Moving Towards and Reaching a 3-D Target by Embodied Guidance: Parsimonious Vs Explicit Sound Metaphors

Coline Fons, Sylvain Huet, Denis Pellerin, Silvain Gerber, Christian Graff

▶ **To cite this version:**

Coline Fons, Sylvain Huet, Denis Pellerin, Silvain Gerber, Christian Graff. Moving Towards and Reaching a 3-D Target by Embodied Guidance: Parsimonious Vs Explicit Sound Metaphors. HCII 2023 - 25th International Conference on. Human-Computer Interaction HCII 2023, Jul 2023, Copenhague, Denmark. pp.229-243, 10.1007/978-3-031-35681-0_15 . hal-04192144

HAL Id: hal-04192144

https://hal.univ-grenoble-alpes.fr/hal-04192144

Submitted on 31 Aug 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Copyright

Moving towards and reaching a 3-D target by embodied guidance: parsimonious vs explicit sound metaphors

Coline Fons[1,2], Sylvain Huet[2], Denis Pellerin[2], Silvain Gerber[2], and Christian Graff[1]

[1] Univ. Grenoble Alpes, Univ. Savoie Mont-Blanc, CNRS, LPNC, 38000 Grenoble France
firstname.lastname@univ-grenoble-alpes.fr

[2] Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France
firstname.lastname@gipsa-lab.grenoble-inp.fr

Abstract. Sensory Substitution Devices (SSDs) assist Visually Impaired People (VIP) by providing information usually acquired by vision through another functional sensory modality. Here, we evaluate a SSD that aims to guide visually impaired people to a target by converting spatial information into sound. A balance must be struck between parsimonious and explicit guidance, so that it is precise but not imposing an excessive cognitive load. Here, we express the deviation from the target to the participant's hand. We compared three sound metaphors: 1) Binary (BI) gives only the right direction by white noise, 2) Angle-to-Pitch (AP) converts the 3-D angular deviation between a pointed direction and the right direction into sound pitch according to a continuous function, 3) Dissociated Vertical-Horizontal (DVH) dissociates the 3-D deviation into two dimensions, using pitch to provide the angular deviation projected on the horizontal plane (a continuous metaphor) and a superimposed white noise to indicate the right height (a binary metaphor). We conducted "hot and cold game" type tests in a 3-D environment. Participants, with eyes closed, moved to search and reach virtual spheres with their index finger. Each metaphor was evaluated by the time taken to reach the target, and by a questionnaire. The results show the advantage of AP over BI: it allowed a faster target hit and was considered easier to use, more efficient, and more comfortable than BI. They also show an advantage of the DVH metaphor over AP: the times to target hit were shorter, and it was considered easier to use than AP. The DVH metaphor is a good trade-off between parsimonious (use of binary information) and explicit (dissociation of the two axes) guidance. It is a first step towards smartphone SSDs applications to help find objects in a situation of visual impairment.

Keywords: Sensory Substitution · Blindness · 3-D Target Reaching · Sound Metaphor · Virtual Prototyping · Spatial Cognition · Hand Guidance

1 Introduction

Visually Impaired People (VIP) interact with their spatial environment through their other senses, such as hearing or touch. But some tasks, including navigation or finding and reaching an object, are challenging without assistance. There is a need to develop accessible assistive technologies for VIP. Various Sensory Substitution Devices (SSDs) have been created to address this need. They provide through another functional sensory modality information captured by an artificial sensor. SSDs can be implemented within navigation applications for smartphones using GPS or synthesizing texts vocally. Navigation tasks are the main focus of most SSDs and often include obstacle avoidance [1, 8, 9, 15, 20]. Here we focus on helping people reach a targeted object with their hand without vision, a task less explored in the literature. The target can be, for example, a door handle, a doorbell button, or an object on a shelf. We aim to create the most efficient and ergonomic sound guidance in 3-D space. Most guidance systems convert the deviation from the target to the manipulated sensor in either sound [4, 9, 10, 26] or vibration [19, 28]. Below, we present arguments for the choice of an egocentric frame of reference, and we will then examine how spatial information can be conveyed through sound guidance.

1.1 Frame of reference

When designing a Sensory Substitution Device (SSD) for guidance, there is a choice of several reference frames in which the target is positioned. The frame of reference for the choice of spatial metrics to be encoded cannot ignore the body-specific logic of the motor and cognitive systems. For example, providing the whole pixel matrix of a depth scene [14] is useless for reaching a single point, and this data's overflow is difficult to interpret. Alternatively, providing only the absolute position of the target also leads to interpretation errors [2], even if the amount of data is reduced to the essential.

Most of the time, the spatial data to be transmitted to the VIPs is collected by one or more cameras. The captured images are used to extract the deviations from the target. The camera's position determines the frame of reference in which the spatial data is captured. Thus, some devices have a camera placed on the participant's head [5, 9, 22], while others position the camera on the hand or forearm of the participant [4, 11, 19]. Some combine the two approaches by using two cameras, one on the head and one on the hand [13, 28], and use one or the other depending on the participant's progression through the task (distal guidance to navigate to the target, then proximal guidance to reach the target with the hand).

When the frame of reference is a camera, it is both the sensor (which collects the data) and the pointer (reference frame in which the deviations from the target are calculated). It is an allocentric frame of reference: the spatial data are encoded in relation to an object, the camera. In contrast, an egocentric frame of reference allows data to be encoded in reference to a coordinate system that places the user's body as the origin and defines the axes of the coordinates in

relation to the user's body orientation [18]. For VIPs, an egocentric frame of reference is more effective [18]: due to the visual feedback deficit, it is indeed difficult for VIPs to match their smartphone's coordinates to their egocentric coordinate system.

An egocentric frame of reference would therefore seem more suitable for a SSD guiding a visually deprived person to a target. The best frame of reference could be the finger because the hand is central in an object reaching task. Here we express the deviations with respect to the hand that has to reach the object and not with respect to the sensor that provides the raw spatial data.

1.2 Sonic guidance

Among the spatial information that can be sent to the user, the challenge is to provide the most useful and simple elements. Two categories of sounds can be used to convey spatial information: verbal description or sonification. Verbal description in a guidance task may correspond to instructions such as "turn right", "turn left", etc [4, 5, 11, 19, 22, 23, 28]. However, verbal instructions are considered slow and cumbersome [13] since they disrupt the perception/action loop. Indeed, if the user moves the sensor too quickly, at the end of the verbal instruction, it may no longer be valid. Sonification, on the other hand, is the use of nonverbal sounds to convey information or perceptual data [16]. There are two kinds of sonification [16]. Spatial sonification consists in using the natural capacities of the auditory system to locate the position of a target by virtually rendering its position. This is a technique used in [5, 9, 10, 13] to localize the target on the horizontal plane. Some [5, 9–11, 13, 23] also rely on non-spatial sonification, which uses the physical characteristics of sounds, such as pitch, intensity, tempo, brightness, etc., to convey guidance information. The link between the input visual data and the output sounds is therefore metaphorical.

In their research on the process of sonification design for guidance tasks, Parseihian et al. [17] found an improvement in distance perception with non-spatial sonification compared to spatial sonification. They also advise to adapt the sonification strategy to the goal of the guidance task. Indeed, they found that participants in a target reaching task were more accurate with "strategies with reference" (adding a sound reference corresponding to the target) than with "basic strategies" (based on varying the basic perceptual attributes of the sound), whereas they were slower with "strategies with reference."

In their work on sonification in three-dimensional space, Ziemer et al. [24–27] identified several requisites for creating an accurate guidance system. One of them is to integrate the three spatial dimensions into a single auditory stream; another one is to allow them to be perceived orthogonally. Perceptual orthogonality means that if two information variables are sonified simultaneously, both can be interpreted, and if one variable changes, the change in sound can be attributed to its corresponding continuum and unambiguously interpreted. It should be taken into account that physically independent sound parameters are not necessarily orthogonal in perception (e.g., pitch and intensity). They also

recommend using continuous dimensions and having a high perceptual resolution (which can be measured by the just-noticeable difference).

While considering these criteria, the user's cognitive load must also be taken into account. The criteria cited by Ziemer et al. [24–27] allow for the creation of a very precise sound guidance, but a balance must be struck between precision and ease of use for the user. Indeed, Ziemer et al. [24–27], taking these criteria into account, created a sonification with a single auditory stream but five psychoacoustic quantities varying continuously along the three dimensions. Having a single stream allows one to hear all the information at once, which allows precise guidance but could also be overwhelming and produce cognitive overload.

To obtain an efficient 3-D guidance but also to decrease the cognitive load and have a comfortable device, choices between parsimonious and explicit information have to be weighed:

- Encoding the three spatial dimensions explicitly (which requires to use several auditory metrics) OR reducing the number of encoded dimension to minimize the amount of information to be processed. In the latter case, the user must infer the implicit information from the information provided.
- Using a single auditory stream for the three dimensions to have all the information at once (no need to switch attention) OR using several separate streams to reduce the amount of information on each stream. In the latter case, the user needs to switch attention between streams.
- Using continuous OR using discrete scales.

Taking into account these different choices and the literature on the subject, we created three sound metaphors, which we compared in pairs:

- Comparison 1: Angle-to-Pitch (AP) VS Binary (BI). BI gives only binary information: sound when pointing in the right direction VS no sound. Studies have shown that simple binary sound cues are enough to create a guidance system [12]. AP gives a directional deviation in angle from the target, expressed by the pitch of the sound, in continuous scales. A single sound parameter is used to express a single global 3-D spatial metric. Each dimension can be inferred by the action-perception loop. AP has several qualities highlighted by Ziemer et al. [24–27]: dimensions integrated into a single stream, high perceptual resolution (pitch), and continuous scales. But the dimensions are not orthogonal: a single sound parameter is used to encode direction on both vertical and horizontal axes. It could imply more efforts to interpret the sonification, and thus a higher cognitive load. To avoid this, the directional deviations on the two axes can be separated into two parameters on a single sound stream, as suggested by Ziemer et al. [24–27]. But integrating several sound parameters on the same stream can lead to interpretation errors and cognitive overload.
- Comparison 2: AP VS Dissociated Vertical-Horizontal (DVH). With DVH, the deviation on the horizontal axis are coded on a continuous scale, by the

pitch of the sound. The deviation on the vertical axis is coded binary by the presence of a white noise of good height. Thus the two dimensions are orthogonal, the user can interpret them separately, which should simplify the sonification and make the guidance more accurate. The use of binary information should further facilitate the sonification and decrease the amount of data to integrate and therefore the cognitive load.

To compare these three metaphors we conducted "hot and cold game" type tests in a 3-D environment. Participants, with eyes closed, moved to search and reach virtual spheres with their index finger. Each metaphor was evaluated by the time taken to reach the target and by a questionnaire.

In the following, we will present the material and method used to conduct this experiment, along with the results, which we will discuss.

2 Methods

2.1 Participants

We conducted two comparisons. For Comparison 1, fourteen sighted participants and one VIP performed the target-reaching task with the first two metaphors: BI and AP. For Comparison 2, eight other sighted participants and one other VIP participant compared the AP metaphor and the DVH metaphor. Most participants were students participating for course credit, others were volunteers (other students, friends, and relatives). All gave informed consent before participating in the study.

2.2 Engineering

Tests take the form of a virtual game consisting in reaching spheres in a 3-D space with one's index finger (see Fig. 1). A Qualisys optical motion capture system locates reflectors fixed on the participant's body. Coordinates relative to the finger and elbow positions are transmitted from the acquisition computer running Qualisys Track Manager (QTM) to the pilot computer through the Virtual Reality Peripheral Network (VRPN) protocol. In the pilot computer, our C++ control software: 1) immerses the participant into the virtual environment together with the target, using the OpenScene Graph (OSG) 3-D toolkit; 2) computes spatial metrics used in sound conversions; 3) transmits them to the PureData sound system, running on the same computer, which synthesizes the sounds accordingly and sends them to the participant; 4) drives the experimental protocol.

The target is a sphere of 30 cm diameter that can be positioned at 27 different locations on a 3*3*3 grid of 82 cm steps in x, 1.35 cm in y and at three different heights in z: 70 cm; 110 cm; 150 cm. A test block consists of a series of 6 consecutive targets to reach, positioned semi-randomly to ensure that the participant always travels a minimum distance between targets. The sequence ensures that the three x; y; z coordinates of the spheres vary between each trial.

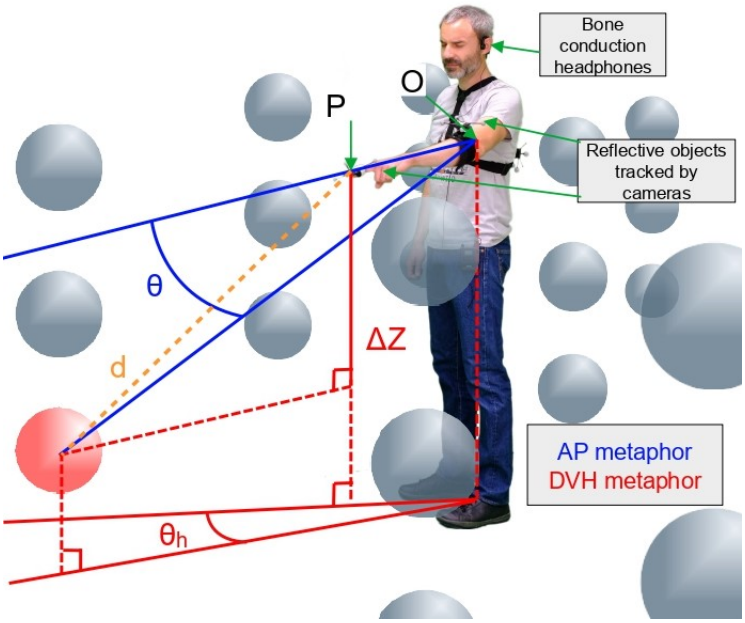

Fig. 1: Photomontage showing the reflectors used to capture the position of the participant's limbs, the virtual spheres' position (the red sphere being the target), and BI, AP and DVH metaphors' metrics. The target sphere is located at one out of 27 possible locations. The points O (elbow) and P (index finger tip) determine the direction pointed by the participant. With BI, a white noise is triggered when the line OP intersects the target (θ null). With AP, the sound's pitch additionally varies according to θ not null. With DVH, the sound's pitch varies according to θ_h, and a white noise is triggered when P is at the same height as the target.

2.3 Participant's Equipment

During the experiment, the participant was equipped with (see Fig. 1):

- Objects with reflectors mounted on them, located by the Qualisys system. They are positioned on anatomical segments corresponding to points O and P located in the frame of reference. A wrap around the finger places point P at the tip of the index finger. An armband places the point O at the elbow. The $\overrightarrow{OP}$ vector defines the direction of the participant's pointing.
- A bone conduction headset (Aftershokz Sportz3) connected to a receiver box (Mipro MI909R) to receive the sound feedback.
- A miniature keyboard (iclever Rii 2.4 GHz L * l * h = 151 * 59 * 12.5 mm) to start the next trial once the target has been reached.

2.4 Spatial metrics

Several features are extracted from the scene to be transcoded into sounds (see Fig. 1). To compare different sound metaphors, we need to compute different geometric quantities.

For the BI metaphor and the AP metaphor, the extracted geometric features are as follows:

- The angle $\theta = \widehat{POT}$ which corresponds to the angle formed by the lines OP and OT (angular deviation).
- The distance d = [PT] (distance deviation).

For the DVH metaphor, we used two additional features:

- The angle $\theta_h = \widehat{POT}_h$, which corresponds to the projection of the angle θ on the horizontal plane parallel to the ground (angular deviation)
- The height difference ΔZ = $|z_P$ - $z_T|$, which corresponds to the projection of the distance [PT] on the vertical axis Z (distance deviation).

These features allow us to dissociate the deviations on the horizontal and vertical axes for the DVH metaphor.

O and P define the direction of the participant's pointing and their distance to the target. For the AP metaphor, we used a polar coordinate system. However, in classical polar coordinates, the angle between the pointed direction and target direction is considered from the same origin O. It is the case in most studies in which the camera is used to estimate both the angle and the distance to the target [4, 11, 19, 23]. Here, we dissociated O and P. O is used to estimate the orientation of the angle θ, while P is used to estimate the distance to the target (see Fig. 2). We could have used O to estimate the distance, but in this target-reaching task with the hand, the distance is intuitively considered null when the finger touches the target.

2.5 Sound metaphors

We transcoded the deviations in angle and distance from the target into sound parameters. The metaphors tested were as follows:

- BI (Table 1): When the participant points in the direction of the target (if the OP line intersects the target), white noise is generated. Otherwise, no sound is generated.

For the AP and DVH metaphors, the extracted spatial information is encoded by generating a sinusoid that varies in pitch. When the angle θ exceeds 90°, the sinusoid is no longer generated, allowing the participant to know that the target is behind them.

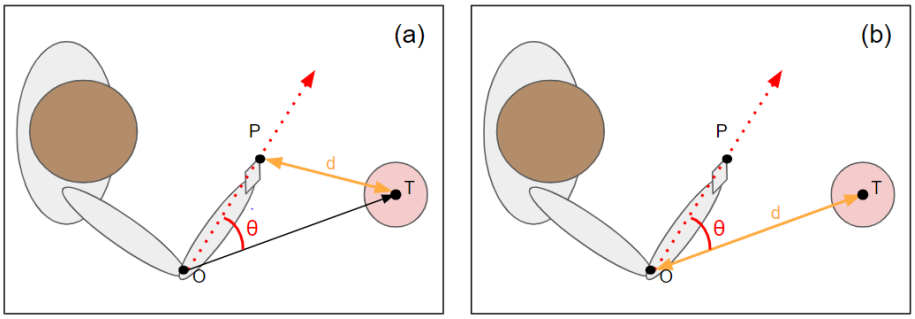

Fig. 2: Representation of classical polar coordinates versus natural pointing. (a) Natural pointing: angle θ has for origin O, and distance d has for origin P. (b) Classical polar coordinates: angle θ and distance d have the same origin O.

– AP (Table 2): The sinusoid's pitch f varies on a continuum from $fmin = 110$ Hz for $\theta max = 90°$ to $fmax = 440$ Hz for $\theta min = 5°$ so that (1) no sound is delivered from the rear (2) f is in the audible spectrum while avoiding higher pitches that are harsh in a constant stimulus. The angle-to-pitch conversion follows Steven's psychophysical law [21], a power law function:

$$log(f) = A * log(\theta) + B$$

with :

$$A = [log(fmax) - log(fmin)]/[(log(\theta min) - log(\theta max)] = 0.48$$

$$B = log(fmin) - A * log(\theta max) = 6.86$$

The intensity is constant. White noise is superimposed on the sinusoid when the OP line intersects the target (as for BI).
– DVH (Table 3): This metaphor dissociates the horizontal and vertical axes. The sinusoid's pitch varies according to the same conversion function as for AP, but according to angle θ_h which is the projected angle θ on the horizontal plane. The horizontal dimension is thus coded by the pitch, while the vertical dimension is coded by the activation of white noise when the participant points at the same height as the target position, i.e. when $\Delta Z < 15cm$ (i.e. the radius of the sphere). The intensity is constant.

2.6 Protocol

The experiment takes place in the motion capture space described above. Participants must find several targets presented in succession. Each participant makes one of two comparisons (BI VS AP or AP VS DHV). The experiment begins with a training phase, consisting of a block of six trials per metaphor. During this phase, the participant receives explanations from the experimenter on the

Table 1: Sound parameters of the metaphor BI

	Pitch	White noise
Horizontal		OP intersects
Vertical		the target

Table 2: Sound parameters of the metaphor AP

	Pitch	White noise
Horizontal	$\theta min = 5°$, $\theta max = 90°$	OP intersects
Vertical	$fmin = 110$ Hz, $fmax = 440$ Hz	the target

Table 3: Sound parameters of the metaphor DVH

	Pitch	White noise
Horizontal	$\theta_h min = 5°$, $\theta_h max = 90°$ $fmin = 110$ Hz, $fmax = 440$ Hz	
Vertical		ΔZ<target's radius

difficulties that may arise and how to overcome them. The session continues with the completion of four experimental blocks alternating the two metaphors. The order of passage of the metaphors is counterbalanced between the participants. Each experimental block is preceded by a refreshment block of the next metaphor. Each participant takes as many trials as necessary, and indicates when ready to start the test block.

Each block takes place as follows: the participant stands in the center of the room and closes their eyes. The participant starts the first trial as soon as they are ready by pressing the marked N key on the mini-keyboard. A start-up sound indicates the beginning of the trial. The sound feedback is triggered and changes according to the participant's movements, depending on the sound metaphor used. For the three metaphors, a buzzer (square signal) is triggered when the participant's finger passes the target ([OP] > [OT]). When the pointer P enters the target ($d < 15$ cm), only white noise is triggered, which intensity is stronger than the white noise of good direction or good height. When P stays inside the target for 400 ms, a sound indicates the victory and the end of the trial, and the sound feedback is switched off. The trial is interrupted if the participant does not reach the target within the 180-second limit, and a bell sound indicating defeat is triggered. After each trial, the participant stays on the spot and starts the next trial with the keyboard. It continues until all six trials in the experimental block are completed.

At the end of the experiment, the participants filled out a questionnaire. They were asked their general opinion about the experiment, the challenges

they encountered and to rate on a Likert scale each of the two metaphors on three aspects: ease, comfort and efficiency.

3 Results

3.1 Quantitative assessment

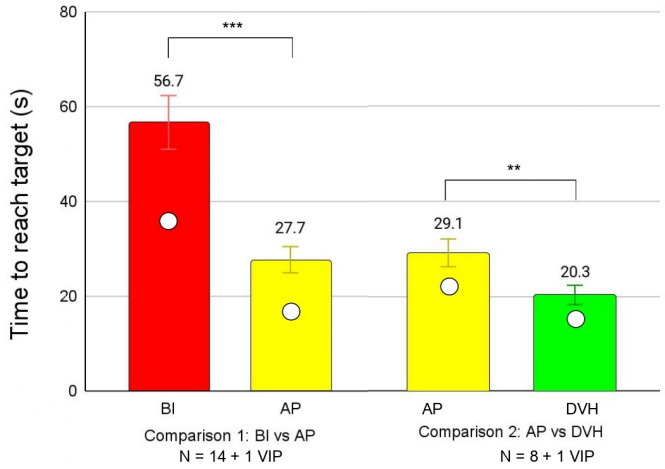

Fig. 3: Mean target-reaching times for each metaphor comparison. The white disks represent the average times obtained by visually impaired participants (VIP, one for each comparison). The bars represent the 90% confidence interval.

To compare metaphors, we measured our participants' objective performance by the time taken to reach the target. Since our response variable (time to target) is a duration with a skewed distribution, the Cox model appeared the most appropriate [7]. It allows us to analyze repeated measures without averaging the data for each participant, so it accounts for intra- and inter-participant variability. To perform our analyses, we used the *coxph* function in the *Survival* package of R software.

Comparison 1 (BI vs. AP): As shown in Fig. 3 (left) participants were 29 seconds faster on average with AP (M = 27.70 s; SD = 12.27 s) than with BI (M = 56.68 s; SD = 33.78 s) (z = -10.41; $p < 0.001$).

Comparison 2 (AP vs. DVH): As shown in Fig. 3 (right) participants were 10 seconds faster on average in the DVH condition (M = 19.70s ; SD = 10.77s) than in the AP condition (M = 29.85s ; SD = 17.70s) (z = 4.80 ; $p < 0.01$).

As we only had one visually impaired participant per metaphor comparison, their results are not included in the analyses. However, we can observe that the pattern of their results is similar to those of the sighted participants, with shorter reaching times for AP than for BI, and for DVH than for AP. Their results are represented by white circles in Fig. 3 and 4.

3.2 Subjective rating

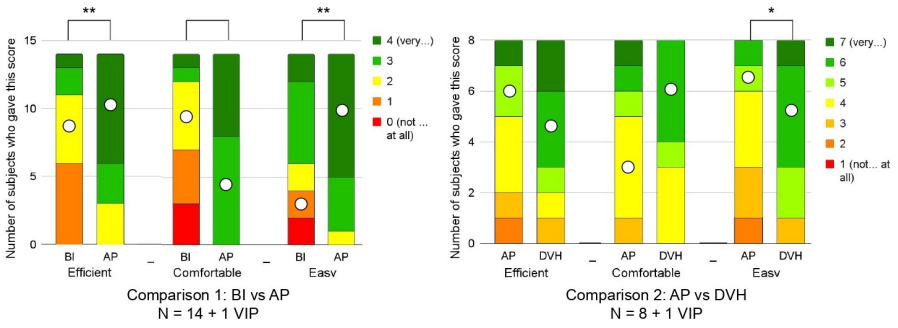

Fig. 4: Distribution of scores about efficiency, comfort, and ease-of-use for each metaphor comparison. The white circles represent the scores made by visually impaired participants (VIP one for each experiment).

Participants also gave a score to each metaphor on three criteria: efficiency, comfort, and ease. For each of the three criteria, we want to explain the form and intensity of the relationship between the score and the metaphor. The score takes as value an integer. We chose to look at the probability distribution for the values taken, and thus consider the score as an ordinal categorical variable. We performed an ordinal regression, using the *clm()* function of the ordinal package of the R software.

Comparison 1 (BI vs. AP): According to the scores given to each metaphor (see Fig. 4, left), AP is considered to be more comfortable than BI (z = -56.62, $p < 0.001$), more efficient ($z = 2.74$, $p < 0.01$) and easier to use ($z = 3.02$, $p < 0.01$).

Comparison 2 (AP vs. DVH): The task was considered easier with DVH than with AP (see Fig. 4, right) (z = -2.29; $p = 0.02$).

The scores given by the visually impaired participants are similar to those given by the sighted participants, with better scores for AP than for BI and better scores for DVH than for AP.

4 Discussion

Our goal was to create the most efficient and ergonomic sound guidance device possible to assist VIPs in target-reaching tasks. Most of the time, distance between the user and the target is calculated from the sensor, often a handheld or head-mounted camera. Here, the frame of reference is the participant's own body, a more natural egocentric frame of reference that is appropriate for the task and the user.

We created three sound metaphors taking into account advice from the literature to design accurate sound guidance, but also considering the cognitive load placed on the user. We tested these metaphors in pairs on a virtual target reaching task with the hand. The participants' performance was measured by the time to reach the target, and their evaluation of each metaphor was collected by a questionnaire.

We first compared BI, which gives a binary information of good direction, to AP, which gives a direction deviation in angle with respect to the target, expressed by the pitch of the sound in a continuous scale. The results showed the superiority of AP over BI in terms of speed, ease, comfort and efficiency. Transmitting only binary information about the direction is not enough. There is an advantage in giving the deviations to the target in the way of a hot-cold game. With BI, the participant has to scan the space for audio information and has to make a lot of unnecessary movements to effectively scan the area. AP quickly and effortlessly provides information on a 3-D spatial metric that separates the participant from the target. AP has several qualities identified by Ziemer et al. [24–27], but its dimensions are not orthogonal. Indeed, a single sound parameter is used to encode direction on both vertical and horizontal axes. We observed that the subjects wandered around a target when they were close to it when using AP. This problem could be due to the lack of orthogonality of the dimensions. At a distance, 3-D angle variations originated essentially from movements on a horizontal plane, while vertical components could be neglected; at proximity, both azimuth and elevation deviations contributed to the same sound continuum. A confound of the gravity axis and the horizontal plane is not consistent with the embodied representation of space [3, 6]. Action-perception loops allow inferring components along either axis with AP, but at the cost of an additional cognitive load. Such cost was alleviated using the DVH metaphor.

Indeed, in a second step, we compared AP to DVH, a metaphor whose deviations on the horizontal axis are coded on a continuous scale by the pitch of the sound, while the deviations on the vertical axis are coded in a binary way, by the triggering of white noise of good height. The results showed the advantage of DVH: participants were faster and found this metaphor easier than AP. With DVH, the two dimensions are orthogonal: the participant can interpret them separately. They are separated into two streams, which overlap only when the correct height is reached. The white noise of right height is an additional flow containing binary information, it does not require one to divert one's attention from the other flow to interpret it. Using binary information simplifies the soni-

fication, reduces the amount of information to be integrated, and decreases the cognitive load.

The results of this second comparison with a metaphor that dissociates horizontal and vertical axes are consistent with those of [10], a study in which target-reaching times were shorter with a sound indicating the elevation of the target (added to spatialized sound). However, this experiment was conducted by moving an avatar in a virtual world, making it impossible to compare the latencies with our study. On the contrary, in [13], the participants were physically moving in a 3*3 m room while being immersed in a virtual environment thanks to a Virtual Reality (VR) headset. The target position on the horizontal axis was coded by spatialized sound and on the vertical axis by sound frequency. The average latency to reach the target was 25 s. We obtained comparable latencies, with 19.70 s for DVH and 29.85 s for AP.

5 Conclusion

Sightless participants reached a target faster and preferred to be guided acoustically by directional deviations from the target than by the right direction only. These advantages are enhanced when the vertical and horizontal dimensions are given by two distinct sound streams. Visually impaired participants behaved similarly to blindfolded ones. Effective sonification of target guidance requires a balance between overly detailed - or strictly necessary - information to achieve sufficient accuracy without creating cognitive overload. The first comparison AP VS BI shows the benefit of additional information over a more parsimonious metaphor. The benefits of DVH that explicitly dissociates the two axes, alleviates the participant from the cognitive load to do it oneself; the load is further reduced by converting verticality into binary information [2], which has shown sufficient to guide someone to a target [12]. The DVH metaphor is a good trade-off between parsimonious (use of binary information) and explicit (dissociation of the two axes) guidance. It is a first step towards smartphone SSDs applications to help find objects in a situation of visual impairment.

Acknowledgements This work was supported by the Agence Nationale de la Recherche (ANR-21CE33-0011-01). It was authorized by the ethical committee CER Grenoble Alpes (Avis-2018-06-19-1).

This work has been partially supported by ROBOTEX 2.0 (Grants ROBOTEX ANR-10-EQPX-44-01 and TIRREX ANR-21-ESRE-0015) funded by the French program Investissements d'avenir.

The authors acknowledge the BIOMECA facility (GIPSA-lab UMR5216) for the experiments.

We would also like to thank Charles Fricaud, Laurent Bourque and Juliette Suslian for their contribution to this project during their internship.

Bibliography

[1] Chang, W.J., Chen, L.B., Sie, C.Y., Yang, C.H.: An artificial intelligence edge computing-based assistive system for visually impaired pedestrian safety at zebra crossings. IEEE Transactions on Consumer Electronics **67**(1), 3–11 (2020)

[2] Gao, Z., Wang, H., Feng, G., Lv, H.: Exploring sonification mapping strategies for spatial auditory guidance in immersive virtual environments. ACM Transactions on Applied Perceptions (TAP) (2022). https://doi.org/10.1145/3528171

[3] Graf, W., Klam, F.: Le système vestibulaire: anatomie fonctionnelle et comparée, évolution et développement. Comptes Rendus Palevol **5**(3-4), 637–655 (2006). https://doi.org/10.1016/j.crpv.2005.12.009

[4] Hild, M., Cheng, F.: Grasping guidance for visually impaired persons based on computed visual-auditory feedback. In: 2014 International Conference on Computer Vision Theory and Applications (VISAPP). vol. 3, pp. 75–82. IEEE (2014). https://doi.org/10.5220/0004653200750082

[5] Katz, B.F., Kammoun, S., Parseihian, G., Gutierrez, O., Brilhault, A., Auvray, M., Truillet, P., Denis, M., Thorpe, S., Jouffrais, C.: Navig: augmented reality guidance system for the visually impaired. Virtual Reality **16**(4), 253–269 (2012). https://doi.org/10.1007/s10055-012-0213-6

[6] Lamy, J.C.: Bases neurophysiologiques de la proprioception. Kinésithérapie scientifique **472**, 15–23 (2006)

[7] Letué, F., Martinez, M.J., Samson, A., Vilain, A., Vilain, C.: Statistical methodology for the analysis of repeated duration data in behavioral studies. Journal of Speech, Language, and Hearing Research **61**(3), 561–582 (2018). https://doi.org/10.1044/2017_JSLHR-S-17-0135

[8] Lin, Y., Wang, K., Yi, W., Lian, S.: Deep learning based wearable assistive system for visually impaired people. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 0–0 (2019)

[9] Liu, Y., Stiles, N.R., Meister, M.: Augmented reality powers a cognitive assistant for the blind. ELife **7**, e37841 (2018). https://doi.org/10.7554/eLife.37841.001

[10] Lokki, T., Grohn, M.: Navigation with auditory cues in a virtual environment. IEEE MultiMedia **12**(2), 80–86 (2005). https://doi.org/10.1109/MMUL.2005.33

[11] Manduchi, R., Coughlan, J.M.: The last meter: blind visual guidance to a target. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 3113–3122 (2014). https://doi.org/10.1145/2556288.2557328

[12] Marston, J.R., Loomis, J.M., Klatzky, R.L., Golledge, R.G.: Nonvisual route following with guidance from a simple haptic or auditory display. Journal of Visual Impairment & Blindness **101**(4), 203–211 (2007). https://doi.org/10.1177/0145482X0710100403

[13] May, K.R., Sobel, B., Wilson, J., Walker, B.N.: Auditory displays to facilitate object targeting in 3d space. In: The 25th International Conference on Auditory Display (ICAD 2019). Georgia Institute of Technology (2019). https://doi.org/10.21785/icad2019.008

[14] Meijer, P.B.: An experimental system for auditory image representations. IEEE transactions on biomedical engineering **39**(2), 112–121 (1992). https://doi.org/10.1109/10.121642

[15] Neugebauer, A., Rifai, K., Getzlaff, M., Wahl, S.: Navigation aid for blind persons by visual-to-auditory sensory substitution: A pilot study. PLoS One **15**(8), e0237344 (2020)

[16] Parseihian, G., Gondre, C., Aramaki, M., Ystad, S., Kronland-Martinet, R.: Comparison and evaluation of sonification strategies for guidance tasks. IEEE Transactions on Multimedia **18**(4), 674–686 (2016). https://doi.org/10.1109/TMM.2016.2531978

[17] Parseihian, G., Ystad, S., Aramaki, M., Kronland-Martinet, R.: The process of sonification design for guidance tasks. J Mob Med **9**(2) (2015)

[18] Ruvolo, P.: Considering spatial cognition of blind travelers in utilizing augmented reality for navigation. In: 2021 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops). pp. 99–104. IEEE (2021). https://doi.org/10.1109/PerComWorkshops51409.2021.9430997

[19] Shih, M.L., Chen, Y.C., Tung, C.Y., Sun, C., Cheng, C.J., Chan, L., Varadarajan, S., Sun, M.: Dlwv2: A deep learning-based wearable vision-system with vibrotactile-feedback for visually impaired people to reach objects. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 1–9. IEEE (2018). https://doi.org/10.1109/IROS.2018.8593711

[20] Spagnol, S., Hoffmann, R., Martínez, M.H., Unnthorsson, R.: Blind wayfinding with physically-based liquid sounds. International Journal of Human-Computer Studies **115**, 9–19 (2018)

[21] Stevens, S.: On the physiological law. Psychol. Rev **64**, 153–181 (1957)

[22] Thakoor, K., Mante, N., Zhang, C., Siagian, C., Weiland, J., Itti, L., Medioni, G.: A system for assisting the visually impaired in localization and grasp of desired objects. In: European Conference on Computer Vision. pp. 643–657. Springer (2014). https://doi.org/10.1007/978-3-319-16199-0_45

[23] Troncoso Aldas, N.D., Lee, S., Lee, C., Rosson, M.B., Carroll, J.M., Narayanan, V.: Aiguide: An augmented reality hand guidance application for people with visual impairments. In: The 22nd International ACM SIGACCESS Conference on Computers and Accessibility. pp. 1–13 (2020). https://doi.org/10.1145/3373625.3417028

[24] Ziemer, T., Nuchprayoon, N., Schultheis, H.: Psychoacoustic sonification as user interface for human-machine interaction. arXiv preprint arXiv:1912.08609 (2019). https://doi.org/10.48550/arXiv.1912.08609

[25] Ziemer, T., Schultheis, H.: A psychoacoustic auditory display for navigation. In: The 24th International Conference on Auditory

Display (ICAD 2018). Georgia Institute of Technology (2018). https://doi.org/10.21785/icad2018.007

[26] Ziemer, T., Schultheis, H.: Psychoacoustical signal processing for three-dimensional sonification. In: The 25th International Conference on Auditory Display (ICAD 2019). Georgia Institute of Technology (2019). https://doi.org/10.21785/icad2019.018

[27] Ziemer, T., Schultheis, H.: Three orthogonal dimensions for psychoacoustic sonification. arXiv preprint arXiv:1912.00766 (2019). https://doi.org/10.48550/arXiv.1912.00766

[28] Zientara, P.A., Lee, S., Smith, G.H., Brenner, R., Itti, L., Rosson, M.B., Carroll, J.M., Irick, K.M., Narayanan, V.: Third eye: A shopping assistant for the visually impaired. Computer **50**(2), 16–24 (2017). https://doi.org/10.1109/MC.2017.36