

Communicating the Angular Distance of a Virtual Target with Spatialised Audio with Varying Pitch

SUBTITLE

Anonymous Author(s)

GENERAL TERMS

KEY TERMS

KEYWORDS

Human-machine interface, multi-modal HMI, visually impaired, navigation, non-verbal instructions, spatialised sound, varying pitch

ACM Reference format:

Anonymous Author(s). 2017. Communicating the Angular Distance of a Virtual Target with Spatialised Audio with Varying Pitch. In *Proceedings of International Conference on Multimodal Interaction, Glasgow, Scotland, November 2017 (ICMI'17)*, 9 pages. DOI: 10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The UK's Royal National Institute for the Blind (RNIB), a leading organisation in the area, has identified a number of challenges for the modern blind and visually impaired (henceforth referred to as the VI) person. These include the latter's ability to safely and independently use public transport and navigate in unfamiliar environments [23]. Recent technological advances in the fields of mobile computing and computer vision have allowed for new and innovative solutions to come to the fore to address these challenges.

To this end, we are developing a mobile device-based navigation system that caters to the needs of the blind and VI that is based on a Google Project Tango device, pictured in Figure 1. 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 to guide a user toward a target destination while providing information on any oncoming obstacles.

In this paper, we discuss in detail how one of these feedback modes are used in our system. We also discuss the experiments we performed to determine how effective this mode is at directing a user to complete a pointing task, as well as how its parameter values affect a user's performance.

A note.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ICMI'17, Glasgow, Scotland

© 2017 ACM. 978-x-xxxx-xxxx-x/YY/MM. . . \$15.00

DOI: 10.1145/nnnnnnn.nnnnnnn

For the pointing task, we asked the subjects to simply point a camera to where they thought a virtual target was. The targets' location on the vertical plane was given to the subjects with a spatial tone with varying pitch to convey its pan and tilt angles respectively, played back through 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, using these kind of headphones bypasses 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.

The contributions of this paper are two-fold:

- these are the first experimental results obtained that provide an indication on how well a tone with varying pitch can convey the tilt angle of a target using bone-conducting headphones with a significant sample size and
- we show that this sound-based human-machine interface conforms to Fitts's Law and can provide a metric of performance for the interface.

Fitts's Law is a predictive model in the field of human-machine interfacing that relates the time it takes a user to direct a pointing device toward a target as a function of the difficulty of finding the target, i.e. the ratio between the distance to the target and its width.

The remainder of this paper is organised as follows: related work is presented and discussed in Section 2, followed by a description of the navigation system and its audio feedback interface in Sections 3 and 4. The experiments and the results they generated are then discussed in Sections 5 and 6. Finally, Section 7 gives a brief summary of the current work and an overview of the future work is then given in conclusion.

2 PREVIOUS WORK

In the past there has been much academic and commercial interest in the problem of empowering the blind and visually impaired to use the tools and systems normally-sighted people take for granted and to allow them to play a more active role in modern society. In this section we discuss some of the relevant work that has been done to enable this.

Delivering a system that will allow the blind and VI to independently navigate and accomplish everyday tasks is not a new proposal; in fact, there are multiple commercial systems 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., [5, 12] 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. These systems are fairly simple, reliable and are typically familiar to the VI, but often they are clunky, advertising the users' disability, with this clunkiness proving to be a major hurdle to market penetration.

Another approach was to equip a normal walking cane with special equipment to make it act as an radio-frequency identification (RFID) antennae to read a set of RFID tags that are placed around the environment at key spots or along a path [8, 30]. This approach to modifying 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, such as the Drishti system [22], while cheap and reliable in open 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 [19]. 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 which will allow a user to safely traverse through it [17, 24]. However, these systems have not yet been thoroughly tested with VI people in a real environment. Another approach is to use object recognition techniques to detect various objects and landmarks, such as doors, staircases, etc. [28], and communicate their relative location to the user. Furthermore, though not strictly useful for navigation but nice to have, are other assistive functionalities, such as face recognition and currency reading, that can also be added to enhance the VI user's experience and add more usability to the system [7].

An important feature of a user-centric navigation system is an effective human-machine interface (HMI) that enables effective and seamless communication between the system and the user. Creating such a system for the VI in particular can prove to be a challenge by itself. Some work has been done in this regard with a fair amount of success, from determining which feedback media the VI prefer to work on how to best translate a visual scene into a format that will be useful to a VI person.

In their survey, Khoo and Zhu [16] 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 [26], and they prefer to give instructions to the system using a familiar QWERTY (or the regional equivalent) keyboard or voice commands to query the system for output. 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.

On the front of user feedback, work has been done in translating a visual scene into format that would be useful to the VI, with so-called 'soundscapes' (e.g. Ward and Meijer [29] and their 'The Voice' system) 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 2D location of a target in the vertical plane and experimenters have previously determined that people are able to correctly find the location of a sound source within an error range of roughly $\pm 35^\circ$ in both the pan and tilt dimensions [33]. Furthermore, Ashmead et al. [3] determined that the minimum difference in the spatial sound's angle to be able to tell whether it has moved, is approximately 1.7° . During these experiments, the sound source (a speaker in this case) was physically manoeuvred to provide the subject with a spatialised sound tone.

Authors such as Holland et al. [14] have tried using simulated spatial audio to instruct the VI user which direction to go in. In their paper 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. Holland et al. report promising results with users being able to tell the direction of the sound source on the panning plane with a fair degree of accuracy (they did not report on the results for the distance or elevation).

There are experimental results to determine how well users can find a target presented with spatial sound in the tilt and panning dimensions [15, 33], but to the authors' knowledge, no extensive work or experiments have been done with a significant sample size to determine how well users respond to tilt adjustment instructions using an audio tone with *varying pitch*.

Researchers have previously used Fitts's Law [9], and more recently MacKenzie's modified version of the law [18], as a metric to evaluate the performance of a spatial audio HMI system. Fitts's Law states that there is a logarithmic relation between the time it takes a subject to point a device to a target on a 2D plane, and the difficulty index of the target, defined as the ration between the travel distance to the target and the width of the target. The performance of the HMI can be summarised as the ratio between the pointing time and the difficulty index.

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, Ahmaniemi and Lantz [1] performed experiments with a vibro-tactile feedback pointing device to determine how effective it is at directing a user to finding a target. They found that the search time is a function of the targets' index of difficulties and therefore adheres to Fitts's Law. However, they also note that it is not a perfect fit and citing the fact that Fitt's Law does not take into account a user's search strategy as a possible reason for this.

Another group of researches conducted an experiment using a spatial audio interface to describe the position of a target on the horizontal plane [20]. Here, a subject had to walk along a path and point to where they thought the targets were on their left or right as they traversed along the path. They reported that their results showed a good Fitts relation between target difficulty and the search time, providing a strong argument that Fitts's Law can be used to describe the performance of a spatial audio interface. 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. These results have since been supported by findings from Wu et al. [31], where they found that Fitts's Law provided a good fit for the results



Figure 1: A picture of the Tango device and the bone-conducting headset.

from an experiment they conducted using visual, limited visual and non-visual feedback cues.

3 PORTABLE NAVIGATION SYSTEM

The system we intend to ultimately deliver is a portable navigation device that caters to the needs of the blind by using 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 that of the human ear. The system is based on work previously done by Bellotto [4].

The system is based on a Google Tango device, pictured in Figure 1 along with the bone-conducting headset that we are using. A Tango device is an Android-based device available in a cellphone or tablet form-factor and comes equipped with an RGB-D camera to estimate depth and combines an inertial measurement unit (IMU) 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 that are placed externally on the user's head, with the reason being that the system should not interfere with a user's normal hearing function. This becomes particularly important when your target user-base has significant visual impairments.

We intend to use multiple feedback modes to provide the VI user with navigation and obstacle avoidance instructions. These modes are spatialised audio, voice prompts and vibration cues. However, for this paper, we only considered the spatialised audio mode in order to determine its effectiveness in conveying the tilt angle 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

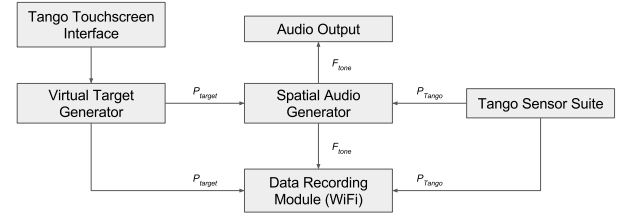


Figure 2: A diagram of the individual system components and their communication pipelines.

difference between these two sets of parameters and send it to the audio output channel that then plays it back to the user. The WiFi recording module is constantly monitoring the different parameter values of the Tango and target's position, 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, we only used the audio feedback mode to interface with the user. Here, the audio component is responsible for conveying the 2D position of a target on the vertical plane relative to the target in terms of pan and tilt angles.

The audio being generated is a sinusoidal sound wave that is constantly generated and played to the user through bone-conducting headphones. We select a sinusoidal wave due to it being relatively simple to manipulate and analyse.

The audio is spatialised using and HRTF provided by a third-party open-source sound library. 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 pinnae of the ears which provide humans their ability to localise an elevated sound source [2, 25], making it necessary to convey the tilt using another method. 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 Description

The pan angle describes the angle which the user needs to rotate the camera vector around the y -axis, i.e. how far the target is to the left or right of the user. To communicate this user, we use a HRTF to add a spatial element 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.

The HRTF functionality is implemented using the OpenAL library¹ [13] to generate a sinusoidal sound wave based on the relative difference in the user and target's positions. We implement the library as a 'black box' where the inputs are position values and it outputs a tone based on the the angle between the two position vectors and we implemented the same HRTF across all of the experiments we conducted.

¹<https://openal.org>

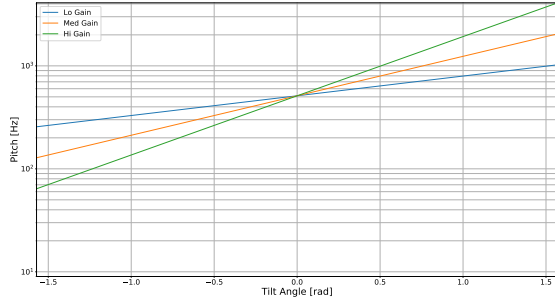


Figure 3: Plot depicting the different pitch gain preset functions.

4.2 Tilt Description

To communicate the tilt angle of the target, the system adjusts the emitted tone's pitch as a logarithmic function of the tilt angle, θ , between the user's camera vector and the target's position. 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 is selected, since humans naturally tend to associate high-pitched sounds with higher objects and lower-pitched noises with lower objects [21]. 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 [27].

We wish to determine what effect the gradient of the pitch gain function has on a user's performance; that is to say, does an increased rate of change in the pitch as a function of the tilt angle lead to an increased target acquisition rate, for example. For this we select three different pitch gain gradients, so-called *lo*, *med* and *hi* high gain presets. To find these gradients, we set the maximum and minimum limits for the tilt and pitch angles. 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 used have low volume gain at very high and low frequencies, making it difficult to hear. Figure 3 shows the *lo*, *med* and *hi* gain preset graphs.

5 EXPERIMENTS

To determine how effective the individual feedback modes of our HMI is at directing a user to perform a given task, we performed a set of experiments with blindfolded users using a limited set of the feedback modes.

In this case, we experimented with the spatialised sound feedback mode; that is to say we determined how effective a spatial tone, with varying pitch, is at directing a user to pan and tilt a camera to find a target. Furthermore, we also carried out a set of pre-screening experiments to determine each subject's hearing characteristics.

We plan to use the results from the experiments we performed to better understand how the users respond to different settings for the feedback stimuli mentioned earlier, 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, with the 10 being male and 32 female. The subjects also reported having no significant sight or hearing issues or any other major disability.

The subjects were asked to participate in 3 experiments, each of which are discussed here. The first 2 experiments were performed to determine each subject's hearing characteristics and capabilities to provide some context to the results generated during the 3rd and 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 evaluate a subject's ability to determine the direction a sound is coming from. To do this, we play a 512 Hz sinusoidal tone to the subject through the headphones and apply an HRTF to it to make it sound like its coming from either the left or right of the subject. The subject must then select whether the sound is coming from the left or right. 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, i.e. if one distance is less four-times bigger or smaller than the other, of each 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. In this experiment, we determine a subject's ability to tell tones with different frequencies apart, i.e. how well can they tell if a tone is high or low pitched? Here we play 2 tones to the subjects, one after the other, with the second being higher or lower-pitched than the first tone. The subjects are then asked to select whether the second tone was higher or lower-pitched than the first tone.

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

As with the spatial experiment, a 2-up, 1-down step process is used: for every 2 consecutive correct answers, the pitch difference between the two tones are 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 (512 Hz) between the tones and the other with a small difference (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. 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 the pan and tilt angle adjustment the subject needs to make the camera point to the target. These instructions are spatialised tones with varying pitch: an HRTF indicates whether the target is to the left or the right and the pitch indicates whether the subject should be looking up (high pitch) or down (low pitch) to find the target.

The distance between the subject and the target is not considered here. Therefore, the targets are generated on the vertical plane at a constant distance from the subject, in this case 2 m. 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.

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 subject have to decide for themselves whether they truly are looking at the target and tap the screen to indicate the direction they believe the target is in. At this point a new target is presented to the subject which they have to search for again.

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 were adjusted. In this case, the rate of change of the tone's pitch was 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 was done to see whether, for example, a more rapid increase in pitch will help the subject find the target faster.

The subjects were given a few minutes without a blindfold prior to the experiment started where they could familiarise themselves with the system where they could 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 the 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 tile 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 took each subject to find a target. However, since each subject was presented with a different, randomly generated set of targets, a direct time comparison is impossible. Therefore, for this analysis, we opt to use Fitts's Law [9] which states that there is a logarithmic 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 in Equation 1.

$$t = a + b \log_2 \left(\frac{d}{w} + 1 \right) \quad (1)$$

Here, t is the time it takes to find the target and d is the distance between the two subsequent targets' centre points, while w is the width of the target and a and b are constants determined through regression.

Fitts's Law uses the target's width as a parameter in its index of difficulty. However, our targets do not have a width, which means we have to use a modified form of Equation 1 that uses an effective width, w_e , as a parameter instead. Using this parameter was originally proposed by MacKenzie [18] in order to provide more accurate results when the process contains noise and uncertainties. In our case, the noise is introduced by the subjects not selecting the target at a constant distance from the target's centre.

w_e is given by

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

where σ is the standard deviation in the error data, taken as the difference in Euclidian distance between the subjects' target selections and the targets' actual positions. This results in the modified form of Fitts's Law, given in Equation 2.

$$t = a + b \log_2 \left(\frac{d}{4.133\sigma} + 1 \right) = a + bID \quad (2)$$

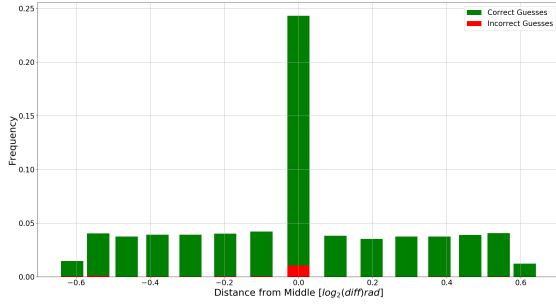


Figure 4: A bar plot showing the subjects' guesses about the tones' locations.

Here ID is shorthand for the Index of Difficulty. Fitts's index of performance [9, p. 390], IP , can then be calculated using the relation from Equation 3.

$$IP = \frac{ID}{t} \quad (3)$$

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 red, 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 0.031 25 m, 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 frequency of correct guesses versus the incorrect guesses for each tone difference level.

As one might expect, the number of samples at the extreme, larger tone differences are significantly higher than for the smaller differences, where around 80% of the respondents' correct answers were given at a difference of 16 Hz or more, with the reason being that it is easier to discriminate between two tones with a large tone difference. The number of incorrect guesses in this interval makes up approximately 50% of the total number of incorrect guesses. For frequencies less than 16 Hz, the number of correct guesses drops down to 20% of the total correct guesses, while the amount of incorrect guesses remain around 50% of the total incorrect guesses made.

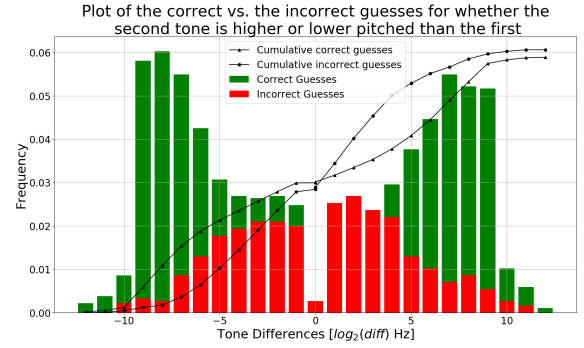


Figure 5: A bar plot showing the subjects' guesses about the tone differences.

Table 1: A summary of the tilt results for the target search experiment. The results include the absolute average, average and standard deviations given in radians, as well as the Pearson and Spearman correlation R-scores.

	μ_{abs}	μ	σ	R_p	R_s
<i>lo</i>	0.40	-0.14	0.46	0.34	0.40
<i>med</i>	0.32	-0.12	0.36	0.44	0.52
<i>hi</i>	0.34	-0.16	0.39	0.48	0.55

This indicates that our subjects' incorrect answers are fairly consistent across the frequency spectrum, but correct tonal discrimination is far more likely to occur with frequency differences larger than 16 Hz.

6.2 Target Search

6.2.1 Tilt Results. Figure 6 shows the results recorded during the target search experiment for the tilt dimension. Plots are presented for each of the three pitch gain gradients, i.e. *lo*, *med* and *hi*. The left-hand plots are 2D frequency histograms plotting the angular errors of the subjects' target selections in the pan and tilt dimension. A 1D histogram is also given for the tilt angle errors along with the 2D plots to help the reader visualise the spread of the error data. Finally, a set of box-plots are given in Figure 7 to convey the average angular error between the subjects' guesses and the targets' true positions. The results are summarised in Table 1.

We found that there is a strong correlation between the subjects' guesses and the actual locations of the targets, shown with strong Pearson and Spearman correlation scores. 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.

Both the absolute and raw averages 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

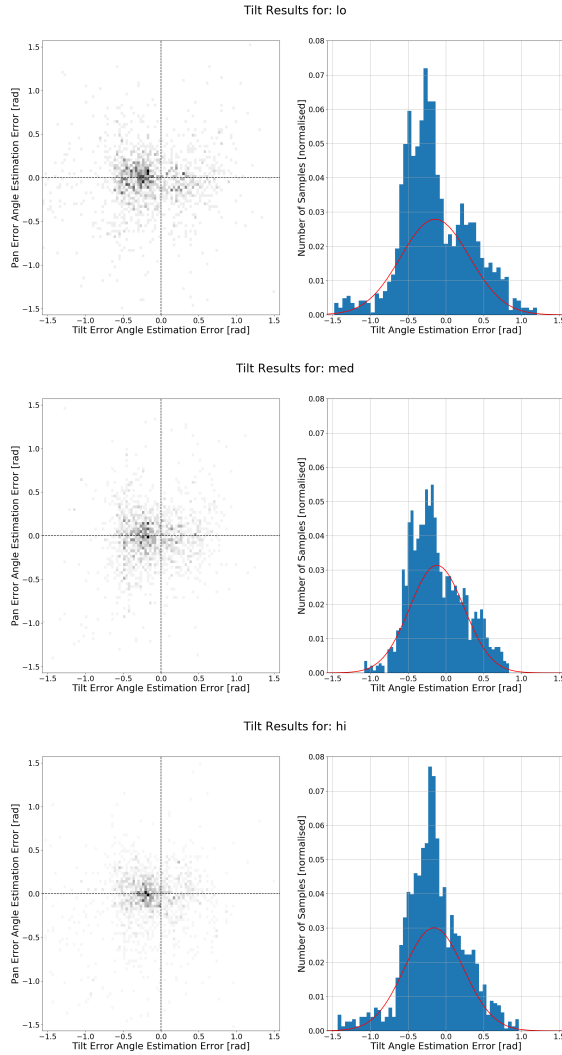


Figure 6: The results for the target search experiment in the tilt dimension. Each plot contains the results for the *lo*, *med* and *hi* pitch gain gradients respectively. The plot on the left is a scatter plot correlation graph between the subjects' guesses and the targets' true positions and the plots on the right are histograms for the error in these guesses.

giving a significantly 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.

It can also be seen that the data is not normally spread and displays 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 unreliable (ANOVA). We therefore use the non-parametric version of the repeated-measure ANOVA test, i.e. the Friedman test [11], on the

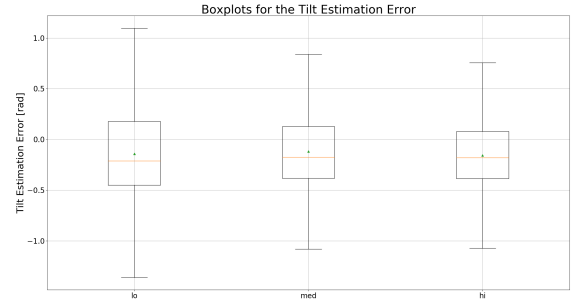


Figure 7: A set of box-plots to summarise the results in the angular error data from the tilt dimension. There is a plot for each of the *lo*, *med* and *hi* configurations.

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 median values. This results in a p -value of 0.06%, which falls below the commonly-used 5% critical threshold and implies that there is statistically significant difference between the three datasets.

At this stage it is unclear what causes this bias, but it is suspected that the ground introduced a position constraint within the subjects' minds, where the target could not appear below the ground, but can potentially appear well above the subjects' head, giving variable upper and lower limits that are dependant on the subjects' height and individual perception. We might have to consider using a non-linear increase in pitch as a function of tilt angle instead of the linear one we used for these experiments to remedy this bias. The skewness of the data makes it impossible to use

These results highlight a clear and significant difference between the three different pitch gain gradients, with the steep-gradient gain in pitch as a function of target tilt angle producing the results closest to the true tilt. Moreover, it shows that a tone with varying pitch can be used to convey the tilt angle of a virtual target to a human user to a degree of accuracy similar to those previously established in literature [6, 15, 32] using spatial sound, either with an HRTF or mobile speakers, and a head-mounted orientation tracker.

6.2.2 Panning Results. The results from the target search experiment in the pan dimension are given in the y -axis of the heatmaps in Figure 6, where the angular errors in the pan and tilt dimensions are plotted against each other in a 2D frequency histogram. A set of box-plots of the angular errors are also given in Figure ?? . A box-plot for each of the *lo*, *med* and *hi* configurations are given. The results are summarised in Table 2.

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, with a statistical level of significance well below 5%.

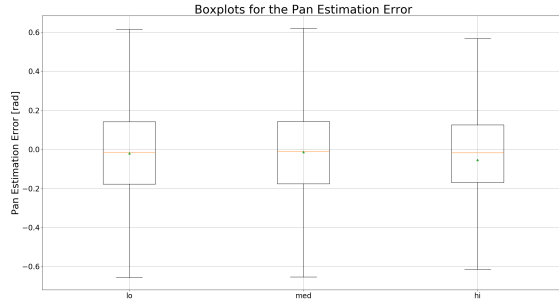


Figure 8: A set of box-plots to summarise the results in the angular error data from the pan dimension. There is a plot for each of the *lo*, *med* and *hi* configurations.

Table 2: A summary of the pan results for the target search experiment. The results include the absolute average, average and standard deviations given in radians, as well as the Pearson and Spearman correlation R-scores.

	μ_{abs}	μ	σ	R_p	R_s
<i>lo</i>	0.23	-0.02	0.32	0.72	0.78
<i>med</i>	0.21	-0.01	0.28	0.76	0.80
<i>hi</i>	0.22	-0.05	0.31	0.71	0.75

The angular errors are roughly normally distributed around zero with 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 a clear effect on the performance.

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

6.3 Time to Target

We used the relation from Equation 2 and fitted a line through the distance-time data we gathered. The raw data is displayed in binned form for the reader's convenience and a logarithmic line of best fit is drawn over the raw data. The resulting plots are given in Figure 9.

From these figures we can see that the data roughly obeys Fitts's Law and the logarithmic line of best fit very closely approximates the median values for the binned data for all three pitch gradient settings, with the exception of the relatively large offset in the last bin of the *hi* configuration. This result enables us to use the index of performance, given by Equation 3. We used this relation to plot Figure 10.

From Figure 10 that the *lo* and *med* configurations give similar results with the *hi* configuration gives the best result, where it took

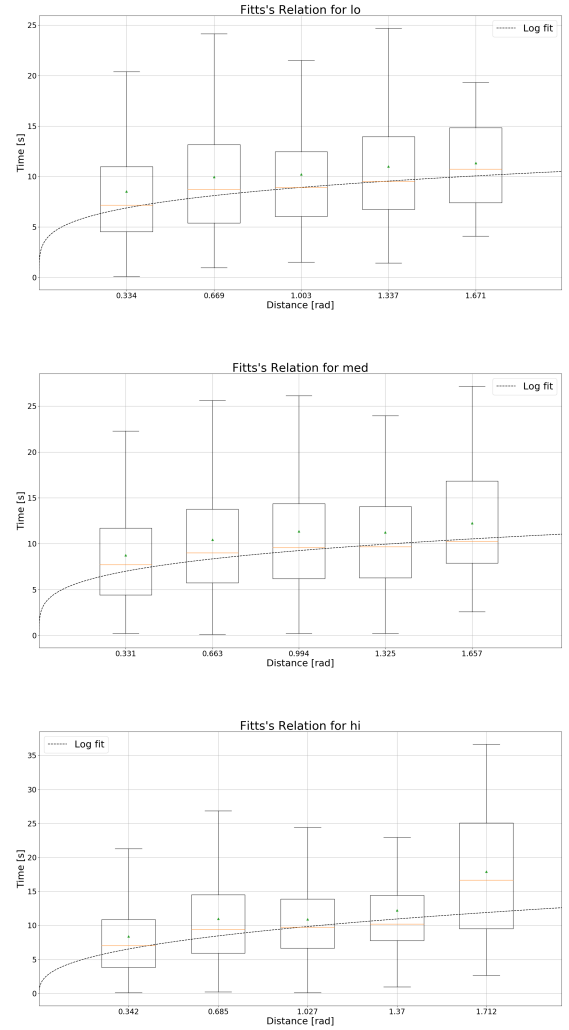


Figure 9: The plots showing Fitts's Law relation between the targets' index of difficulty and the time it took the subjects to find the target. There is a plot for the *lo*, *med* and *hi* pitch gain gradient configurations.

the average subject between 7 to 9 seconds less to find the most difficult target.

7 CONCLUSION AND FUTURE WORK

The next phase of the project is to add a co-adaptive module to the feedback interface, the goal of which is to refine the parameters of the feedback interface to better match the individual user's navigation habits and capability thereby increasing navigation performance and user satisfaction.

8 ACKNOWLEDGEMENTS

We would like to acknowledge and thank Google for their funding and support, as well as the UK's Engineering and Physical Sciences

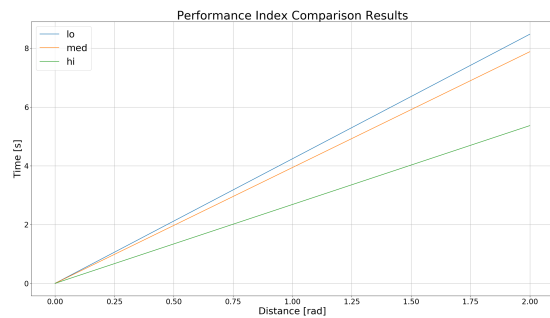


Figure 10: A plot comparing the indices of performance for the three different pitch gradient configurations.

Research Council (EPSRC) and Visual Image Interpretation in Humans and Machines (ViHM) for their funding and aid in carrying out these experiments.

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] 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.
- [6] 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.
- [7] Manuela Chessa, Nicoletta Noceti, Francesca Odone, Fabio Solari, and JoanLee Sosa-Garcia. 2016. An Integrated Artificial Vision Framework for Assisting Visually Impaired Users. *Computer Vision and Image Understanding* (2016), 209 – 228.
- [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] Marion A. Hersch and Michael A. Johnson. 2008. *Assistive Technology for Visually Impaired and Blind People* (1 ed.). Springer-Verlag London.
- [13] Garin Hiebert. 2005. Openal 1.1 specification and reference. (2005).
- [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] Brian FG Katz and Lorenzo Picinali. 2011. *Spatial audio applied to research with the blind*. INTECH Open Access Publisher.
- [16] 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.
- [17] 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.
- [18] I Scott MacKenzie. 1992. Fitts' law as a research and design tool in human-computer interaction. *Human-computer interaction* 7, 1 (1992), 91–139.
- [19] 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.
- [20] 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.
- [21] Carroll C Pratt. 1930. The spatial character of high and low tones. *Journal of Experimental Psychology* 13, 3 (1930), 278.
- [22] 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.
- [23] RNIB. 2016. *UK Vision Strategy*. Technical Report. Accessed: 19-07-2016.
- [24] Alberto Rodriguez, 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.
- [25] 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.
- [26] 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.
- [27] Roger N Shepard. 1964. Circularity in judgments of relative pitch. *The Journal of the Acoustical Society of America* 36, 12 (1964), 2346–2353.
- [28] Yingli Tian, Xiaodong Yang, Chucai Yi, and Aries Arditii. 2013. Toward a Computer Vision-Based Wayfinding Aid for Blind Persons to Access Unfamiliar Indoor Environments. *Machine Vision and Applications* (2013), 521 – 535.
- [29] Jamie Ward and Pieter Meijer. 2010. Visual Experiences in the Blind Induced by an Auditory Sensory Substitution Device. *Consciousness and Cognition* (2010), 425 – 500.
- [30] 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.
- [31] Jinglong Wu, Jiajia Yang, and Taichi Honda. 2010. Fitts's law holds for pointing movements under conditions of restricted visual feedback. *Human movement science* 29, 6 (2010), 882–892.
- [32] 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.
- [33] 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