3D Cues for Human Control of Target Acquisition in Auditory Augmented Reality [⋆]

Konstantinos Dadamis, John Williamson, Roderick Murray-Smith

School of Computing Science, University of Glasgow, Scotland.
(e-mail: Roderick.Murray-Smith@glasgow.ac.uk).

Abstract: We compare the effectiveness of different auditory cues for attracting attention to spatial targets around a mobile user, using a commercial 3D audio headset instrumented with GPS and inertial sensors. We compare two approaches to spatial audio feedback with a baseline case that only provides 'on target' feedback: 1. hints as single sounds played from a 3D location and 2. frequency modulation of inter-pulse gaps based on proximity. We illustrate the difference in user control behaviour created by the different forms of feedback with phase plots. Single 3D sound hints provided the best improvement over the baseline case of no hint. Frequency modulation of pulses performed more poorly for larger targets. The choice of sound has a significant effect on targeting performance and there is a significant trade-off between efficient targeting and aesthetically-pleasing audio.

1. INTRODUCTION

Instrumented headsets which can sense orientation, location and bearing can be used to augment the user's experience of the world with a virtual audio layer. Fusing location awareness with orientation sensing allows accurate alignment of the virtual layer with real-world objects. In mobile contexts visual attention is a scarce resource, and navigation systems based on audio and vibrotactile cues Williamson et al. (2010); Holland et al. (2002) have successfully provided spatial guidance without overloading the visual channel. Positional audio could increase the efficiency of these navigation mechanisms.

Aside from the benefits of disengaging from the visual display, the advantages of positional audio cues are twofold. Firstly, audio cues extend the field of awareness of the user, presenting information close to their current location, but out of their current field of view, as discussed in Bolia et al. (1999). Secondly, audio cues function as effective attention management elements. Animation is an essential part of modern interfaces as it directs user attention to key UI components; the audio counterparts of animation cues can apply this attention management for entities out of view. We explore a range of possible solutions for effectively and efficiently informing the user about the location of nearby points of interest.

We used the Jabra Intelligent Headset [1] which includes 3D audio, GPS location, magnetometers, accelerometers and gyroscopes in iOS with the Jabra API. This commercially-available, integrated hardware package simplifies the equipment requirements for spatial audio target acquisition and provides a potential mass market for spatial audio applications.

2. TARGET ACQUISITION

There are several challenges when it comes to designing auditory cues for spatialised content. The cues need to be both efficient and result in a pleasant user experience. Key aspects of auditory target display are informing the user about the nature and number of targets nearby and their bearing and distance from the user. This paper explores different feedback mechanisms for informing users about the bearing of a single given target. The purpose of guidance feedback is to ease (quicker, requiring less effort) aligning head orientation with that of targets around the user. It needs to give a user hints about which direction to turn, and how close the target is. This feedback can be a single event (e.g. a "ping" in the target direction), while in others it is an ongoing process providing gradient information to ease acquisition. For a review of the spatial audio targeting literature see Marentakis (2006); Gröhn et al. (2005); Strachan et al. (2005); Sandberg et al. (2006); Eriksson (2008); McGookin et al. (2009). We explore three feedback conditions:

1. Baseline condition. No Hint In this experiment, the user is given no cue about the target direction, to explore the user's behaviour and performance on the simplest scenario, as a baseline. Instead, she only relies upon the simple feedback when on-target. Consequently, in order to acquire the target, she turns her head until feedback is heard.

2. Single-sound Hint from 3D location For each trial, a 3D pulse sound (with duration of 0.3-1.00s) is played once, from the direction of the target. We experiment with a number of different sounds.

3. Frequency modulation of pulses based on proximity In this case, the feedback given to the user simulated the behaviour of feedback from parking sensors available in many modern cars, where the delay between short pulses represents the distance to another car. Assume that the feedback pulse has a duration of τ, and the angular size of the target is w_t. Within the target area, feedback is played continuously, as in the previous experiments. When the user is in the opposite direction (180° from the target's centre), the pulse's period was set to be $k\tau$, where k is a specified multiplier. Consequently, when the user's distance from the target is $\psi \in (w_t/2, 180^\circ]$, the pulse period is $\frac{\psi - w_t/2}{180^\circ - w_t/2}(k-1)\tau + \tau$. k was set to 20 and the pulse used was a sinusoidal tone of 261Hz of duration $\tau = 0.1$s.

[⋆] We acknowledge funding from GN Store Nord and EPSRC grant EP/R018634/1, *Closed-loop Data Science*

[1] **`https://www.jabra.co.uk/supportpages/jabra-intelligent-headset`**. Last accessed 25/1/2022.

As a pre-experiment, we investigated the sensitivity of faster cue-based acquisition to the specific sound used by exploring the impact of the types of sounds played on localisation speed. Our baseline was a simple sinusoidal tone (261Hz). When applying filters to white noise sounds, the widest filters (350-8000Hz) give the best localisation results Susnik et al. (2003), so the ideal sound should contain a wide range of frequencies. We tested two "blowing bottle" sounds from Cook (2002), and recorded two voiced vowel sounds. 4 further synthesized sounds were tested, Buzz-0004, Buzz-0035, Buzz-0036 and pulse1sec. 3D positional audio works best with sounds with strong transients and significant high-frequency content, for a clear inter-aural time delay and a perceptible effect of the head-related transfer function, which primarily modulates high-frequency components. As expected, the pure tone sound did not perform well, but surprisingly the richer "blowing bottle" sounds and the voiced sound performed worse. The sounds with fastest responses are Buzz-0004, Buzz-0036, the 'a' vowel recording and 'pulse1sec', which was the overall best. The aesthetic aspect of the sounds is important for the user experience, but the best performing sounds, apart from Buzz-0004, were considered to be somewhat robotic, squeaky or eccentric for a mainstream target acquisition application.

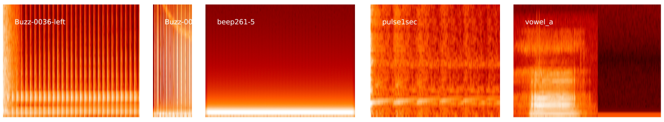


Fig. 1. Spectrograms of sounds used. Clipped to 8KHz max. freq, NFFT=1024, 1000 sample overlap.

3. EXPERIMENTAL SCENARIO

We investigate the impact of feedback choice on user behaviour, in terms of speed of action, nature of movement and user experience. In all cases, a simple 261Hz feedback tone indicates that the user is 'on-target'. The user is considered to be located at a fixed position, so that the GPS location uncertainty does not affect the results. As such the results provide a 'best case' scenario. The experimental task was performed on-campus, and the targets used are parts of the University of Glasgow's Main Building, as shown in Figure 2. Successful acquisition is defined as being achieved when the user looks towards the target bearing ψ_t for 3 seconds consecutively. The target is considered to be missed when the user has not acquired the target within 20 seconds. The targets are considered to be at the same distance d from the user and have the same angular w_t size (in degrees). We consider the user to be looking at the target when the direction is within the angle range $\psi_t \pm \frac{w_t}{2}$.

Fig. 2. Panoramic view of experiment location.

5 users aged between 23-46 and self-declared normal hearing tested the system. Each test for each of 7 target sizes w_t consists of the same ten targets in sequence, and the order of condition was cycled through participants. After acquiring or missing one target, the user was immediately presented with the next target. After all 10 targets at a given size w_t, the user rested for two minutes, then continued with the next sub-test of a different target size. Each experiment consists of 7 tests, for target sizes of $w_t = 5$, 10, 20, 40, 60, 90 and 180°, ordered from easiest (largest) to smallest, to give users progressively more challenging tasks. Within a particular feedback condition (e.g. frequency modulation of pulses) the order of the targets was kept constant. The users were not able to memorise the sequence. The sequence was varied across conditions. In most trials the users moved their whole body and not just the head in order to acquire the target. The sounds used are shown in the spectrogram plots in Figure 1.

4. ANALYSIS OF RESULTS

We view the user's behaviour from the perspective of a control system minimising the 'error' between the current bearing angle ψ and the reference or target bearing ψ_t, such that error $e = \psi_t - \psi$. The different feedback mechanisms will change the overall control system behaviour, Jagacinski and Flach (2003); Poulton (1974). In the 'no hint' case, feedback is only provided when over the target, so the user has an exploration behaviour. In the '3D hint' case feedback is provided once, at the start, to help the user infer target location so any error minimisation is being done by the user, with respect to the user's inferred target location. In the 'frequency modulation of pulses' method, explicit error feedback is provided in an ongoing fashion.

The experiments are sampled at 20Hz, and the data is smoothed using a Savistky-Golay filter (length 41 samples, order 4 polynomial), equivalent to least-squares polynomial fitting of a quartic polynomial to the last 2.05 seconds of data. This filter structure better preserves edges and transients than standard low-pass filters and can be used to robustly estimate derivatives.

The phase and polar plots shown in Fig. 3 provide a visual summary of acquisition performance. Target overshooting, oscillation and under-damped behaviour are all clearly visible. In contrast to time-series, phase plots make it easier to align and compare the dynamics of multiple acquisitions as time offsets are ignored. An example of an unsuccessful acquisition is illustrated in Figure 3a, where the user receives no hint about the location of the target, and fails to acquire it within the 20s time limit. The user enters and exits the target zone associated with the feedback tone starting and stopping, but continues to overshoot and 'hunt' around the small target.

4.1 No 3D hint

User performances are summarized in Figure 4, where the standard error of the mean time is indicated by error bars. The acquisition time decreases as the target size increases. No hint was slowest for all users apart from User 2 who's slowest condition was frequency modulation of pulses. The users missed significant numbers of targets on the 5 & 10° tests. Figure 3b shows a successful acquisition for a larger, easier target size of 20°. The user was initially close to the target (˜50°), but with no cue, chose the longer, slower way (310°).

4.2 With single sound 3D hint

Figure 4 shows a significant improvement in the acquisition times compared to the baseline case of *no hint*. The 3D hint that the user receives provides the approximate direction of the target and lets the user turn immediately towards that direction. As expected, the 3D audio made it clear whether a target is on the left or right of the user, but it was not as easy to perceive whether it was in front of or behind the user. The user's search for the target became faster on average for all users, with larger improvements for smaller targets.

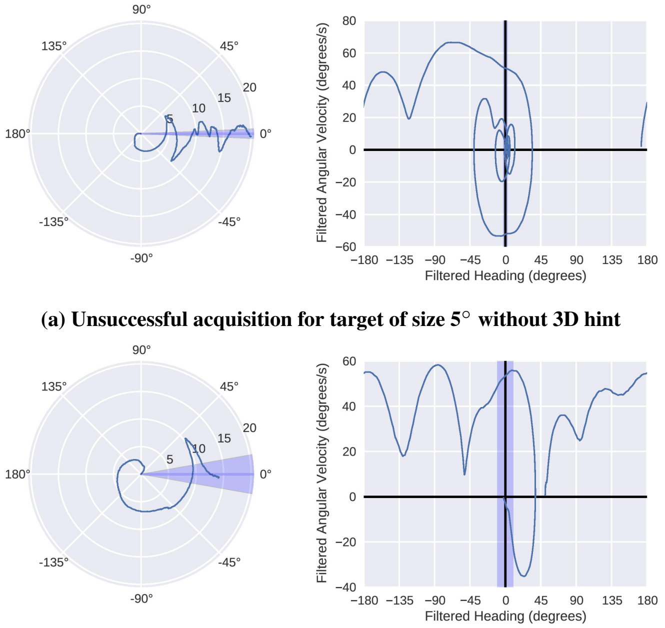
4.3 Frequency modulation of pulses based on proximity

Fig. 4 shows that for small targets (w_t 5-20°), this approach can speed responses over *No 3D hint*, but for larger targets (w_t 40-180°), it adds little, or gets worse. User 2 was slower throughout with this approach. Users had no initial hint of the target's direction, so to find the shortest path, some scanned the area around them, by quickly turning to the left and right (as shown in the edges of Fig. 5f), and then followed the path which increased the pulse frequency. Others went for one direction or the other until they heard the first target cues. The proximity indication of target reduces the velocity near the target.

Summarising the statistics, comparing ratios of means for each type of trial, the relative speed up using 3D hints over no hint is 40% ($\mu = 1.40$, $\sigma = 0.55$). The speed up of 3D hints over the Frequency approach is 25% ($\mu = 1.25, \sigma = 0.74$). The No Hint case had most misses ($\mu = 1.39$), followed by Frequency ($\mu = 1.21$) and fewest was 3D hint ($\mu = 0.79$).

4.4 Comparison of Phase plots

The use of phase plots to represent the error convergence allows us to show multiple acquisitions of different time lengths on a single plot which allows us to test for consistent changes in approach depending on the feedback style. We have grouped the responses to small ($w_t = 5°$) targets and large ($w_t = 60°$) targets. Smaller targets in Figure 5 show underdamped responses where the user oscillates around the target, whereas larger targets show overdamped responses where the user hits the target and stays there. For larger initial errors, the velocity decrease slows already before the target zone is reached in Figure 5b, suggesting that the user enters a different control mode (akin to Costello's *Surge Model* Costello (1968)), however for larger targets, in Figure 5e there is less anticipatory change, and no further oscillatory control near the target, leading to larger final errors. The 3D hint led to smoother velocity profiles outside the target zone, suggesting that the user has a good sense of target location, where other feedback mechanisms, especially no feedback and pulse frequency modulation, have more variation in bearing velocity. Oscillation around the target is worst with no hint. Surges from initial conditions to close to the target are larger for the single sound hint (the 3 largest velocities when crossing 0° are for the 3 initial conditions closest to the target) suggesting scope for improving performance for nearby targets.



(a) Unsuccessful acquisition for target of size 5° without 3D hint

(b) Successful acquisition for target of size 20° without 3D hint

Fig. 3. User behaviour without an audio cue, during one target acquisition. Left: evolution of the bearing error angle e on polar axes. The radial part r represents time, the angle e is the user's bearing error. The $0 \pm \frac{w_t}{2}$ target zone is shaded to ease comparison. Right: phase plots of time derivative $\dot{e}$ against the user's bearing error e.

4.5 Movement time and difficulty of task

We investigated the relationship between the movement time and the index of difficulty for the "No 3D hint" and "With single sound 3D hint" systems. Meyer et al. Meyer et al. (1990), developing earlier work Crossman and Goodeve (1983), proposed that the time (MT) to move to a target area is a function of the distance to the target (A) and the size of the target (W), $MT = a + bID$, where the index of difficulty, $ID = (\frac{A}{W})^{\frac{1}{n}}$, where n relates to the upper limit on submovements. $n = 2.6$ minimised the RMS error. Figs. 6a and 6b show the linear relationship between the MT and ID. The circles' radii $r \propto W$. Blue circles indicate that the user did not go past 180°, while red circles indicate the user took the long way round, with higher MT. Large targets have lower ID and lower MT.

5. CONCLUSIONS

We demonstrate auditory targeting behaviour with a commercial, instrumented headset. The headset and API provided a practical development platform for spatial audio systems. Experimental results demonstrate that the use of 3D hints for auditory targeting is an improvement over no feedback. User feedback indicated that this provides a simple, intuitive, aesthetically pleasing way for users to locate targets and required the least mental and physical workload. The pulse-frequency approach was less effective, slowing users for larger targets.

Visualisation of experimental results based on phase plots standard in control applications, can aid the design of bearing-based interaction. Phase plots allow rapid comparison of behaviour from time-series of varying lengths and present a clear visual summary of the dynamics of target acquisition, where the distance cue leads to a ballistic 'surge' phase Costello (1968) where the user moves towards the target zone to the final control phase. The additional 'radar' visualisation approach to time-series representation gives a clear representation of the head movement during the acquisition process, highlighting areas of high activity, which are likely to lead to lower usability results.

REFERENCES

Bolia, R.S., D'Angelo, W.R., and McKinley, R.L. (1999). Aurally aided visual search in three-dimensional space. *Human Factors*, 41(4), 664–669.

Cook, P.R. (2002). *Real Sound Synthesis for Interactive Applications*. A K Peters, Wellesley, Massachusetts.

Costello, R. (1968). The surge model of the well-trained human operator in simple manual control. *IEEE Transactions on Man–Machine Systems*, 9(1).

Crossman, E.R.F.W. and Goodeve, P.J. (1983). Feedback control of hand-movement and fitts' law. *The Quarterly Journal of Experimental Psychology*, 35(2), 251–278.

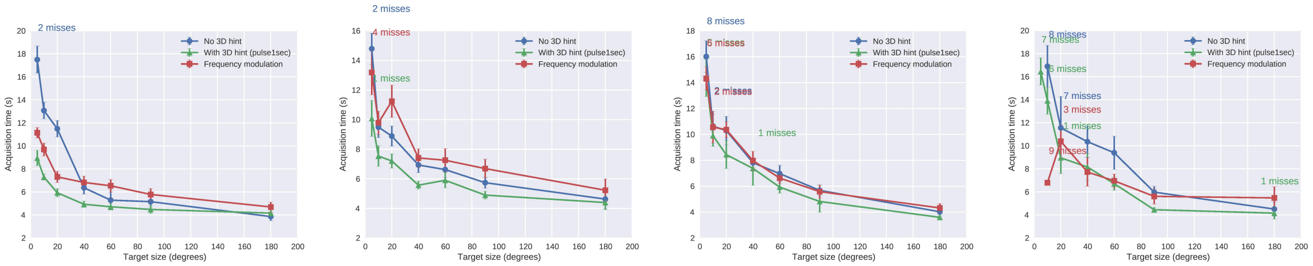
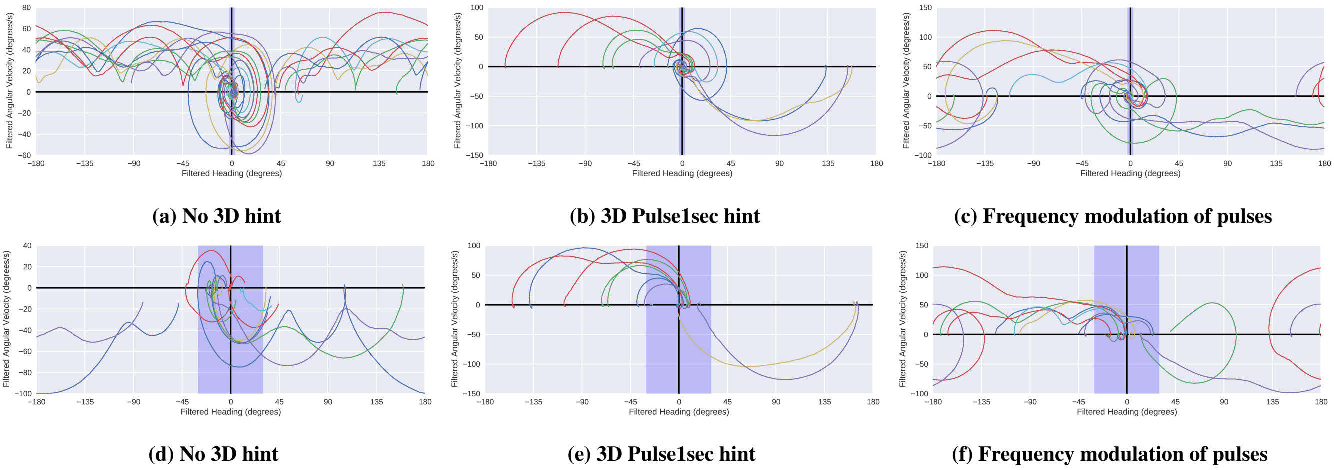


Fig. 4. Time-to-target results for users 1-4 in experiments with and without 3D hint and "Frequency modulation of pulses". Only successful selections are included in the mean & std. err. # of misses are shown beside each point.



(a) No 3D hint

(b) 3D Pulse1sec hint

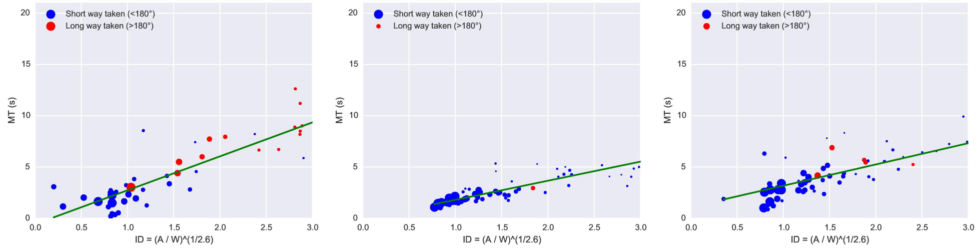
(c) Frequency modulation of pulses

(d) No 3D hint

(e) 3D Pulse1sec hint

(f) Frequency modulation of pulses

Fig. 5. Phase plots for User 1 for small ($w_t = 5°$, upper) and large ($w_t = 60°$, lower) targets. Blue = feedback zone.



(a) Without 3D hint. $a = -0.57, b = 3.31$

(b) With single 3D Pulse1sec hint. $a = -0.06, b = 1.86$

(c) Frequency modulation of pulses. $a = 1.11, b = 2.07$

Fig. 6. Meyer's Power Law analysis of successful targeting for User 1. Movement time vs Index of Difficulty

Eriksson, L.*et al.*. (2008). On visual, vibrotactile, and 3D audio directional cues for dismounted soldier waypoint navigation. In *Proc. HF & Erg. Soc.*, volume 52, 1282–1286. SAGE.

Gröhn, M., Lokki, T., and Takala, T. (2005). Comparison of auditory, visual, and audiovisual navigation in a 3D space. *ACM Trans. Applied Perception (TAP)*, 2(4), 564–570.

Holland, S., Morse, D.R., and Gedenryd, H. (2002). AudioGPS: Spatial audio navigation with a minimal attention interface. *Personal and Ubiquitous computing*, 6(4), 253–259.

Jagacinski, R.J. and Flach, J.M. (2003). *Control Theory for Humans: Quantitative approaches to modeling performance*. Lawrence Erlbaum, Mahwah, New Jersey.

Marentakis, G. (2006). *Deictic Spatial Audio Target Acquisition in the Frontal Horizontal Plane*. Ph.D. thesis, Univ. Glasgow.

McGookin, D., Brewster, S., and Priego, P. (2009). Audio bubbles: Employing non-speech audio to support tourist wayfinding. In *HAID*, 41–50. Springer.

Meyer, D., Keith-Smith, J.E., Kornblum, S., Abrams, R.A., and Wright, C.E. (1990). Speed-accuracy trade-offs in aimed movements: Toward a theory of rapid voluntary action. *M. Jeannerod (Ed.), Attention and Performance XIII*, 173–226.

Poulton, E.C. (1974). *Tracking skill and manual control*. Academic Press, New York.

Sandberg, S., Håkansson, C., Elmqvist, N., Tsigas, P., and Chen, F. (2006). Using 3D audio guidance to locate indoor static objects. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 50, 1581–1584.

Strachan, S., Eslambolchilar, P., Murray-Smith, R., Hughes, S., and O'Modhrain, S. (2005). gpsTunes: controlling navigation via audio feedback. In *MobileHCI*, 275–278. ACM.

Susnik, R., Sodnik, J., and Tomazic, S. (2003). Sound source choice in HRTF acoustic imaging. *HCI Int. adj. proc.*, 101–2.

Williamson, J., Robinson, S., Stewart, C., Murray-Smith, R., Jones, M., and Brewster, S. (2010). Social gravity: a virtual elastic tether for casual, privacy-preserving pedestrian rendezvous. In *Proc. ACM SIGCHI*, 1485–1494.