Attracktion: Field Evaluation of Multi-Track Audio as Unobtrusive Cues for Pedestrian Navigation

Florian Heller
Jelco Adamczyk
Kris Luyten
florian.heller@uhasselt.be
me@jelcoadamczyk.eu
kris.luyten@uhasselt.be
Hasselt University - tUL - Flanders Make
Expertise Centre for Digital Media
Diepenbeek, Belgium

ABSTRACT

Listening to music while being on the move is common in our headphone society. However, if we want assistance in navigation from our smartphone, existing approaches either demand exclusive playback through the headphones or impact the listening experience of the music. We present a field evaluation of Attracktion, a spatial audio navigation system that leverages the access to single stems in a multi-track recording to minimize the impact on the listening experience. We compared Attracktion against current turn-by-turn navigation instructions in a field-study with 22 users and found that users perceived acoustic overlays with additional navigation information to have no impact on the listening experience. In terms of path efficiency, errors, and mental workload, Attracktion is on par with spoken turn-by-turn navigation instructions, and users liked it for the aspect of serendipity.

CCS CONCEPTS

• **Human-centered computing → Mixed / augmented reality**; *Sound-based input / output*; *Auditory feedback*; Field studies; Ubiquitous and mobile computing.

KEYWORDS

Spatial Audio; Music; Navigation; Pedestrian.

ACM Reference Format:
Florian Heller, Jelco Adamczyk, and Kris Luyten. 2020. Attracktion: Field Evaluation of Multi-Track Audio as Unobtrusive Cues for Pedestrian Navigation. In *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '20), October 5–8, 2020, Oldenburg, Germany*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3379503.3403546

MobileHCI '20, October 5–8, 2020, Oldenburg, Germany

ACM ISBN 978-1-4503-7516-0/20/10…$15.00
https://doi.org/10.1145/3379503.3403546

1 INTRODUCTION

Spatial audio rendering applies special filters to recorded sounds to create the impression that these originate from a specific location around the listener. Together with position and orientation sensors, this technology is used to create audio augmented reality systems, where the virtual sound sources appear to come from a location situated in physical space. One example for which spatialized audio is especially suited are pedestrian navigation systems. While early systems were subject to the limited capabilities of the then-available hardware and required the user to carry additional components for localized audio, today's smartphones provide the required sensors and processing power for complex auditory scenes.

Auditory navigation systems, however, conflict with the fact that many people already listen to music while on the move [6]. To be able to play the instructions or orientation cues, either exclusive playback is required, or the listening experience of the music is altered through overlayed playback. Recent projects investigated the use of multi-track audio recordings to reduce the impact of integrating navigation cues on the listening experience [9]. According to a lab-experiment, people can localize the voice of a singer in a spatial mix with an accuracy of about 30°, which is sufficient for pedestrian navigation [30]. Heller and Schöning [9], however, only investigated the technique for one dedicated song, and only in a lab and a web study, not in a real pedestrian context.

In this paper, we present Attracktion, a multi-track record-based pedestrian navigation system. Attracktion indicates in which direction to walk by placing either the drums, the voice of the singer, or a mix of vocals and lead instruments in the direction of the next navigation waypoint while the other instruments remain static (Figure 1). We evaluated the feasibility of this approach on a 1.2km path through an urban area with a variety of music. We compared Attracktion with musical playback, with and without an additional auditory confirmation when reaching a waypoint to traditional spoken turn-by-turn instructions. Our results show that Attracktion with a minimal auditory waypoint confirmation is on par with turn-by-turn navigation in terms of path efficiency, navigation errors, and mental workload.

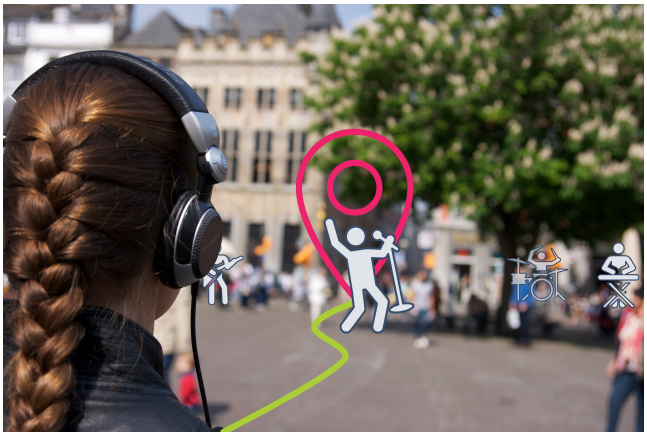

Figure 1: Attracktion indicates the direction of the next navigation waypoint by moving a single stem of a multi-track recording using spatial audio rendering. For example, the user would follow the voice of the singer.

2 RELATED WORK

Using auditory cues as means to indicate the direction for navigation dates back to AudioGPS [11]. Compared to the mainly visual (e.g., map-based or instruction-based navigation systems such as [27]) or haptic [20] pedestrian navigation systems, using audio has the advantage of leveraging our natural capability of localizing sound sources in space and thereby reduces the load on the visual (or haptic) senses [10] which are already used for the primary task of, e.g., walking or riding a bike [34].

The early AudioGPS [11] panned a beacon sound in the stereo spectrum and used different sounds to differentiate between locations in the frontal or rear hemisphere of the user. Current smartphones bring enough processing power to use spatial audio rendering [7] instead of simple stereo panning, which applies special filters to an audio signal to make it appear from a specific direction in space. This improves the speed and accuracy of the auditory localization task and the overall navigation performance [18]. The Heare App[1] uses this technology to create curated routes for blind users. It has also been used in a series of Mobile Audio Augmented Reality installations to create an environment for serendipitous discovery [2, 8, 16, 24]. In this virtual audio space, certain target locations are augmented with a sound which the user localizes and tries to reach. However, these approaches require the user to wear headphones to listen to a single beacon sound, which blocks this channel for other sources of information relevant to the user [6] (e.g., music, phone calls, audiobooks). To address this issue, several systems implemented the navigation function into a music player. ONTRACK [13] indicates the direction of the next waypoints of a navigation route by panning a song from left to right and communicates the distance to that waypoint by increasing the volume the closer you get to the target destination. GpsTunes [22] minimizes the impact on the music listening experience by inverting the distance cue and dims the audio in close proximity of waypoint.

[1]http://www.heareapp.com

For sources in the rear hemisphere, gpsTunes applied a low-pass filter to give the music a more muffled character. The approach to pan the entire track into a specific direction, however, can result in the audio being played on one ear only if the target is on the far left or right [1, 4, 13, 22, 32], which affects how music is actually perceived [15]. Shifting the phase between the two signals of a stereo recording to indicate the direction of the next waypoint [32] is interesting because it picks up one of the physical cues humans use to locate sound in space [3] and is applicable to any kind of music. However, the localization accuracy of this technique and its impact on the listening experience still need to be investigated.

3 ATTRACKTION

Attracktion aims at presenting navigation cues with a minimal impact on the listening experience by taking advantage of multi-track recordings. A multi-track recording allows access to the individual stems of a song, thus does not contain the combined stems in a fixed format. In contrast to traditional stereo recordings where we can only shift the playback from one side to another, multi-track recordings allow us to move, for example, the voice of the singer around the user's head using spatial audio rendering, while all remaining instruments remain static relative to the user's head. With this approach, we avoid the need for an explicit beacon sound being played exclusively or as an overlay, and instead, merge two functionalities into a single enjoyable listening experience. The result is a mixture of exocentric audio-augmented reality, where the source is fixed to an absolute position in the physical space, in our case, the position of the next waypoint, and an egocentric representation for the remaining instruments.

In a lab-setting, the vocals have been shown to be the preferred orientation cue [9]. However, using the singer's voice is not possible for all songs, nor for the whole duration of a song. While modern pop-songs put the emphasis on the singer, older pop-songs often have a longer vocal break to make room, e.g., for an instrumental solo. In our experiment, we also considered drums or a mix of vocals and lead instruments as cue type to overcome this problem.

4 IMPLEMENTATION

For our experiment, we implemented Attracktion as an Android application. Spatial audio rendering was performed using the Resonance Audio framework [5]. We set the renderer to use its high-quality rendering mode, which enables spatialization through a generalized head-related transfer function (HRTF) [25]. An HRTF describes the modification of sound on its path to the ear, depending on its origin relative to the listener's head. This function is individual to every listener, making it unfeasible to use the user's own HRTF for rendering on a larger scale. Generalized or non-individual HRTFs are more generic but apply to a large number of listeners at the expense of a slight loss of realism [29]. To assess the spatial resolution of the audio rendering framework, we ran a small experiment with four people. We played a series of sonar pulses, moving around in the azimuth plane. For each pulse, participants were asked to indicate whether the sound shifted left, right, or did not move at all. For an angle of 6° and above, all participants localized the shift correctly, which is consistent with the human capability to localize sounds as reported in the literature [28]. Regarding

Figure 2: The headset we used for head-orientation measurement. From left to right: Adafruit HUZZAH32 Microcontroller board, Bosch BNO055 9-DoF IMU, Battery.

possible front-back confusions which can occur when using generalized HRTFs [29], we argue that the constant movement while walking will cause slight noise in the sensor readings, which simulates the small head movements humans do to eliminate front-back confusions with real, physical sound sources. We used the same headphones as in the larger experiment. In this experiment, we did not include elevation into the rendering [7], meaning that all sources appear in the horizontal plane running through the user's ears.

Our software ran on a Motorola Moto Z Play smartphone running Android 8.0.0. To track the location, we used the smartphone's onboard GPS unit. Head-orientation was measured by placing a Bluetooth enabled microcontroller board (Adafruit HUZZAH32 with an Espressif ESP32 MCU) running the MotionHeadset firmware[2] and a 9-DOF inertial measurement unit (IMU) (Bosch BNO055) onto the head strap of regular headphones (Sony MDR-ZX770BN) (Figure2). The fused heading of the IMU is reported to the rendering engine with a refresh rate of 100Hz through a Bluetooth LE connection, while the headphones were wired to the smartphone to minimize audio latency.

The capture radius was set to 5 meters. While prior research indicates that a human-sized capture radius (about 2 meters) is optimal for efficient navigation [19, 26], larger radii tend to promote faster navigation [31]. The larger capture radius should result in faster progress around the waypoints as they do not need to be reached exactly. When reaching a waypoint, a confirmation consisting of a triple sonar pulse with a delay of 500ms could be played in the direction of the new waypoint. This sound has ideal characteristics for localization [23].

The app offered a large "help me" button that could be used to get automated guidance by reveling the direction of the next waypoint using an arrow for a few seconds. The experimenter followed the participants to ensure their safety and to intervene in the case of navigation errors. If a participant strayed from the route, the observer would get their attention and tell them to press the "help" button provided on the smartphone. Using the help button, however, is considered as a navigation error.

[2]https://github.com/florianheller/MotionHeadset

5 EVALUATION

To evaluate Attracktion, we performed an in-the-wild experiment where participants had to walk a 1.2km path through an urban area using our navigation system while holding the phone in their hands. The path consisted of 14 waypoints, including the starting point. Waypoints were mostly located at basic and easy intersections with two or three clearly distinguishable choices even taking into account potential sensor inaccuracies. There are two curved road sections and one complicated intersection consisting of five possible options, some of which are at small angles. We added an additional waypoint for the curved road on the left of Figure3 as a straight connection to the next waypoint would have led to a very small difference between two possible road choices at the complicated intersection. The participants were expected to cross the road on three separate occasions, two of which were at a pedestrian crossing. To minimize the effect of possible front-back confusions caused by the spatial audio rendering, we designed the route such that the following waypoint always was in the frontal hemisphere or on the far ear side of the user. As can be seen with the curved road on the left and the parking lot on the bottom of Figure3, the path did not always precisely follow the road, meaning that participants were left with some autonomy in choosing their paths. To mitigate learning effects, participants could practice by walking along a simple route consisting of a square with four waypoints on each corner of a plaza.

For the Attracktion conditions, we prepared eight different popular pop songs based on publicly available remix stems. The playback order of the songs was randomized at the start of each trial, and playback ended as the trial was concluded (within-subjects), so participants did not necessarily hear every song. For each song, we chose one or more stems to act as orientation cue. For more modern pop songs where the emphasis is usually mostly on the singer, we used the vocal track throughout the entire song. For (older) pop or rock songs in which the vocals would stop in favor of an instrumental riff or solo (for example, 'Enjoy the Silence' by Depeche Mode or 'Enter Sandman' by Metallica), we manually decided what the current "focus" track was and used this as orientation cue. Finally, for songs in which the drum score plays a central and fundamental role (for example, 'Take on Me' by A-ha or 'Smooth Criminal' by Michael Jackson) we opted for an exclusive drums orientation cue. When entering the capture radius around a waypoint, the auditory notification overlay was enabled for half of the participants that used the Attracktion playback.

We implemented a baseline condition with spoken turn-by-turn instructions overlaid over the music playback. While the instructions played, the music volume was lowered. In the *baseline* condition, the notification of reaching a waypoint was given implicitly when the navigation instructions were played.

In summary, we have one between-subject factor CONDITION, which could be *baseline*, *Attracktion with notification*, and *Attracktion without notification*. In the *baseline* condition, the waypoint confirmation was given implicitly by playing the next navigation instruction. In the two Attracktion conditions, we have the within-subjects factor CUE TYPE, which was either *drums*, *voice*, *mix*, and was randomized at the start of the experiment.

Figure 3: Logged paths of our participants. The intended route is plotted in blue. Participant's actual paths are plotted in red (Attracktion with notification), green (Attracktion without notification), and yellow (text-to-speech). Image data: Google.

We logged the user's GPS position to calculate the path efficiency, i.e., the ratio between the length of the actual path and the ideal path consisting of straight connections between the waypoints. We recorded the orientation of the head and the smartphone to derive the amount of head-turns a user made as an indicator for how difficult it is to localize a sound source. As users tend to keep the device in front of their body [8], we calculated the difference between the two points of measurement and summarized this as root mean square (RMS) head-yaw deviation. We also logged the current walking speed as reported by the Android Location Service, and the number of navigation errors.

The experiment closed with a NASA TLX questionnaire along with a comparison of the perceived listening experience to the regular experience, followed by asking the participants for some open feedback.

Overall, we recruited 25 participants among fellow students, friends, and family to participate in the experiment. When asked for a known hearing disorder or problems with spatial hearing, two mentioned having a tinnitus, and two mentioned having a minor presbycusis, but none mentioned having known problems with spatial hearing. Two participants mentioned having previous experience from consumer surround sound systems, and one of these also mentioned having experience with spatial audio from VR headsets. Three users could not complete their trial. One participant was unable to notice any change in the presented audio and two because of technical failures during the experiment. The results of the remaining 22 participants (14 female, 8 male, average age 36 years) that finished the experiment were used for statistical analysis. Participants were assigned to the conditions one after the other to keep the group sizes balanced. Seven participants completed the Attracktion condition with notifications, seven the Attracktion condition without notifications, and eight participants completed the experiment in the baseline condition.

6 RESULTS

The following section contains a detailed discussion of the various effects and results. For a better overview, the results are summarized in Table 1.

Path Efficiency. We calculated the path efficiency (PE) of each trial by comparing the walked distance to the length if the path connecting each intermediate waypoint through straight lines. In theory, if participants walk exactly from waypoint to waypoint without following curved roads or avoiding road blocks, their PE would be 100% and by cutting some corners, they could even achieve a path efficiency above 100%. A Shapiro-Wilk W test for normality indicates that the distribution of path efficiency can be considered normal ($p = .1178$) but a Levene's test for homogeneity of variances indicates that the data seem to violate homoscedasticity ($F_{2,19} = 4.46, p = .0258$). According to a Kruskal-Wallis test, there is a significant effect of CONDITION on path efficiency ($\chi^2(2, N = 22) = 10.5716, p = .0051$). A *post hoc* pairwise comparison using a Tukey HSD test showed the path efficiency to be significantly higher in the *baseline* ($M = 91.6$ $SD = 1.7$) condition than with *Attracktion without notifications* ($M = 83.5$ $SD = 4.4$ $p = .0016$), but not significantly higher than *Attracktion with notifications* ($M = 87.5$ $SD = 4.8$ $p = .1171$).

The path efficiency was highest with the *vocal* ($M = 77$ $SD = 15.5$) and *mix* ($M = 76.4$ $SD = 12.8$) cues while the *drum* cue lead to a slightly lower efficiency ($M = 69.6$ $SD = 21.6$). This is caused by two users who had particular problems with one specific song, although the average PE was comparable across all songs. If we exclude these two outliers, the PE for the *drum* cue is similar to the other ones ($M = 74.8$ $SD = 12.6$). Even taking the outliers into account, the CUE TYPE did not have a statistically significant effect on path efficiency ($F_{2,16.31} = 2.8464, p = .087$).

Errors. During the navigation, participants sometimes missed a turn or went into the wrong street or direction. While we notified them of their mistake and helped them resume navigation, we logged the number of errors. The most errors occurred in the *Attracktion* condition *without waypoint notifications* with an average of $M = 4.6$ ($SD = 1.3$, $min = 3$, $max = 7$) errors. With waypoint notifications enabled, the number of errors $M = 2.3$ ($SD = 1.6$, $min = 0$, $max = 5$) dropped to the same level as in the *baseline* condition $M = 2.1$ ($SD = 1.2$, $min = 1$, $max = 4$). The distribution of errors can be considered normal (Shapiro-Wilk test, $p = .3279$) and a Levene's test found no violation of homoscedasticity ($F_{2,19} = .21, p = .8123$). There is a significant effect of CONDITION on the number of errors ($F_{2,19} = 7.105, p = .005$). A *post hoc* pairwise comparison using a Tukey HSD test showed that there is a significant difference only between *Attracktion without notification* and the other conditions ($p < .0153$).

While using the drum cue, participants made significantly more errors (M = .83 SD = .82, $F_{2,73.02} = 3.8546, p = .0256$) than with vocals ($M$ = .36 SD = .6) and mix (M = .55 SD = .5).

Head Turns. The head-yaw data shows a lognormal distribution (Kolmogorov's D test, $p = .15$). A repeated-measures ANOVA with CUE TYPE and NOTIFICATION as fixed factors and user as a random factor showed a significant effect of CUE TYPE on log-transformed RMS head-yaw deviation ($F_{2,18.93} = 6.93, p = .0055$) We performed a *post hoc* pairwise comparison using a Tukey HSD test. While the results for Voice (M = 50.8 SD = 7.6) and Mix (M = 50.9 SD = 10) are very similar (n.s., $p = .784$), the drum cue, although actually well suited for localization tasks for its strong onsets and transient nature, caused significantly more head turns (M = 54.9 SD = 12) than the other two ($p < .0121$). Turning notifications on or off in the *Attracktion* conditions had no significant influence on the amount of head-turns ($F_{1,10.66} = .59, p = .46$).

Questionnaire. In a post-experiment questionnaire we asked the participants to rate their perceived listening experience on a scale of 1 to 5 compared to how they would have experienced the same music using their own devices on their own accord. While a Wilcoxon Rank Sums Test found no significant difference between the overall *Attracktion* condition (Mdn = 4, IQR = 1.25) and the *baseline* condition (Mdn = 5, IQR = .75), participants who received *no notification* did rank their listening experience significantly lower (Mdn = 4, IQR = 1) than the participants with *notification* (Mdn = 5, IQR = 1) ($p = .353$) and the *baseline* condition ($p = .0083$).

The workload we measured using the NASA TLX questionnaire was higher in the *Attracktion* condition *without notifications* (M = 35.7 SD = 4.75) than *with notifications* enabled (M = 28.1 SD = 9.97) and lowest in the *baseline* condition (M = 17.7 SD = 7.29). An ANOVA showed a significant effect of CONDITION on workload ($F_{2,1229.18} = 10.58, p = 0.0008$). According to a *post hoc* Tukey HSD test, the workload is significantly lower in the *baseline* condition than in the two *Attracktion* conditions ($p < .0415$). If we look at the individual scales of the questionnaire, we see significant differences in the ratings for mental load, effort, and frustration. The mental load of using the *baseline* implementation (M = 31.25 SD = 30) is significantly lower than in the *Attracktion without notifications* (M = 73.6 SD = 7.5, $p = .0048$) condition, but not than in the *Attracktion* condition *with notification* (M = 60 SD = 22.4, n.s.). Effort was perceived significantly lower in the *baseline* condition (M = 13.75 SD = 6.4) than in both *Attracktion* conditions (No notifications M = 33.6 SD = 14.4, with notifications M = 38.6 SD = 16; $p < .0185$). Frustration was rated significantly lower in the *baseline* condition (M = 5.6 SD = 1.8) compared to *Attracktion without notifications* (M = 30 SD = 18 $p = .009$), but not significantly lower than with *Attracktion with notifications* (M = 18.6 SD = 17.3, n.s.).

When asked whether they would like to have distance information encoded in the audio stream, 4 out of 14 participants agreed, for example, to estimate if they are going into the right direction.

Table 1: Summary of our experimental results based on 22 participants. An asterisk* denotes a statistically significant difference to the other values.

	Attracktion without notifications	**Attracktion** with notifications	**Baseline**
Path efficiency	$M = 83.5^*$ $SD = 4.4$	$M = 87.5$ $SD = 4.8$	$M = 91.6$ $SD = 1.7$
Errors	$M = 4.6^*$ $SD = 1.3$	$M = 2.3$ $SD = 1.6$	$M = 2.1$ $SD = 1.2$
Listening experience	$Mdn = 4^*$ $IQR = 1$	$Mdn = 5$ $IQR = 1$	$Mdn = 5$ $IQR = .75$
Workload	$M = 35.7$ $SD = 4.75$	$M = 28.1$ $SD = 9.97$	$M = 17.7^*$ $SD = 7.29$

6.1 Discussion

From observing the trial and discussing it with the participants afterward, we noticed that switching beacon tracks on a song-by-song basis was confusing for some participants. Identifying what stem they had to focus on for the navigation cues, was sometimes cumbersome. A recurring example was that participants were following the singer's voice while actually the drums were supposed to be the beacon. Even when turning their heads, when they should have noticed the vocals staying in position while the drums shifted, some participants did not catch up on this, probably due to lack of training.

Before deploying this approach as a real-world system, the manual selection of the cue-type as done for our experiment needs to be automated such that it can cover a large variety of music. Algorithms that can separate the voice of the singer [12] or drums [33] from the remaining elements of a music track produce good results. This means that these could be extracted for an automated analysis prior to playback to judge whether they are usable for navigation, e.g., if they contain long gaps.

When asked for their preferred beacon track, 12 out of 14 participants in the Attracktion conditions chose the vocals, and only four opted for the mix and two for the drums (multiple selections were possible). Five participants also explicitly mentioned in their feedback that the vocals were much easier to localize, and to a lesser extent, the mix cue because they automatically focused on those tracks to begin with. This is in line with the results from Heller & Schöning [9]. In contrast to the analytical assumption of drums being suited because of their high amount of transients and harsh onsets, users have difficulty localizing them in the musical mix. This is partially due to the fact that the drums, although a crucial part of a song, are often not the most distinct instrument in the mix.

Environmental noise is a critical factor for successfully using the system. Three participants mentioned that they had to turn up the volume because they could not hear the song at busy intersections. This might interfere with the navigation cues that are embedded in the song. In real-world implementations, users would likely be able to take a look at their smartphone's screen. Headphones with active noise cancellation could also alleviate the impact of noisy environments, as isolation is one of the reasons people listen to music while being on the move [6]. Environmental noise is,

however, also a very important source of information to create awareness of one's own environment. If we eliminate this channel, safety-relevant acoustic signals, such as those from an approaching car, might be missed. The use of bone-conductance headphones (BCH) can alleviate this problem as they do not cover the ears [17]. While there are no spatial audio rendering algorithms specifically tweaked towards this technology, the perception in the 2D plane is comparable to normal headphones [21].

The capture radius of 5m we used was too small in that it required the users to closely reach the waypoint. While we already used a larger radius than other pedestrian navigation systems [19, 26], relating it to the local environment, e.g., the width of the sidewalk might increase the navigation efficiency.

The low ratings for listening experience in the Attracktion condition without waypoint notifications are surprising, as it was designed to provide a listening experience close to the original. From our experimental data, we have to conclude that participants do not perceive overlays or turn-by-turn instructions as part of the music that is playing, and therefore not interfering with it. In contrast to turn-by-turn instructions, Attracktion provides continuous feedback on the direction to the next waypoint.

Overall, most participants indicated that they would use a system like this for their own leisure (71%), with multiple remarks on how the trial was experienced as 'fun' and 'interesting'. One participant, in particular, mentioned that she usually does not enjoy walking, but that a system like this could encourage her to walk more in a fun way. This underlines the usefulness of mobile audio-augmented reality systems for serendipitous discovery, for example, in a touristic audio guide or an audio-based game. Among the reasons that people would not use a system like this was that they do not like following a planned route, or that they just do not like walking in general.

One limitation of this work is certainly the public availability of multi-track recordings. For the case of common pop-songs, neural networks achieve good results in separating the singer's voice from the rest [12], which would be sufficient for our application. For other genres, the filter algorithms would need to be optimized towards other stems and instruments.

The route we used in our experiment did not contain hairpin turns or waypoints that lead in the opposite direction. Such a simplification might not always be possible in a real-world implementation. Nevertheless, a navigation system using our approach could mitigate the effects by never placing two consecutive waypoints in the opposite direction, but instead replace a sharp turn with a series of smaller path segments. Some segments of the path required users to navigate in more autonomously than others. For example, the straight connection between the two last waypoints in Figure3 crosses a building; thus, the participants had to deviate from the optimal path. We argue that this is not much different from the curved street on the left of Figure3. Such special cases should be investigated further, i.e., how much autonomous navigation users are willing to make while using such a system. However, none of the participants mentioned this path segment to be particularly complex or confusing.

7 SUMMARY & FUTURE WORK

In this paper, we presented a field evaluation of a mobile audio-augmented reality music player for pedestrian navigation. Without waypoint confirmation, it could not compete with spoken turn-by-turn navigation. However, with a brief acoustic confirmation of reaching a waypoint, Attracktion is on par with turn-by-turn navigation in terms of path efficiency, error rate, and mental workload. Additionally, participants enjoyed the aspect of serendipitous discovery of our system. As users did not perceive overlayed cues to influence the listening experience, adding a minimal waypoint confirmation re-assures users that they are still on the right path.

Using dynamic capture radii can increase navigation efficiency without losing the precision necessary at complex intersections. Using a large radius allows users to cut corners in more open areas, while the system can revert to small radii, e.g., if the user needs to discern between to close paths.

Covering the ears during pedestrian navigation might also shift focus away from the potentially harmful surroundings, e.g., because the user does not hear an approaching car. Current high-end earables, small individual earpieces, therefore include microphones on the outside of the casing to make them acoustically transparent if needed. As earables also provide increasing processing capabilities [14], we envision an automatic recognition of approaching sounds and automatic switching between the closed and the transparent listening mode.

While in this experiment, we used custom hardware, such sensor hardware will become more prevalent in future headsets [14].

ACKNOWLEDGMENTS

We would like to thank the participants of our study and the anonymous reviewers for their valuable feedback.
This work was funded by the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme and by Flanders Make (ICON project OperatorKnowledge).

REFERENCES

[1] Robert Albrecht, Riitta Väänänen, and Tapio Lokki. 2016. Guided by music: pedestrian and cyclist navigation with route and beacon guidance. *Personal and Ubiquitous Computing* 20, 1 (2016), 121–145. https://doi.org/10.1007/s00779-016-0906-z

[2] Anupriya Ankolekar, Thomas Sandholm, and Louis Yu. 2013. Play It by Ear: A Case for Serendipitous Discovery of Places with Musicons. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI '13)*. Association for Computing Machinery, New York, NY, USA, 2959–2968. https://doi.org/10.1145/2470654.2481411

[3] Jens Blauert. 1996. *Spatial Hearing: Psychophysics of Human Sound Localization* (2 ed.). MIT Press, Cambridge, MA.

[4] Richard Etter and Marcus Specht. 2005. Melodious walkabout: Implicit navigation with contextualized personal audio contents. In *Adjunct Proceedings of the Third International Conference on Pervasive Computing* (Munich, Germany) *(Pervasive '05 Adjunct)*. 43–49.

[5] Marcin Gorzel, Andrew Allen, Ian Kelly, Alper Gungormusler, Julius Kammerl, Hengchin Yeh, and Francis Boland. 2019. Efficient Encoding and Decoding of Binaural Sound with Resonance Audio. https://resonance-audio.github.io/resonance-audio/.

[6] Gabriel Haas, Evgeny Stemasov, and Enrico Rukzio. 2018. Can't You Hear Me? Investigating Personal Soundscape Curation. In *Proceedings of the 17th International Conference on Mobile and Ubiquitous Multimedia* (Cairo, Egypt) *(MUM 2018)*. Association for Computing Machinery, New York, NY, USA, 59–69. https://doi.org/10.1145/3282894.3282897

[7] Florian Heller, Jayan Jevanesan, Pascal Dietrich, and Jan Borchers. 2016. Where Are We? Evaluating the Current Rendering Fidelity of Mobile Audio Augmented Reality Systems. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services* (Florence, Italy) *(MobileHCI*

'16). Association for Computing Machinery, New York, NY, USA, 278–282. https://doi.org/10.1145/2935334.2935365

[8] Florian Heller, Aaron Krämer, and Jan Borchers. 2014. Simplifying Orientation Measurement for Mobile Audio Augmented Reality Applications. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 615–624. https://doi.org/10.1145/2556288.2557021

[9] Florian Heller and Johannes Schöning. 2018. NavigaTone: Seamlessly Embedding Navigation Cues in Mobile Music Listening. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3173574.3174211

[10] Eve Hoggan, Andrew Crossan, Stephen A. Brewster, and Topi Kaaresoja. 2009. Audio or Tactile Feedback: Which Modality When?. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, MA, USA) *(CHI '09)*. Association for Computing Machinery, New York, NY, USA, 2253–2256. https://doi.org/10.1145/1518701.1519045

[11] 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. https://doi.org/10.1007/s007790200025

[12] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde. 2017. Singing voice separation with deep U-Net convolutional networks. (October 2017). https://openaccess.city.ac.uk/id/eprint/19289/

[13] Matt Jones, Steve Jones, Gareth Bradley, Nigel Warren, David Bainbridge, and Geoff Holmes. 2008. ONTRACK: Dynamically adapting music playback to support navigation. *Personal and Ubiquitous Computing* 12, 7 (2008), 513–525. https://doi.org/10.1007/s00779-007-0155-2

[14] Fahim Kawsar, Chulhong Min, Akhil Mathur, and Alessandro Montanari. 2018. Earables for Personal-Scale Behavior Analytics. *IEEE Pervasive Computing* 17, 3 (2018), 83–89. https://doi.org/10.1109/MPRV.2018.03367740

[15] Doreen Kimura. 1964. Left-right differences in the perception of melodies. *Quarterly Journal of Experimental Psychology* 16, 4 (1964), 355–358. https://doi.org/10.1080/17470216408416391

[16] Andreas Komninos, Peter Barrie, Vassilios Stefanis, and Athanasios Plessas. 2012. Urban Exploration Using Audio Scents. In *Proceedings of the 14th International Conference on Human-Computer Interaction with Mobile Devices and Services* (San Francisco, California, USA) *(MobileHCI '12)*. Association for Computing Machinery, New York, NY, USA, 349–358. https://doi.org/10.1145/2371574.2371629

[17] Andreas Kratky. 2019. Walking in the Head: Methods of Sonic Augmented Reality Navigation. In *Human-Computer Interaction. Recognition and Interaction Technologies*. Springer, Berlin Heidelberg, 469–483. https://doi.org/10.1007/978-3-030-22643-5_37

[18] Camilla H. Larsen, David S. Lauritsen, Jacob J. Larsen, Marc Pilgaard, and Jacob B. Madsen. 2013. Differences in Human Audio Localization Performance between a HRTF- and a Non-HRTF Audio System. In *Proceedings of the 8th Audio Mostly Conference* (Piteå, Sweden) *(AM '13)*. Association for Computing Machinery, New York, NY, USA, 8. https://doi.org/10.1145/2544114.2544118

[19] Nicholas Mariette. 2010. Navigation Performance Effects of Render Method and Head-Turn Latency in Mobile Audio Augmented Reality. In *Auditory Display (ICAD '09)*. Springer, Berlin, Heidelberg, 239–265. https://doi.org/10.1007/978-3-642-12439-6_13

[20] Enrico Rukzio, Michael Müller, and Robert Hardy. 2009. Design, Implementation and Evaluation of a Novel Public Display for Pedestrian Navigation: The Rotating Compass. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, MA, USA) *(CHI '09)*. Association for Computing Machinery, New York, NY, USA, 113–122. https://doi.org/10.1145/1518701.1518722

[21] Raymond M. Stanley and Bruce N. Walker. 2006. Lateralization of Sounds Using Bone-Conduction Headsets. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 50, 16 (2006), 1571–1575. https://doi.org/10.1177/154193120605001612

[22] Steven Strachan, Parisa Eslambolchilar, Roderick Murray-Smith, Stephen Hughes, and Sile O'Modhrain. 2005. GpsTunes: Controlling Navigation via Audio Feedback. In *Proceedings of the 7th International Conference on Human Computer Interaction with Mobile Devices & Services* (Salzburg, Austria) *(MobileHCI '05)*. Association for Computing Machinery, New York, NY, USA, 275–278. https://doi.org/10.1145/1085777.1085831

[23] Tuyen V. Tran, Tomasz Letowski, and Kim S. Abouchacra. 2000. Evaluation of acoustic beacon characteristics for navigation tasks. *Ergonomics* 43, 6 (2000), 807–827. https://doi.org/10.1080/001401300404760 PMID: 10902889.

[24] Yolanda Vazquez-Alvarez, Ian Oakley, and Stephen A. Brewster. 2012. Auditory display design for exploration in mobile audio-augmented reality. *Personal and Ubiquitous Computing* 16, 8 (2012), 987–999. https://doi.org/10.1007/s00779-011-0459-0

[25] Michael Vorländer. 2007. *Auralization: Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality* (1st ed.). Springer.

[26] Bruce N. Walker and Jeffrey Lindsay. 2006. Navigation Performance With a Virtual Auditory Display: Effects of Beacon Sound, Capture Radius, and Practice. *Human Factors* 48, 2 (2006), 265–278. https://doi.org/10.1518/001872006777724507

[27] Dirk Wenig, Johannes Schöning, Brent Hecht, and Rainer Malaka. 2015. StripeMaps: Improving Map-Based Pedestrian Navigation for Smartwatches. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services* (Copenhagen, Denmark) *(MobileHCI '15)*. Association for Computing Machinery, New York, NY, USA, 52–62. https://doi.org/10.1145/2785830.2785862

[28] Elizabeth M. Wenzel, Marianne Arruda, Doris J. Kistler, and Frederic L. Wightman. 1993. Localization using nonindividualized head-related transfer functions. *The Journal of the Acoustical Society of America* 94, 1 (1993), 111–123. https://doi.org/10.1121/1.407089

[29] Elizabeth M. Wenzel, Frederic L. Wightman, and Doris J. Kistler. 1991. Localization with Non-Individualized Virtual Acoustic Display Cues. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New Orleans, Louisiana, USA) *(CHI '91)*. Association for Computing Machinery, New York, NY, USA, 351–359. https://doi.org/10.1145/108844.108941

[30] John Williamson, Simon Robinson, Craig Stewart, Roderick Murray-Smith, Matt Jones, and Stephen Brewster. 2010. Social Gravity: A Virtual Elastic Tether for Casual, Privacy-Preserving Pedestrian Rendezvous. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. Association for Computing Machinery, New York, NY, USA, 1485–1494. https://doi.org/10.1145/1753326.1753548

[31] Jeff Wilson, Bruce N. Walker, Jeffrey Lindsay, Craig Cambias, and Frank Dellaert. 2007. SWAN: System for Wearable Audio Navigation. In *Proceedings of the 11th IEEE International Symposium on Wearable Computers (ISWC '07)*. IEEE, 91–98. https://doi.org/10.1109/ISWC.2007.4373786

[32] Shingo Yamano, Takamitsu Hamajo, Shunsuke Takahashi, and Keita Higuchi. 2012. EyeSound: Single-Modal Mobile Navigation Using Directionally Annotated Music. In *Proceedings of the 3rd Augmented Human International Conference* (Megève, France) *(AH '12)*. Association for Computing Machinery, New York, NY, USA, 4. https://doi.org/10.1145/2160125.2160147

[33] Aymeric. Zils, François Pachet, Olivier Delerue, and Fabien Gouyon. 2002. Automatic extraction of drum tracks from polyphonic music signals. In *Proceedings of the Second International Conference on Web Delivering of Music (WEDELMUSIC 2002)*. IEEE, 179–183. https://doi.org/10.1109/WDM.2002.1176209

[34] Matthijs Zwinderman, Tanya Zavialova, Daniel Tetteroo, and Paul Lehouck. 2011. Oh Music, Where Art Thou?. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services* (Stockholm, Sweden) *(MobileHCI '11)*. Association for Computing Machinery, New York, NY, USA, 533–538. https://doi.org/10.1145/2037373.2037456