. 57th Audio Engineering Society Conference: 57th International Conference*, 2015.

[22] D. Lee, J. D. Lewis, P. M. Johnstone, and P. N. Plyler, "Acceptable noise levels and preferred signal-to-noise ratios for speech and music," *Ear and Hearing*, vol. 43, no. 3, pp. 1013–1022, 2022.

[23] H. Fletcher and W. A. Munson, "Loudness, its definition, measurement and calculation," *Bell System Technical Journal*, 1933.

[24] B. C. Moore and B. R. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," *The journal of the acoustical society of America*, vol. 74, no. 3, pp. 750–753, 1983.

[25] G. Galimberti, "Auditory feedback to compensate audible instructions to support people with visual impairment," in *The 23rd International ACM SIGACCESS Conference on Computers and Accessibility*, 2021.

[26] A. Sears and V. Hanson, "Representing users in accessibility research," in *Conference on Human factors in computing systems.* ACM, 2011.

[27] 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*, 2019.

[28] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *Journal of the american statistical association*, vol. 32, no. 200, pp. 675–701, 1937.

[29] O. J. Dunn, "Multiple comparisons among means," *Journal of the American statistical association*, 1961.

[30] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal statistical society: series B (Methodological)*, 1995.

[31] H. B. Mann, "Nonparametric tests against trend," *Econometrica: Journal of the econometric society*, pp. 245–259, 1945.

[32] M. G. Kendall, *Rank correlation methods.* Griffin, 1948.

[33] B. Thylefors, A. Negrel, R. Pararajasegaram, and K. Dadzie, "Global data on blindness." *Bull of the WHO*, vol. 73, no. 1, 1995.

[34] "Blindsquare," 2021. [Online]. Available: https://www.blindsquare.com/

[35] "Lazarillo," 2022. [Online]. Available: https://lazarillo.app

[36] "Nearby explorer, american printing house for the blind, inc., 2017. [online]." 2022. [Online]. Available: https://tech.aph.org/ne_info.htm

[37] "Ariadne gps," 2021. [Online]. Available: https://www.ariadnegps.eu/

[38] "Viaopta nav," 2021. [Online]. Available: https://www.itcares.it/portfolio/viaoptanav/

[39] "Tomtom navigator app," 2022. [Online]. Available: https://www.tomtom.com/it_it/navigation/mobile-apps/go-navigation-app

[40] "Microsoft soundscape." [Online]. Available: https://www.microsoft.com/en-us/research/product/soundscape/

[41] "Seeingai," 2021. [Online]. Available: https://www.microsoft.com/en-us/ai/seeing-ai

[42] D. McShefferty, W. M. Whitmer, and M. A. Akeroyd, "The just-noticeable difference in speech-to-noise ratio," *Trends in hearing*, vol. 19, 2015.

[43] A. Weisser and J. M. Buchholz, "Conversational speech levels and signal-to-noise ratios in realistic acoustic conditions," *The Journal of the Acoustical Society of America*, vol. 145, no. 1, pp. 349–360, 2019.

[44] S. Mascetti, L. Picinali, A. Gerino, D. Ahmetovic, and C. Bernareggi, "Sonification of guidance data during road crossing for people with visual impairments or blindness," *Int. J. of Human-Computer Studies*, 2016.