Audio Interface of a Navigation Aid for the Visually Impaired

Anonymous Author(s)

ABSTRACT

Our aim is to build a navigation system for the visually impaired that uses a combination of feedback modes to guide the user to his/her destination. In this paper, we investigate the effectiveness of a spatial audio tone with a varying pitch component, played with bone-conducting headphones, in conveying the pan and tilt angles of a target to the user in a pointing task. We also wish to see how changes in the behaviour of the pitch affects a user's performance. We conducted a set of experiments with blindfolded users and found that the varying pitch component works well in conveying the tilt angle of a target. Furthermore, we were able to determine that the audio interface adheres to Fitts's Law and used it as a metric to determine which pitch setting produces the best results. We discovered a trade-off between the speed and accuracy in the pointing task, which are maximised when the tone-settings is adjusted to low and high respectively.

Figure 1: The Tango device and bone-conducting headset (right) used by a blindfolded subject (left) during our experiments.

CCS CONCEPTS

• **Human-centered computing** → **Usability testing**; *Empirical studies in HCI*; User interface toolkits;

KEYWORDS

Human-machine interface, visually impaired, navigation aid, spatialised sound, Fitts's Law, pointing task

ACM Reference format:
Anonymous Author(s). 2017. Audio Interface of a Navigation Aid for the Visually Impaired. In *Proceedings of International Conference on Multimodal Interaction, Glascow, Scotland, November 2017 (ICMI'17),* 10 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

In recent years, governments have passed numerous laws to support the disabled and enable them play a more active role in modern society and the Royal National Institute for Blind People has prioritised enabling the visually impaired (VI) to use some of the services and products many people take for granted, such as public transport and cellphones [31]. Improvements in modern computing have made it possible for new and innovative solutions for these problems come to the fore.

To this end, we are developing a mobile device-based navigation system based on a Google Project Tango device, pictured in Figure 1, that caters to the needs of the VI by using a non-visual interface, instead of a visually-driven one that many systems currently use, and allow them to reach a target location in the final leg of their journey, i.e. the so-called 'last 10 yards' problem, although the ideas here can be extended and applied to macronavigation tools (e.g. Google Maps) as well.

A Tango-enabled device comes pre-equipped with powerful image-processing, localisation and depth-perception capabilities and is built on top of a standard Android platform, providing access to the entire set of input/output options that Android has to offer. The final system will use multiple feedback modes (vibration, audio and voice) to guide a user toward a target destination while providing information on any oncoming obstacles.

In this paper, we focus on the audio feedback mode and discuss how it is used in our system, the experiments we performed to determine how effective this mode is at directing a user to complete a pointing task and how its parameter values affect a user's performance.

For the pointing task, we presented the subjects with a set of virtual targets, one of another, and asked the subjects to simply point a camera to where they thought the targets were. The targets' pan and tilt angles w.r.t. the Tango's vertical plane were given to the subjects through a spatial tone with varying pitch, played via a set of bone-conducting headphones.

We use external bone-conducting headphones since we do not wish to interfere with the user's normal hearing function which the VI tend to rely upon. Furthermore, these headphones bypass the external structure of the ear responsible for localising a sound source's elevation, making it necessary to convey the tilt angle using another method.

Unfortunately there is a gap in literature regarding the use of a tone with a varying pitch component to convey a target's tilt angle for pointing tasks. It is also unclear whether popular metrics, such as Fitts's Law, can be applied in this case. Fitts's Law is a predictive model in the field of psychomotor research and used in the human-machine interfacing field that relates the time it takes a user to direct a pointing device toward a target as a function of the

ICMI'17, Glascow, Scotland
© 2017 ACM. 978-x-xxxx-xxxx-x/YY/MM...$15.00
DOI: 10.1145/nnnnnnn.nnnnnnn

difficulty of finding the target, i.e. the ratio between the distance to the target and its width.

The contributions of this paper are two-fold:

- we provide the first experimental results on how well a tone with varying pitch can convey a target's tilt angle;
- we show that this sound-based human-machine interface conforms to Fitts's Law and can provide a metric of performance for the interface.

The remainder of this paper is organised as follows: Section 2 provides an overview of existing navigation systems for the VI as well as existing audio interfaces. Section 3 and Section 4 discusses the Tango and the navigation system, along with a description of how the pan and tilt angles are conveyed to the user. The three experiments that were conducted are then discussed and explained in Section 5, while the results are presented and discusses in Section 6. Finally, Section 7 concludes this paper with a brief summary of the findings made in this paper.

2 PREVIOUS WORK

Delivering a system that allows the VI to independently navigate and accomplish everyday tasks is not new; in fact, there are multiple commercial systems and research prototypes currently available. These products vary from sonar, radar and GPS-based systems, to some of the more recent systems which use computer vision techniques to detect and avoid obstacles in the user's path.

One approach that has been investigated is to outfit the existing white walking cane with various sensors, such as sonar, radar, motor encoders, etc., [6, 13] to warn the user of upcoming obstacles from a distance instead of relying on haptic feedback from the impact between the cane and the obstacle.

Another approach is to outfit a walking cane to act as a radio-frequency identification (RFID) antenna that can read a set of RFID tags that are placed around the environment at key spots or along a path [8, 38]. This modification to the traditional cane is more discreet than the systems mentioned earlier and has been shown to work well. However, the major drawback here is the significant cost of modifying existing infrastructure with RFID tags and maintaining them to keep up with a changing environment. GPS systems [15, 22, 29], while cheap and reliable in outdoor environments, are not applicable in built-up urban areas and indoors where GPS signals are notoriously unreliable.

Computer vision-based systems provide a good compromise between usability, cost and accuracy and has been the focus of much research in the recent past [24]. One popular solution is to use an RGB-D depth sensing camera, which are becoming increasingly more accurate and cheaper, to build a 3D image map of an environment which will allow a user to safely traverse through it [18, 32]. Another approach is to use object recognition techniques to detect various objects and landmarks, such as doors, staircases, etc., and communicate their relative location to the user [36].

An important feature of user-centric systems is a human-machine interface (HMI) that enables effective and seamless two-way communication between the system and the user. In their surveys, researchers found that the VI prefer receiving feedback and instructions in the form of speech and haptic feedback cues, preferring the haptic feedback to the audio feedback [17, 34]. However, haptic feedback modes typically have a lower data bandwidth when compared to audio feedback and also requires the user to wear a special device in order to transmit the haptic signals to the user effectively. Work has also been done in translating a visual scene into format that is useful to the VI, with so-called sensory substitution systems (e.g. 'The Voice' [37]) and virtual audio reality (VAR) systems [10] reporting favourable results. However, The Voice, while helpful, has a very steep learning curve that has proven to be a significant barrier to entry, and with the VAR system it is not clear how unknown environments, where markers have not yet been encoded, will be handled and described to the user.

Spatial audio has also been considered to convey the direction of a target and experimenters have previously determined that people are able to find the location of a stationary sound source with an error of $\pm 35°$ in both the pan and tilt dimensions [42] for both early-blind and normally-sighted people. However, in [19, 20] it was found that that the blind have a clear advantage in localisation accuracy over sighted people when presented with a more difficult task, such as targets in motion and narrow-band stimuli, with [20] reporting an average absolute localisation error of approximately $10°$.

Other authors determined that the minimum difference in the spatial sound's angle for the user to be able to perceive movement is approximately $1.7°$. During these experiments, a speaker was physically manoeuvred to provide the subject with a spatialised sound tone [3]. Researchers have also tried using simulated spatial audio to inform the user which direction to go in [14, 15, 26, 30, 39]. In these works, a sound is played through a set of headphones and the source is spatialised with a head-related transfer function (HRTF) in order to trick the listener into thinking the sound source is located at some arbitrary 3D location. Various audition techniques and methods are then used to guide the user along a path. Other authors have also provided a framework for evaluating the quality of this 3D sound in terms of its psychoacoustic properties [12, 27].

There are experimental results that determined how well users can find targets presented with spatial sound in the tilt and panning dimensions [16, 42], but to our knowledge, no extensive work or experiments have been done to determine how well users respond to tilt adjustment instructions using a tone with *varying pitch*.

Researchers have previously used Fitts's Law [9], and more recently MacKenzie's modified version of the law [23], as a metric to evaluate the performance of a spatial audio HMI system.

Fitts's Law was originally proposed for visual target search tasks. However, it has been applied in non-visual target search tasks as well. For example, experiments with a vibro-tactile feedback pointing device have been performed to determine how effective it is at directing a user to finding a target [1]. The authors found that the search time adheres to Fitts's Law. However, they also note that it is not a perfect fit, citing the fact that Fitt's Law does not take into account a user's search strategy as a possible reason.

Another group of researchers conducted experiments using a spatial audio interface to describe the position of a target on the horizontal plane [25]. Here, a subject pointed to where they thought the targets were on their left or right as they traversed along a path. Their results show a good Fitts relation between target difficulty and search time, providing a strong argument that Fitts's Law can be used to describe the performance of a spatial audio interface. These results have since been supported by findings from other authors, where they found that Fitts's Law provided a good fit for the results from an experiment they conducted using visual, limited visual and non-visual feedback cues [40]. However, Fitts's Law has not yet been shown to apply to a spatial tone that uses varying pitch to convey the target's tilt angle.

3 PORTABLE NAVIGATION SYSTEM

The system we intend to ultimately deliver is a portable navigation device that is capable of guiding a VI user on the last leg of the journey, e.g. a specific aisle in a shop. To do this, we will use a combination of different feedback modes to facilitate two-way communication between the user and the device. A large amount of data needs to be translated from a visual form into a format that is useful to the VI. We therefore plan to use a combination of voice, audio and vibration cues to translate the visual navigation data as effectively as possible and overcome the data bandwidth limitation of the human ear.

The system is based on a concept proposed in [4, 21], which uses a Google Tango device, pictured in Figure 1. This is an Android-based cellphone or tablet device that comes equipped with an RGB-D camera to estimate depth. It combines an inertial measurement unit with powerful and robust landmark recognition and image processing algorithms to localise itself. An added benefit of this platform is its familiar, compact form-factor which will help overcome the hurdle of user-acceptance and usability.

We use a set of bone-conducting headphones (Figure 1) that are placed externally on the user's head so that the system does not occlude a user's perception of real-world sounds and does not interfere with the normal hearing function of a VI user. In the future, our system will use multiple feedback modes to provide the VI user with navigation and obstacle avoidance instructions. However, for this paper, we only considered the spatialised audio mode and its variation in pitch in order to determine its effectiveness in conveying pan and tilt angles to a user.

A diagram of the entire system pipeline is shown in Figure 2. Here, the arrows indicate the direction of the flow of information. When the user taps the Tango's screen, a new virtual target is generated and its coordinates are sent to the audio generation module, along with the Tango's current position and orientation. The audio generator then produces an audio tone based on the difference between the Tango and the target's positions in the Tango's global coordinate system and sends it to the audio output channel that plays it back to the user. The WiFi recording module is constantly monitoring the different parameter values of the Tango and target's positions, as well as the system's output, and records it to a remotely stored datafile.

4 AUDIO INTERFACE

For the series of experiments performed in this work, only the audio feedback mode was used to interface with the user. The audio component is responsible for conveying the 2D position of a target on the vertical plane w.r.t. the Tango's coordinate system, showed in Figure 3, in terms of pan and tilt angles. In this work, the distance to the target is not rendered. The audio signal is a sinusoidal

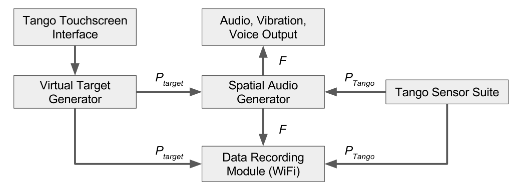

Figure 2: A diagram of the individual system components and their communication pipelines. F indicates a feedback signal and P and pose signal.

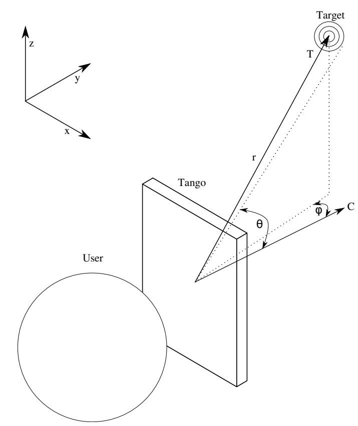

Figure 3: A diagram showing the Tango's frame of reference and the target pan and tilt angle that it conveys to the user.

sound wave that is continuously played to the user through bone-conducting headphones. We select a sinusoidal wave because it is relatively simple to manipulate and analyse.

The audio is spatialised using an HRTF provided by the OpenAL sound library [1]. However, the audio is only spatialised in the pan dimension, while the tilt angle is conveyed by varying the pitch of the audio tone. We use this approach because the external set of bone-conducting headphones plays the sound through the user's cheekbones instead of their outer ears, bypassing the pennae of the ears which provide humans their ability to localise an elevated sound source [2, 33], making the HRTF less effective in conveying the target's tilt angle and making it necessary to convey the tilt using another method, a varying tone pitch in this case. The difference between the target's angular position and the angular orientation of the Tango device are used to generate the audio navigation cues.

4.1 Pan Direction

The pan angle describes the angle which the user needs to rotate the camera vector around the vertical z-axis, how far the target is to the left or right of the user. We use an HRTF to add a spatial element

[1] openal.org

to the audio tone that the system plays to the user, making the tone sound user like it is coming from the direction of the target.

We implement OpenAL's default HRTF, based on the MIT's KEMAR dataset [2], to generate a sinusoidal sound wave based on the relative difference between the user and target's positions. We implement the library as a 'black box' where the inputs are position values and the output is a tone based on the angle between the two position vectors.

4.2 Tilt Direction

The system adjusts the tone's pitch to convey the target's tilt angle w.r.t. the camera's current pointing vector, C, as shown in Figure 3. Here, a high pitch means the target is above the camera vector and the user should look up, whereas a low pitch means the target is below the camera vector and the user should look down. This high/low association scheme was chosen because humans naturally tend to associate high-pitched sounds with higher objects and lower-pitched noises with lower objects [5, 28]. We also opt for a logarithmic, octave-based gain function for the pitch, since an increase in octave provides a distinct perceptible change while keeping the timbre roughly similar [35].

We wish to determine how the gradient of the pitch gain function affects a user's performance. For example, does an increased rate of change in the pitch as a function of the tilt angle lead to an increased target acquisition rate? For this we select three different pitch gain gradients, so-called *lo*, *med* and *hi* gain presets. To find these gradients, we set the maximum and minimum limits for the tilt angle and the maximum and minimum frequencies for the pitch. Furthermore, for the sake of consistency, each gradient is set to pass through the same pitch value at the 0 rad tilt angle.

The neutral, 0 rad, position is set to be directly in front of the user and we limit the angles between $\pm\frac{\pi}{2}$ rad, requiring the system to be able to communicate angles within a range of π rad. Anything outside of this range implies that the target is behind the user.

After practical tests with the Tango and the headphones, we set the neutral, on-target tone to a frequency of 512 Hz for its audibility. For the *med* preset, we set the maximum and minimum pitches to be two octaves higher and lower than the neutral tone, giving limits of 2048 Hz and 128 Hz respectively. The *lo* preset is set to one octave higher and lower than the neutral tone (1024 Hz and 256 Hz) and the *hi* to 3 octaves higher and lower (4096 Hz and 64 Hz) than the neutral tone. We selected these limits for practical reasons, given the fact that the bone conducting headphones we use have low volume gain at very high and low frequencies, making it difficult to hear. The frequency is selected using Equation 1, where θ is the tilt angle and m and c are parameters dependant on the cut-off frequencies.

$$f = 2^{-m\theta+c} \tag{1}$$

5 EXPERIMENTS

We performed a set of experiments with blindfolded users using the spatial audio feedback mode to determine how effective it is at directing a user to perform a given task. Here we determined how effective a spatial tone with varying pitch is at directing a user to adjust the pan and tilt angles of a camera to point it at a target. Furthermore, we also carried out a set of pre-screening experiments to determine each subject's hearing characteristics. The subjects were given time before they were blindfolded to familiarise themselves with the device and the tones it emits, as well as what the 'on-target' tone sounds like while using their sight.

We plan to use the results from the experiments we performed to better understand how the users respond to different settings for the spatial audio feedback stimulus in order to improve and optimise the behaviour of the feedback modes.

5.1 Experimental Procedure

For the experiments we used 42 sighted, but blindfolded volunteers and had them perform a series of experiments using our system and a pair of bone-conducting headphones. The subjects were recruited on a volunteer-basis and consisted of a diverse group of undergraduate students with ages ranging between 18 and 27 years (10 male, 32 female). The subjects also reported having no significant sight or hearing issues or any other major disability.

The subjects participated in 3 experiments, each of which are discussed here. The first 2 experiments were performed to determine each subject's hearing characteristics to provide some context to the results generated during the final, target-search experiment. These characterisation experiments were performed to check if the subjects had any pre-existing biases in the modes or dimensions we were going to perform our target search experiment in.

5.2 Subject Characterisation

5.2.1 Spatial Awareness. In this experiment, we evaluated a subject's ability to determine the direction a sound is coming from. To do this, we played a continuous 512 Hz sinusoidal tone to the subject through the headphones and applied an HRTF to it to make it sound like its coming from the left or right of the subject. The subject then had to select the direction the sound came from. The longer the experiment is run, the closer the source moves to the centre-front of the subject, making it more difficult to localise the sound source.

For this progressive increase in difficulty, a 2-up, 1-down step process is used, meaning that for every 2 correct answers, the distance to the centre halves, making the process harder. Conversely, it becomes easier for each incorrect answer by doubling the sound source's distance from the centre. We also use 2 different step sequences, one starting at a large distance (2 m) from the user and the other at the minimum distance (approximately 3 cm), giving an 'easy' and 'hard' step respectively. The terminating condition for the experiment is when the 2 step sequences are within 2 step ranges of one another, i.e. if one distance is less than four-times bigger or smaller than the other, for 3 consecutive guesses. This gives a distance band within which the subject is capable of localising the sound source. Each subject performs this experiment three times.

5.2.2 Pitch Discrimination. Here we determined a subject's ability to tell tones with different frequencies apart, i.e. how well they can tell if a tone is high or low pitched. Here we play 2 tones to the subjects in succession with the second tone being higher or lower-pitched than the first. The subjects were then asked to select whether the second tone was higher or lower-pitched than the first.

[2] sound.media.mit.edu/resources/KEMAR.html

The first tone is randomly generated while the second tone is generated by adding or subtracting the difference from the first tone. This difference determines the difficulty and is based on an exponential function, $f(n) = 2^n$, where n is increased or decreased to adjust the differentiation difficulty.

As with the spatial awareness experiment, a 2-up, 1-down step process is used: for every 2 consecutive correct answers, the pitch difference between the two tones is halved, increasing the difficulty, and the difference is doubled, i.e n is incremented by 1, for every incorrect answer, making the tones easier to differentiate. Two step sequences are again used here, one starting with a large pitch difference ($2^9 = 512$ Hz) between the tones and the other with a small difference ($2^1 = 2$ Hz). The termination condition is when the two step sequences are within one octave of each other for 3 consecutive answers. Each subject performed this experiment twice.

5.3 Target Search

5.3.1 Task Description. The final experiment is the main one and will answer the question we are most interested in: how well does a spatial tone with varying pitch direct a user to look in a specific direction, and how do the parameters of this tone affect the user's performance in this task?

For this experiment, the subject is blindfolded and given a Tango device running an app written specifically for this experiment. When started, a set of virtual targets are presented one at a time to the subject on the Tango device. Then, depending on the direction the subject is currently pointing the camera relative to the target's position, the Tango generates and plays a tone via the bone-conducting headphones to indicate to the subject the pan and tilt angle adjustment required to make the camera point to the target. These instructions are spatialised tones with varying pitch: an HRTF indicates the pan direction (left or right) and the pitch indicates whether the subject should be looking up (high pitch) or down (low pitch) to find the target.

Once the subject has pointed the camera toward the target, the HRTF centres the tone in front of the subject with a neutral pitch of 512 Hz, which we use as the 'on-target' pitch for all of our experiments. However, the subjects had to decide for themselves whether they truly were looking at the target and tap the screen to indicate the direction they believe the target was in. At this point a new target was presented to the subject which they had to search for.

28 targets are presented to each subject per round. The positions of these targets are randomly generated and are equally spread across the 4 quadrants on the vertical plane to prevent a lumping of targets at one location. After every round of experiments, the parameters controlling the tone's behaviour are adjusted. In this case, the rate of change of the tone's pitch is adjusted to make the pitch increase at a lower or higher rate as a function of the tilt angle between the target and the subject's current gaze direction. This is done to see whether, for example, a more rapid increase in pitch will help the subject find the target faster.

The distance between the subject and the target is not considered here and the 3D targets are therefore generated at a constant distance from the subject. Throughout the experiment, various parameters of the target and the subject are recorded and streamed in real-time to a laptop computer via a WiFi connection.

The subjects were given a few minutes without a blindfold prior to the experiment started where they could familiarise themselves with the system and confirm the target's location with their own eyes.

5.3.2 Metrics. We use two different metrics to compare the three different pitch gradient settings: the accuracy and search time.

The accuracy is given as the difference between the Tango's angular orientation at the time the subject confirmed they were on target, and the target's actual angular position. We separate the results from the tilt and pan dimensions in order to see how the different pitch gradients affect a subject's pointing accuracy.

We also compare the performance of the three pitch gradient settings in terms of the time it takes each subject to find a target. However, since each subject is presented with a different, randomly generated set of targets, a direct time comparison is not possible. Therefore, for this analysis, we opt to use Fitts's Law [9], modified by MacKenzie to give better results when working with uncertain target sizes and noisy data [23], which states that there is a relation between the time it takes to find a target and the index of difficulty of the target (the ratio between the distance to the target and its width). It also gives us a so-called 'index of performance' that we can use as a metric to compare the results between the three configurations. Fitts's Law is given by

$$t = a + bID. \tag{2}$$

Here t is the time it takes to find a target, a and b are constants determined through regression and ID is a description of the difficulty of the target, given as logarithmic function of the ratio between the distance to the target and the target's width. In our case, the targets have no width, since they are points in space, and we therefore use MacKenzie's modified form for ID, given in Equation 3, to provide a better approximation of ID. Also, for our experiments we investigate angles. Since we set the targets a constant distance from the user, this is a simple trigonometric transformation.

$$ID = \log_2 \left(\frac{\theta}{w_e} + 1 \right) \tag{3}$$

Here θ is the angular distance between subsequent targets' centres and w_e is the target's effective angular width, given by

$$w_e = \sqrt{2\pi e}\sigma = 4.133\sigma, \tag{4}$$

where σ is the standard deviation of the error data, taken here as the angle between the subject's target selection and target's actual angular position. Fitts's index of performance [9, p. 390], IP, can then be calculated using

$$IP = \frac{ID}{t}, \tag{5}$$

where $IP = \frac{1}{t}$ when $\sigma = 0$.

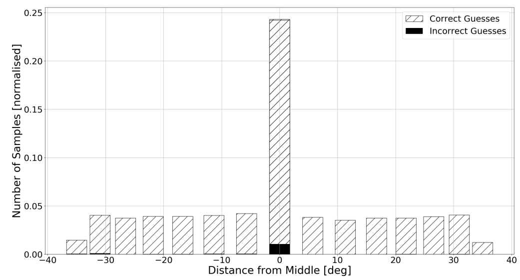

Figure 4: The subjects' guesses about the tone locations.

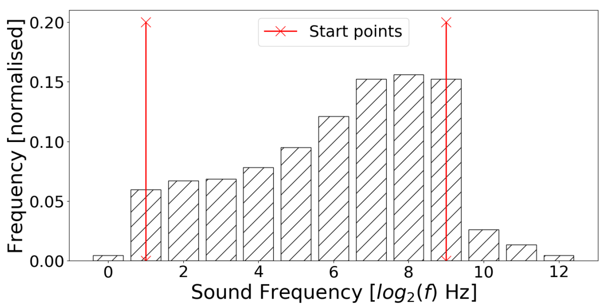

Figure 5: The guesses about the tone differences for the spatial awareness test.

6 RESULTS

6.1 Subject Characterisation

Figure 4 shows the results for the spatial awareness experiment, where the subjects had to determine the location of the sound they were played.

With the relatively low number of incorrect guesses, we see that the subjects were far more successful in correctly guessing the location of the sound source. This is further supported by the large number of samples at the minimum difference level of ~3 cm, indicating that the subjects reached this level more frequently and consistently. Here we can see that the subjects had little problem localising the left-right direction of a sound source.

These results are in line with what we expected and is supported by literature which indicates that humans are very adept at localising the location of a sound source and this ability was apparent for HRTF-generated pan location.

The results recorded during the pitch discrimination experiment are shown in Figure 5 where a bar plot is used to show the number of correct guesses for each tone difference level.

From Figure 5 we can see that the number of the correct guesses gradually decreases as the tone difference decreases and the majority of the samples in contained within the initial frequencies the subjects were presented with at the start of each step-sequence.

Along with this, we fitted a cumulative distribution function (CDF) over each subject's set of results and used the CDFs parameters to determine the cut-off threshold for each subject where the subject could no longer reliably tell two tones apart. We set this threshold at 75% of the correct guesses, starting from the largest tone differences, and we found that mean cut-off threshold frequency is approximately 13.4 Hz with an upper and lower 95% confidence interval of 2.6 Hz and 67.7 Hz respectively. The plot showing the subjects' threshold distribution can be seen in the top plot in Figure 8.

Table 1: A table of the pan and tilt angular error results for the target search experiment. The results include the absolute average error (μ_{abs}), average error (μ) and standard deviations (σ) given in radians, as well as the Pearson (R_P) and Spearman (R_S) correlation R-scores.

		μ_{abs}	μ	σ	R_P	R_S
Pan	*lo*	0.23	-0.02	0.32	0.72	0.78
	med	0.21	-0.01	0.28	0.76	0.80
	hi	0.22	-0.05	0.31	0.71	0.75
Tilt	*lo*	0.40	-0.14	0.46	0.34	0.40
	med	0.32	-0.12	0.36	0.44	0.52
	hi	0.34	-0.16	0.39	0.48	0.55

6.2 Target Search

6.2.1 Panning Results. The results from the target search experiment in the pan dimension are given on the abscissa of the 2D histograms in Figure 6, where the angular errors in the pan and tilt directions are plotted against each other in a 2D frequency histogram. A set of box-plots of the pan errors are also given in Figure 7 for each of the *lo*, *med* and *hi* configurations. The results are summarised in Table 1.

There is a very strong linear correlation between the subjects' guesses and the targets' true pan angles, shown by the high Pearson and Spearman correlation scores of above 0.7 for all of the datasets, with the *med* configuration displaying the best result at 0.75 for the Pearson score and 0.8 for the Spearman score and a statistical level of significance well below 5%. Figure 6 and the median and mean points from the box-plots in Figure 7 show that the data in the pan dimension is approximately normally distributed around zero.

The three settings have comparable average errors and standard deviations, with the *med* configuration producing the best results with the smallest average error and standard deviation. However, the differences are not big enough (approximately 6% difference, Friedman test p-value greater than 84%) to conclude that the pitch gradient has some effect on the performance in the pan dimension.

Based on previous research, these results were somewhat expected. However, they also confirm that as subject's target search capability in the pan dimension is fairly robust to the changing pitch we used to convey the target's tilt angle, which is a useful result going forward.

6.2.2 Tilt Results. The ordniates in Figure 6 also show the results recorded during the target search experiment for the tilt direction. A set of box-plots are also given in Figure 7 to convey the average tilt error between the subjects' guesses and the targets' true positions. All of the results are summarised in Table 1.

We found that there is a significant correlation between the subjects' guesses and the actual locations of the targets, shown by

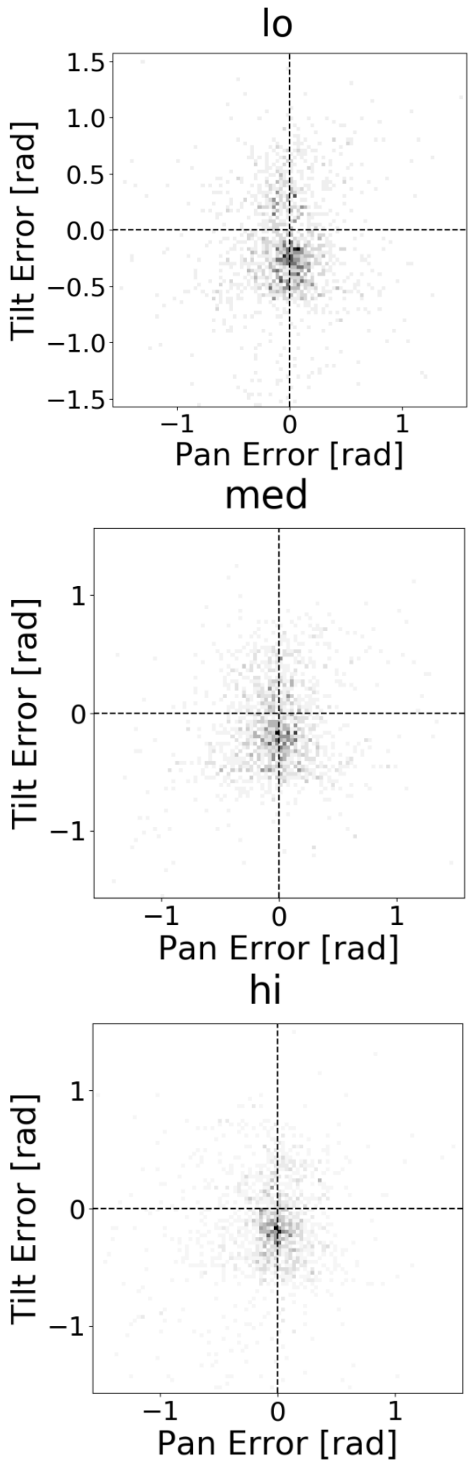

Figure 6: Plots containing the 2D frequency histogram plots for both the pan and tilt angular errors for the *lo*, *med* and *hi* pitch gain gradients respectively.

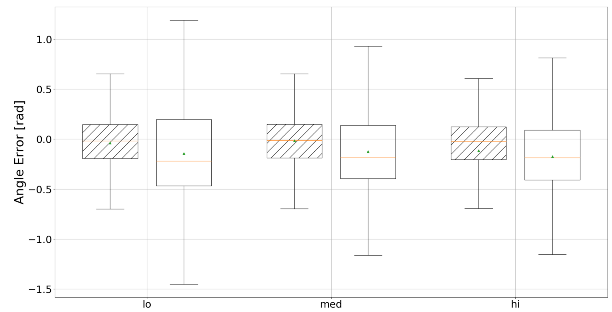

Figure 7: A set of box-plots to summarise the results in the angular error data for the pan and tilt dimensions. There is a plot for each of the *lo*, *med* and *hi* configurations. The filled box on the left is for the pan dimension and the empty box on the right for the tilt dimension.

relatively high shown Pearson and Spearman correlation scores of approximately 0.4 for each dimension, with the *hi* gradient giving the strongest R-scores of 0.48 and 0.55 respectively. This indicates that the varying pitch is working as expected and the subjects in general are interpreting the cues correctly. Both the Spearman and Pearson correlation scores have statistical significance below the critical 5% threshold level, indicating that it is reasonable to trust these correlation scores.

The average errors of the datasets are relatively close to one another, with the *lo* gradient giving the largest absolute error and standard deviation at 0.40 rad and 0.46 rad. The *med* and *hi* have similar absolute errors of 0.32 rad and 0.34 rad respectively. This is in line with the correlation scores, with the *lo* gradient giving a worse result than the *med* and *hi* gradients and the *hi* gradient giving the best results overall with its high correlation score and relatively low absolute angular errors and standard deviation.

From Figure 7 it can also be seen that the data is not normally spread, with a relatively large offset between the mean and median values. This shows a significant skewing to the negative side indicating a potential bias amongst the experiment subject-base that must be taken into account. This non-normality makes analysing the mean data with the conventional t-test and analysis of variance (ANOVA) unreliable. We therefore use the non-parametric version of the repeated-measure ANOVA test, i.e. the Friedman test [11], on the medians of the subject data to find the statistical significance of the differences between the datasets. We use the median values here since the data is not normally spread and there is significant noise within the data which may contaminate the mean values. This results in a p-value of 0.06%, which falls below the commonly-used 5% critical threshold and implies that there is a statistically significant difference between the three settings.

At this stage, it is unclear what causes this bias, but it is suspected that the floor introduces a direction constraint within the subjects' minds, where the target cannot appear below the ground, but can potentially appear well above the subjects' head, giving variable upper an lower limits that are dependant on the subjects' height and individual perception. Going forward, we may have to consider using a non-linear increase in pitch as a function of tilt angle instead

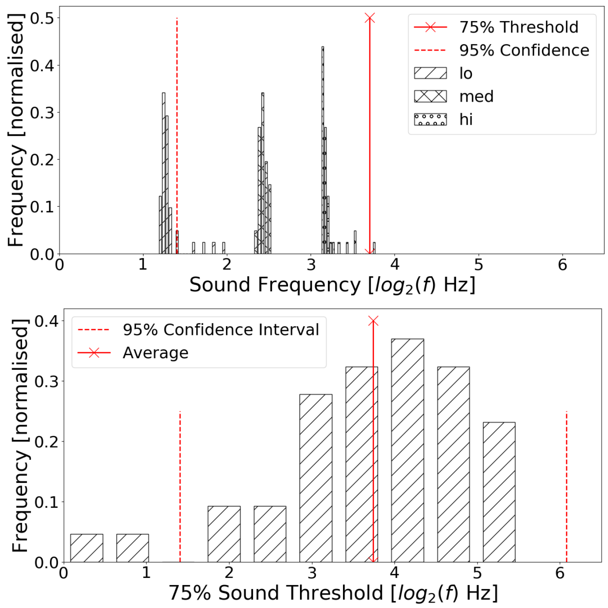

Figure 8: A plot of the distribution of the cut-off thresholds (top) and the medians of the error data in the tilt dimension with a frequency scale (bottom).

of the linear one we used for these experiments, to remedy this bias.

Furthermore, Figure 8 shows the distribution of the cut-off frequency thresholds that were found in Section 6.1, as well as a plot of the median values of the errors in the tilt dimension, transformed to a frequency using each dataset's respective pitch-gain gradient.

Figure 8 indicates that the subjects, on average, searched for the target until they could no longer detect a difference between the tone they were played and the 512 Hz on-target tone they were searching for, shown by the vast majority of the error data being located below the 75% cut-off threshold frequency. Furthermore, it can be seen that the *hi* dataset comes the closest to the cut-off frequency. This coult explain why it produces the smallest error: since the *hi* pitch gradient is the most sensitive to changes in the tilt angle, it allowed the subjects to get closer to the true tilt by playing a more easily distinguishable tone.

These results show that a tone with varying pitch can be used to convey the tilt angle of a virtual target to a human user using a set of bone-conducting headphones to a degree of accuracy similar to those previously established in literature [7, 16, 41]. They also highlight a clear and significant difference between the three different pitch gain gradients, with the *hi* pitch-gain gradient producing the results closest to the true tilt.

6.3 Time to Target

Figure 9 shows the box-plots of the time it takes to find a target as a function of the angular distance between the targets. Here, the bin interval is based on the smallest effective width from Equation 5 for the three datasets. We used the relation from Equation 2 and fitted a logarithmic line through regression for each subject then

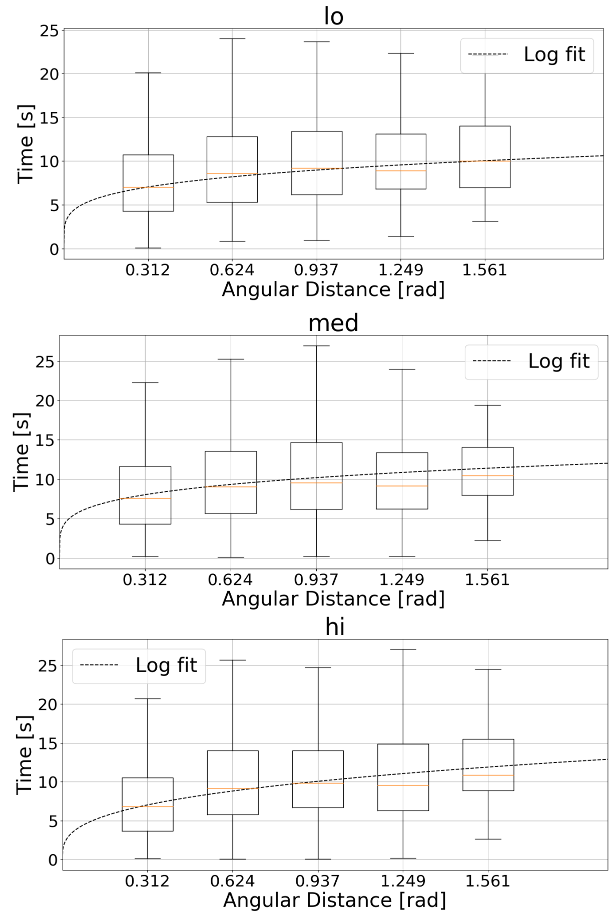

Figure 9: The plots showing Fitts's Law relation between the targets' index of difficulty and the time it took the subjects to find the target. There is a plot for the *lo*, *med* and *hi* pitch gain gradient configurations.

used the medians of the a and b parameters from Equation 2, found with the regression process, to plot the resulting lines of best fit shown in Figure 9.

From these figures we can see that the data approximates Fitts's Law, where the logarithmic line of best fit very closely approximates the median values of the binned data for all three pitch gradient settings. This result enabled us to use the index of performance, IP, given by Equation 5, which was used to plot Figure 10.

It can be seen that that the *lo* and *med* configurations give similar results, while the *hi* gives the worst results with the lowest slope.

A possible explanation for this behaviour is that the more extreme changes in the audio pitch with the *hi* configuration does a better job of informing the user when they are on target, leading to a more accurate estimate, but also increases the time it takes to point in the right direction. Conversely, the *lo* gradient makes it more difficult for the user to know when they are on target, leading to a shorter search time, but at the cost of a lower accuracy. This can be confirmed by the results obtained in Section 6.2.2.

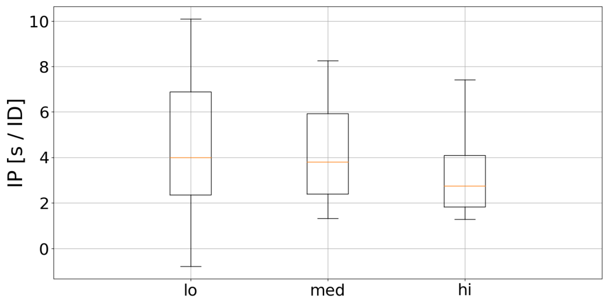

Figure 10: A comparison between the indices of performance for the three different pitch gradient configurations.

7 CONCLUSION AND FUTURE WORK

In this paper we presented a spatial audio interface to direct a subject to point a camera toward a virtual target. We also discussed a set of experiments we performed to determine its effectiveness and performance.

We found that a spatial audio tone with a varying pitch can be used to convey the pan and tilt angles of a target to a user using a set of bone-conducting headphones, and the angular errors made by the subjects are in line with the errors found in previous studies using similar audio interfaces. We also found that varying the pitch-gain gradient of our interface influences the accuracy of the system in the tilt dimension, as well as the time to target, without affecting the performance in the pan dimension. The steeper, *hi*, pitch-gain was found to produce the best results in this respect. Furthermore, we discovered a logarithmic relationship between the index of difficulty of a target and the time taken by a subject to find it, confirming that our interface adheres to Fitts's Law. However, there is a trade-off to be made between speed and accuracy, with the *hi* pitch-gain gradient directing the user to the target in the longest time.

Future research will focus on integrating voice and vibration feedback cues into the system, and to add the ability to automatically refine the parameters of the HMI to better match the individual user's navigation habits and capability, thereby increasing navigation performance and user satisfaction.

REFERENCES

[1] Teemu Tuomas Ahmaniemi and Vuokko Tuulikki Lantz. 2009. Augmented reality target finding based on tactile cues. In *Proceedings of the 2009 international conference on Multimodal interfaces*. ACM, 335–342.

[2] V Ralph Algazi, Carlos Avendano, and Richard O Duda. 2001. Elevation localization and head-related transfer function analysis at low frequencies. *The Journal of the Acoustical Society of America* 109, 3 (2001), 1110–1122.

[3] Daniel H Ashmead, Robert S Wall, Kiara A Ebinger, Susan B Eaton, Mary-M Snook-Hill, and Xuefeng Yang. 1998. Spatial hearing in children with visual disabilities. *Perception* 27, 1 (1998), 105–122.

[4] N. Bellotto. 2013. A Multimodal Smartphone Interface for Active Perception by Visually Impaired. In *IEEE SMC International Workshop on Human Machine Systems, Cyborgs and Enhancing Devices (HUMASCEND)*.

[5] Jens Blauert. 1997. *Spatial hearing: the psychophysics of human sound localization*. MIT press.

[6] J Borenstein and I Ulrich. 1997. The GuideCane - A Computerised Travel Aid for the Active Guidance of Blind Pedestrians. In *Proceedings of IEEE International Conference on Robotics and Automation*. 1283 – 1288.

[7] Michal Bujacz, Andrzej Materka, Michal Pec, Pawel Strumillo, and Piotr Skulimowski. 2011. *Sonification of 3d scenes in an electronic travel aid for the blind*. INTECH Open Access Publisher.

[8] José Faria, Sérgio Lopes, Hugo Fernandes, Paulo Martins, and João Barroso. 2010. Electronic white cane for blind people navigation assistance. In *World Automation Congress (WAC), 2010*. IEEE, 1–7.

[9] Paul M Fitts. 1954. The information capacity of the human motor system in controlling the amplitude of movement. *Journal of experimental psychology* 47, 6 (1954), 381.

[10] C. Frauenberger and M. Noisterig. 2003. 3D Audio Interface for the Blind. In *Proceedings of the International Conference on Auditory Display*. 280 – 283.

[11] Milton Friedman. 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association* 32, 200 (1937), 675–701.

[12] Catherine Guastavino and Brian FG Katz. 2004. Perceptual evaluation of multi-dimensional spatial audio reproduction. *The Journal of the Acoustical Society of America* 116, 2 (2004), 1105–1115.

[13] Marion A. Hersh and Michael A. Johnson. 2008. *Assistive Technology for Visually Impaired and Blind People* (1 ed.). Springer-Verlag London.

[14] Simon Holland, David R Morse, and Henrik Gedenryd. 2002. AudioGPS: Spatial audio navigation with a minimal attention interface. *Personal and Ubiquitous computing* 6, 4 (2002), 253–259.

[15] Slim Kammoun, Gaétan Parseihian, Olivier Gutierrez, Adrien Brilhault, Antonio Serpa, Mathieu Raynal, Bernard Oriola, MJ-M Macé, Malika Auvray, Michel Denis, and others. 2012. Navigation and space perception assistance for the visually impaired: The NAVIG project. *Irbm* 33, 2 (2012), 182–189.

[16] Brian FG Katz and Lorenzo Picinali. 2011. *Spatial audio applied to research with the blind*. INTECH Open Access Publisher.

[17] Wai Lun Khoo and Zhigang Zhu. 2016. Multimodal and alternative perception for the visually impaired: a survey. *Journal of Assistive Technologies* 10, 1 (2016), 11–26.

[18] Young Hoon Lee and Gerard Medioni. 2015. RGB-D Camera Based Wearable Navigation System for the Visually Impaired. *Computer Vision and Image Understanding* (2015), 3 – 20.

[19] Nadia Lessard, Michael Pare, Franco Lepore, and Maryse Lassonde. 1998. Early-blind human subjects localize sound sources better than sighted subjects. *Nature* 395, 6699 (1998), 278.

[20] Jörg Lewald. 2013. Exceptional ability of blind humans to hear sound motion: implications for the emergence of auditory space. *Neuropsychologia* 51, 1 (2013), 181–186.

[21] Jacobus Lock, Grzegorz Cielniak, and Nicola Bellotto. 2017. Portable navigations system with adaptive multimodal interface for the blind. AAAI Spring Symposium.

[22] Jack M Loomis, Roberta L Klatzky, Reginald G Golledge, and others. 2001. Navigating without vision: basic and applied research. *Optometry & Vision Science* 78, 5 (2001), 282–289.

[23] I Scott MacKenzie. 1992. Fitts' law as a research and design tool in human-computer interaction. *Human-computer interaction* 7, 1 (1992), 91–139.

[24] Roberto Manduchi and James M Coughlan. 2014. The last meter: blind visual guidance to a target. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 3113–3122.

[25] Georgios N Marentakis and Stephen A Brewster. 2006. Effects of feedback, mobility and index of difficulty on deictic spatial audio target acquisition in the horizontal plane. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. ACM, 359–368.

[26] Bob Menelas, Lorenzo Picinalli, Brian FG Katz, and Patrick Bourdot. 2010. Audio haptic feedbacks for an acquisition task in a multi-target context. In *3D User Interfaces (3DUI), 2010 IEEE Symposium on*. IEEE, 51–54.

[27] Rozenn Nicol, Laetitia Gros, Cathy Colomes, Markus Noisternig, Olivier Warusfel, Hélène Bahu, Brian FG Katz, and Laurent SR Simon. 2014. A roadmap for assessing the quality of experience of 3d audio binaural rendering. *10.14279/depositonce-4103* (2014).

[28] Carroll C Pratt. 1930. The spatial character of high and low tones. *Journal of Experimental Psychology* 13, 3 (1930), 278.

[29] Lisa Ran, Sumi Helal, and Steve Moore. 2004. Drishti: an integrated indoor/outdoor blind navigation system and service. In *Pervasive Computing and Communications, 2004. PerCom 2004. Proceedings of the Second IEEE Annual Conference on*. IEEE, 23–30.

[30] Marc Rébillat, Brian FG Katz, and Etienne Corteel. 2009. SMART-I 2:âĂIJSpatial multi-user audio-visual real-time interactive interfaceâĂİ, A broadcast application context. In *3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video, 2009*. IEEE, 1–4.

[31] RNIB. 2016. *UK Vision Strategy*. Technical Report. Accessed: 19-07-2016.

[32] Alberto Rodríguez, Luis M Bergasa, Pablo F Alcantarilla, Javier Yebes, and Andrés Cela. 2012. Obstacle avoidance system for assisting visually impaired people. In *Proceedings of the IEEE Intelligent Vehicles Symposium Workshops, Madrid, Spain*, Vol. 35. 16.

[33] Suzanne K Roffler and Robert A Butler. 1968. Factors that influence the localization of sound in the vertical plane. *The Journal of the Acoustical Society of America* 43, 6 (1968), 1255–1259.

[34] David A Ross and Bruce B Blasch. 2000. Wearable interfaces for orientation and wayfinding. In *Proceedings of the fourth international ACM conference on Assistive technologies*. ACM, 193–200.

[35] Roger N Shepard. 1964. Circularity in judgments of relative pitch. *The Journal of the Acoustical Society of America* 36, 12 (1964), 2346–2353.

[36] Yingli Tian, Xiaodong Yang, Chucai Yi, and Aries Arditi. 2013. Toward a Computer Vision-Based Wayfinding Aid for Blind Persons to Access Unfamiliar Indoor Environments. *Machine Vision and Applications* (2013), 521 – 535.

[37] Jamie Ward and Pieter Meijer. 2010. Visual Experiences in the Blind Induced by an Auditory Sensory Substitution Device. *Conciousness and Cognition* (2010), 425 – 500.

[38] S Willis and S Helal. 2005. RFID Information Grid for Blind Navigation and Wayfinding. In *Proceedings of the 9th IEEE International Symposium on Wearable Computers*. 34 – 37.

[39] Jeff Wilson, Bruce N Walker, Jeffrey Lindsay, Craig Cambias, and Frank Dellaert. 2007. Swan: System for wearable audio navigation. In *Wearable Computers, 2007 11th IEEE International Symposium on*. IEEE, 91–98.

[40] Jinglong Wu, Jiajia Yang, and Taichi Honda. 2010. FittsâĂŹ law holds for pointing movements under conditions of restricted visual feedback. *Human movement science* 29, 6 (2010), 882–892.

[41] Dmitry N Zotkin, Ramani Duraiswami, and Larry S Davis. 2004. Rendering localized spatial audio in a virtual auditory space. *IEEE Transactions on Multimedia* 6, 4 (2004), 553–564.

[42] MP Zwiers, AJ Van Opstal, and JRM Cruysberg. 2001. A spatial hearing deficit in early-blind humans. *Journal of Neuroscience* 21, 9 (2001), RC142–RC142.

Received May 2017