NavigaTone: Seamlessly Embedding Navigation Cues in Mobile Music Listening

Florian Heller
Hasselt University - tUL - imec
3590 Diepenbeek, Belgium
florian.heller@uhasselt.be

Johannes Schöning
University of Bremen
Bremen, Germany
schoening@uni-bremen.de

ABSTRACT

As humans, we have the natural capability of localizing the origin of sounds. Spatial audio rendering leverages this skill by applying special filters to recorded audio to create the impression that a sound emanates from a certain position in the physical space. A main application for spatial audio on mobile devices is to provide non-visual navigation cues. Current systems require users to either listen to artificial beacon sounds, or the entire audio source (e.g., a song) is re-positioned in space, which impacts the listening experience. We present NavigaTone, a system that takes advantage of multi-track recordings and provides directional cues by moving a single track in the auditory space. While minimizing the impact of the navigation component on the listening experience, a user study showed that participants could localize sources as good as with stereo panning while the listening experience was rated to be closer to common music listening.

ACM Classification Keywords

H.5.5. Information Interfaces and Presentation (e.g. HCI): Sound and Music Computing; Systems

Author Keywords

Virtual Audio Spaces; Spatial Audio; Mobile Devices; Audio Augmented Reality; Navigation.

INTRODUCTION

Portable music players are popular since they first appeared over 30 years ago. As part of every smartphone they are ubiquitous companions of our mobile lifestyles. While in the early days, wearing headphones in public was rather unusual, it has become a common sight and headphones are, more than ever, a fashion accessory [19]. The rich built-in sensors of current smartphones and the omnipresence of headphones allow us to think of new audio-based applications. Using auditory beacons as navigational aids in mobile audio augmented reality systems has been shown to be a powerful and unobtrusive concept for pedestrian navigation. A large body of related work exists that focusses on how to use spatial audio for pedestrian navigation. These either represent the navigation target using a dedicated beacon sound [10, 18, 25] which limits the use of headphones to that single purpose, or by placing the source of a music track at the target location [2, 21, 27] which impacts the listening experience. Alternatively, orientation cues could be provided using a bone conductance headphone [14, 26] or an acoustically transparent augmented reality audio headset [22], both of which do not block the perception of the surrounding soundscape. However, these do not solve the issue of blending navigation cues into the music the user is listening to.

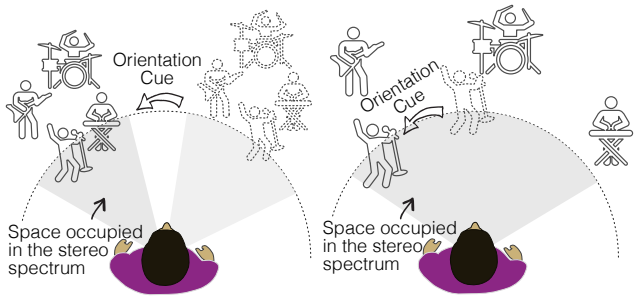

Figure 1. NavigaTone leverages the possibilities of multi-track recordings for spatial audio navigation. Instead of moving around the entire track in the stereo spectrum (left), NavigaTone only moves the voice of the singer or another instrument around the users' head.

In this note we present NavigaTone, that overcomes this problem by leveraging the potential of multi-track recordings. NavigaTone integrates the needed navigational cues into the regular stream of music in an unobtrusive way. Instead of moving the entire track around in the stereo panorama, we only move a single voice, instrument, or instrument group (cf. Figure 1). This is possible with the recent appearance of commercially available multi-track recordings, e.g., in Native Instruments' STEM format [7]. This allows us to balance between the impact of the navigation cue on the overall perception of the audio track and the ability to localize the cue, with the goal to minimize the effect on perception while still providing a good sense of orientation. We compared the aesthetic appearance and localizability of moving single sources of such a multitrack recording in space against overlaying the song with additional beacon sounds. Users clearly preferred the moving voice of the singer over an overlaid beacon sound. To further analyze the localization precision and perception of such an orientation cue, we performed a controlled user study to compare our NavigaTone system against the baseline of moving the entire track in the stereo spectrum. Our participants reported that

CHI 2018, April 21–26, 2018, Montreal, QC, Canada

ISBN 978-1-4503-5620-6/18/04…$15.00

DOI: https://doi.org/10.1145/3173574.3174211

NavigaTone felt significantly more natural and intuitive compared to the baseline, while being accurate in differentiating the origin of a sound at 30° resolution, which is sufficient for a pedestrian navigation scenario. In terms of other performance variables, the more obtrusive baseline of moving the entire track in the stereo spectrum performed slightly better, but was perceived to negatively impact the listening experience.

RELATED WORK

AudioGPS [10] was the first system using auditory cues for navigation by indicating the direction of the target. To differentiate between sources in front or in the back of the user, it used a harpsichord and a trombone timbre as navigational beacons. Instead of using different beacon sounds, spatial audio rendering applies special filters to an audio signal to make it appear from a certain position in space. This can be used to augment a certain target location with a sound which the user localizes and tries to reach [2, 25]. However, as those approaches require the user to wear headphones, listening to a single beacon sound blocks this channel for other sources of information (e.g., music, phone calls, audio books). To overcome this limitation, several systems integrated the navigation function into mobile music players. ONTRACK [12] augments the waypoints of a navigation route with a song that is panned from left to right, and that becomes louder the closer you get to it. GpsTunes [21] inverses the distance cues and dims the audio in close proximity of the target destination in order to minimize the impact on the music listening experience. For sources in the back of the head, gpsTunes applied a low-pass filter to give the music a more muffled character. The approach to pan the entire track into a certain direction, however, can result in the audio being played on one ear only if the target is on the far left or right [1, 8, 12, 21, 31], which affects how music is actually perceived [13].

The main advantage of this technology, compared to the mainly visual (e.g., map-based or instruction-based navigation systems such as [28]) or haptic [20] pedestrian navigation systems, is that it leverages our natural capability of localizing sound sources in space and thereby reduces the load on the visual (or haptic) senses [9] which are already used for the primary task of, e.g., walking or riding a bike [32].

ORIENTATION QUEUES

To be able to move a certain instrument or the voice of a singer around in the stereo spectrum, we need to separate it from the rest. To separate the voice of a pop-song from the rest, the algorithm presented in [11] produces very good results. The separation into single instruments stems, however, remains more complicated and can result in artifacts and noise. In contrast to that, multi-track recordings provide these separate stems for the different instruments, which means that you can move, e.g., the hi-hat independently of the lead guitar. We take advantage of such recordings to reduce the impact of the navigation component on the music listening experience. Instead of moving all sources around, or even worse, cut off certain instruments by using the balance control, we only move a single source, e.g., the voice of the singer (cf. Figure 1). In order to achieve a good localizability of the orientation cue, the choice of the beacon sound is very important. Transient sounds, i.e., sounds with a short duration, high amplitude onset, are best to localize by human listeners [4]. Furthermore, the larger the frequency range of the sound, the more information it can carry that can be used for localization. Most of the lab studies on the localizability of sound, therefore, include bursts of white noise as beacon sounds as they incorporate all of the aforementioned qualities. Such synthetic sounds, however, while being technically optimal, are not necessarily a pleasing listening experience. As a natural signal, human speech covers a large spectrum and contains repeated transient elements, making it a suitable cue without exhibiting the problem of repetition as faced by synthetic beacons sounds [23]. Furthermore, our auditory perception is optimized to localize and identify human speakers, even in complex auditory environments [3, 17]. Alternatively, the drums in a song are suitable as localization cue because these are strong transient signals in an inherently repetitive pattern.

ORIENTATION SURVEY

To find out whether there is an opportunity to integrate our envisioned navigation cues into peoples' everyday listening habits, we conducted an online survey. The first section of our online survey was targeted at understanding the listening behavior of our potential users, and was completed by 21 participants. A majority listens to music at least sometimes (3), very often (10), or always (3) while being on the move, and only very few never (1) or rarely (4) do so. Nearly all use headphones or earphones to listen to their music, and all but one listen to the music with both ears. Music stored on the device (13) and music from a streaming service like Spotify (13) are the preferred sources of content. Audio books (7) or speech podcasts (6) are popular as well, whereas only one mentioned to listen to musical podcasts. In most cases (14), the phone is in a pocket of their trousers. Only few keep the phone in their hand (2), while the rest places it in some other pocket or bag, depending on the situation.

Balancing Qualities

In the second part of the survey, we further evaluated the use of different types of cue sounds, with the goal of finding the right balance between an aesthetically pleasing presentation and a good localizability. We used an excerpt of Carlos Gonzalez's 'A Place For Us', a vocal pop-song available as multitrack recording[1] as base for this experiment. We used four different orientation cues: two overlays and two integrated ones. Overlay cues certainly affect the listening experience, nevertheless, we used them as their potential better localizability could compensate for the degraded aesthetic perception. A majority of the lab experiments on source localization have been performed with noise bursts as they provide the largest frequency range. Therefore, our first overlay consists of pink noise bursts at $1.7Hz$ repetition rate ($\frac{1}{4}$ bar at 102bpm) and a duty cycle of 50%. Our second overlay consists of an 880Hz pure sine wave at $0.85Hz$ repetition rate ($\frac{1}{2}$ bar at 102bpm). The overlays were beat-synchronized with the rest of the music. As integrated cues, we used the lead voice and the snare drums. All sources were mixed in the KLANG:app for iOS.

[1] cambridge-mt.com/ms-mtk.htm

The audio files and the mixing configuration are available online[2].

Study

We evaluated the perception of these directional cues both in a static and a dynamic use, meaning the sources staying at a fixed position or moving from one point to another, respectively. We created samples with a fixed orientation cue at the following angles in the frontal hemisphere: far left (-90°), -45°, center (0°), 45°, far right (90°). The dynamic cues moved by 90° from left to center, from center to left, from right to center, from center to right, and from 45° right to 45° left. Participants had to select the correct position or the correct movement range. The static and dynamic conditions were counterbalanced and the order of the samples randomized. In total, n=13 participants (3 female, average age 30) completed this part of the survey.

Results

As hypothesized, in terms of recognition rate, the dynamic cues (M=.82, SD=0.14) outperformed the static ones (M = .53, SD = .1) by far (t(24))=-5.66, p<.0001). Both in the static and the dynamic cases, the voice cue achieved the highest recognition rate (dynamic: M = .92, SD = .19, static: M = 0.66, SD = .22), followed by the overlaid noise (dynamic: M = .92, SD = .19, static: M = 0.57, SD = .23) and beep (dynamic: M = .86, SD = .24, static: M = 0.55, SD = .23) while the snare drum achieved significantly lower rates in the dynamic case (M = .55, SD = .36, $p < 0.05$, static: M = 0.35, SD = .23, *n.s*). We also asked the participants to rate three aspects of the orientation cue on a five-point Likert scale. When asked how well they could localize the orientation cue, the voice (Mdn = 4, IQR = 2.5) and noise (Mdn = 4, IQR = 2) cues achieve the best ratings in both static and dynamic cases while the snare drum (Mdn = 2, IQR = 1), again, got significantly ($p < 0.002$) lower ratings. Participants also perceived the concentration required to localize the source to be equally low for voice and noise cues in the dynamic case (Mdn = 3, IQR = 2.5, whereas beep (Mdn = 4, IQR = 1.5) and snare (Mdn = 5, IQR = 0) demand high attention. The ratings for the static cues are a little better (although not significantly), but the task also was simpler (beep: Mdn = 3, IQR = 2, snare: Mdn = 4, IQR = 1, voice: Mdn = 2, IQR = 2, Noise: Mdn = 3, IQR = 1.5). While in the static case, the listening experience for snare (Mdn = 4, IQR = 1) and voice (Mdn = 4, IQR = 1) was rated as *good*, they received lower ratings in the dynamic examples (snare: Mdn = 3, IQR = 1.5 *n.s.*, voice: Mdn = 3, IQR = 1.5 $p = 0.0137$). According to a repeated measures ANOVA, there is a significant effect of the beacon sound on the perceived listening experience. A post-hoc Tukey HSD test showed that the two overlay sounds receive significantly lower ratings than the two integrated ones ($p < .0014$). For the dynamic cues: noise Mdn = 2, IQR = 2, beep Mdn = 2, IQR = 1.5 and for the static ones: noise Mdn = 2, IQR = 1.5, beep Mdn = 3, IQR = 2.

At the end of the survey we asked which of the orientation cues the participants preferred. A majority opted for the singer's voice (8), followed by the two overlay cues noise (4), and beep (1). The snare drum was the least preferred cue of all, although theoretically, it is a good choice of an orientation cue. The ratings are probably low because it is much less present in the mix than the other signals. We did not optimize the mix for the orientation task, but chose one common for this musical genre. The presence of the snare drum in the mix can, however, be emphasized by increasing its volume relative to the other stems.

[2] heller-web.net/navigatone

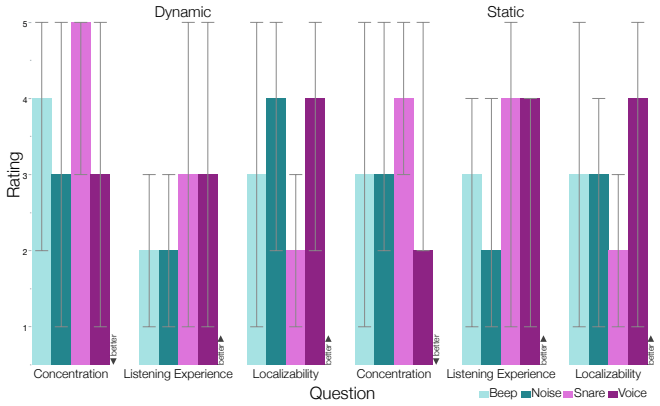

Figure 2. Median ratings of the different cue sounds. We asked how well participants could localize the cue, how much they have to concentrate to localize the sound, and their perception of the overall listening experience. Error bars indicate range.

LOCALIZATION PRECISION

While in the survey mentioned above, the goal was to determine which kind of cue achieves an acceptable balance between localization performance and aesthetic presentation, the following experiment aims at determining the localization performance more precisely.

Implementation

NavigaTone was implemented on an iPad Air 2 running iOS with an attached motion-tracking headset. Spatial audio rendering was performed in the KLANG:app, which uses a generalized Head-Related Transfer Function (HRTF) for spatialization with a resolution of 1° in horizontal and 5° in vertical direction. In a small experiment with 5 users, we determined a minimum audible angle of around 6° in horizontal and 16° in vertical direction, which is in line with our human capabilities to locate sound sources [29]. We loaded the multitrack recording in the software and placed the different sources around the listener's head. To track head movements, we used the Jabra Intelligent Headset (intelligentheadset.com) which comes with a motion tracker, that reports changes in head orientation at a rate of around 40 Hz and has a specified latency of around 100 ms, which is noticeable [6] but well below the threshold of 372 ms defined in [16]. The listener orientation and other relevant playback parameters were sent to the KLANG:app through OSC commands. While the audio data could also have been transmitted to the headphones via Bluetooth along with the sensor data, we used a wired connection between the tablet and the headset to minimize latency. To allow simple replication of this experiment, we used the demo track "Unsere Stadt" that is part of the KLANG:app.

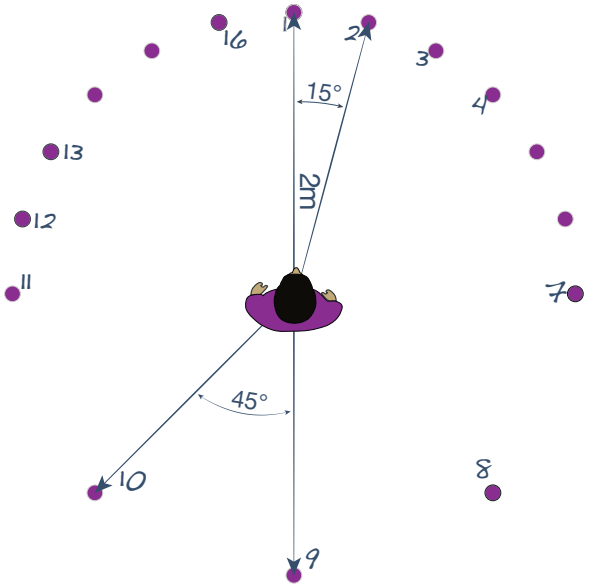

Figure 3. We placed 16 cardboard tubes with 15° spacing at 2m distance from the participant. As in a real world scenario, you would turn around to precisely localize a source in the back of your head, we only tested the ability to detect whether a source is in the parietal hemisphere with three sources.

STUDY

To evaluate the feasibility of our approach in terms of the user experience as well as the performance, we compared NavigaTone to the standard stereo-panning approach as used in gpsTunes [21] or ONTRACK [12, 27]. This baseline is very simple to implement, yet, with the interaural level difference (ILD), covers the most important cue for lateralization. In the NAVIGATONE condition, the sources are arranged around the listener's head using an HRTF based rendering, and only the lead vocal track is moved to communicate the direction for navigation. To differentiate between sources that are in the back or in front of the user, the existing stereo-panning approaches applied a low-pass filter to muffle the sound of sources in the occipital hemisphere of the listener. For reasons of comparability, we used the same rendering as in the NAVIGATONE condition for the STEREO condition, and simulated panning by placing all sources at the same position and moving them in parallel. The hypothesis is, that people can distinguish the source orientation with the same accuracy and that the listening experience is more pleasant in the NAVIGATONE condition.

We placed 16 numbered cardboard tubes with 15° spacing at 2m distance of the listener (Figure 3). Participants were standing at the center of the source circle looking at source number 1. Before every trial we muted the lead vocal track and in the STEREO condition, placed all sources at position no. 1. We then moved the position of the lead vocal track (NAVIGATONE) or all tracks (STEREO) (Figure 4) to one of the 16 cardboard tubes, un-muted the lead-vocal track and participants had to name the tube from which they perceived the voice to come from as fast and accurate as possible. To evaluate the risk of front-back confusions, we allowed the participants to move their head and upper body, as they would

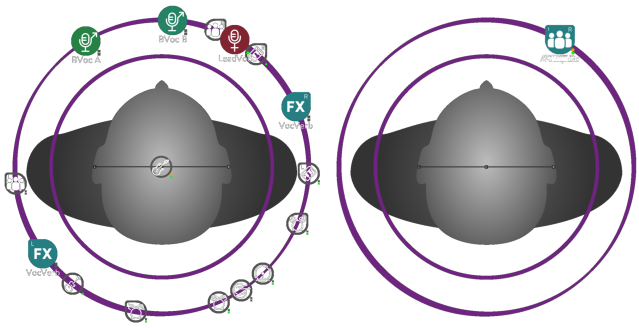

Figure 4. The setup of the user study. (left) In the NAVIGATONE condition just the vocal track was shifted to source no. 3, while the other tracks were positioned in space to create an immersive user experience. (right) in the STEREO condition all tracks were shifted to appear at source no. 3.

do in a real-world application, but not to move their feet. Once a source is perceived in the back, users would turn around until they have the active source in their frontal hemisphere, which is also the reason why we only use three sources in the back of the listener (Figure 3). As the IMU in the intelligent headset tends to drift after fast and large head-turns, we checked the calibration after every trial and realigned the virtual sources if necessary. After both conditions, participants had to fill out an adapted version of the presence questionnaire [30]. We reduced the questionnaire to items applicable to our system and additionally asked how similar the listening experience was compared to regular music listening. All answers had to be given on a five point Likert scale. Finally, participants were asked for feedback in semi-structured interviews.

Results

In total, 16 participants (3 female, average age 27) completed the experiment. The conditions where counterbalanced and the sequence of sources was randomized using Latin squares. None of the participants reported having a hearing disorder, problems with spatial hearing, and three reported having previous experience with audio augmented reality systems.

According to our users, both conditions offered a pleasant music listening experience. Nevertheless, NavigaTone outperformed the baseline in some dimensions. According to a Wilcoxon singed rank test the NavigaTone condition was rated significantly more intuitive and introducing less mental workload. This was also backed up by participants' comments after the test, saying that they found the NAVIGATONE condition to be more natural. One user said "In the first condition *(stereo)*, it was easier, because the sound is just there *(pointing at one of the cardboard tubes)*, but it is also quite narrow. I preferred the second condition *(NavigaTone)* because the sound is all around you." Interestingly, the question "How natural did your interactions with the environment seem?" received the same rating for both conditions (Mdn=2, IQR=1.75). However, the rating for "How similar was the music listening experience compared to regular music listening?" was slightly better in the NAVIGATONE condition (Mdn=1, IQR=3) than in the STEREO condition (Mdn=2, IQR=3), but the difference was not statistically significant.

After they completed the experiment, we asked participants for feedback on the listening experience. Nearly all participants mentioned that they found it easier to localize the sources in the STEREO condition, but that they found the listening experience to be more enjoyable in the NAVIGATONE condition. The STEREO condition has the advantage of providing a very strong cue, whereas NavigaTone aims to provide navigation cues, that are not impacting the music listening experience and are very unobtrusive.

Those findings were also reflected in the variables that compared the performance of both approaches. The average task completion time was 13.3s (SD=7.6) in the NAVIGATONE condition and 10.97s (SD=4.9) in the STEREO condition. A repeated measures ANOVA on the log-transformed task completion times with user as random factor showed that users were significantly faster in the STEREO condition than in the NAVIGATONE condition ($F(1,507) = 11.8$, $p = 0.0006$). No significant effect of source position on the task completion time could be found. Again, as the baseline condition provided a very prominent cue, this result was not surprising.

More interestingly, the recognition rate was slightly better in the STEREO condition ($M= 0.49$, $SD = 0.5$) than in the NAVIGATONE condition ($M=0.41$, $SD = 0.49$), but overall rather low and with a large spread. We calculated the offset between the number of the source actually playing and the given answer and found an average error of 0.83 (SD=0.96) in the NAVIGATONE condition and 0.73 (SD=1.13) for the STEREO condition. This shows that the answers were mostly only off by one source or 15° respectively, which is in the range of human lateralization error [15]. Furthermore, localization performance decreases in the presence of other, competing sound sources [5], which is not the case in the STEREO condition. In a pedestrian navigation scenario, it is rarely necessary to differentiate between two paths at such angular resolution, making both implementations well suitable in practice. If we count the off-by-one answers as correct, then we achieve a recognition rate of 86% (SD=35) for the NAVIGATONE and 90% (SD=30) for the STEREO condition. Interestingly, we observed two cases of front-back confusion in the STEREO condition, but none in the NAVIGATONE condition.

DISCUSSION

Overall, our results of the controlled experiment confirm that it is feasible with NavigaTone to provide navigational cues while listening to music, by shifting just one single track. While the listening experience was rated better, the performance was comparable to the baseline.

Coming from a lab experiment, our results do not account for situations with higher cognitive load as they would be encountered in a real-world navigation scenario. Participants could fully concentrate on determining the origin of the sound, without having to ensure their personal safety by paying attention to traffic lights, pedestrians, or other obstacles. This more complex setting might influence the perception of our cues [24], which is why we plan to run further studies with the presented system as pedestrian navigation system under more realistic conditions.

Again, as the baseline provides a very strong cue, it is not surprising, that NavigaTone did not outperform it. We believe that, in the future, we can further tweak the NavigaTone approach to even outperform the baseline, e.g., by using two tracks that span a navigation vector (one source moves in front of the user, the other one moves in the back) or by finding the sweet spot between both approaches. As participants are familiar with listening to stereo music, the perception of standing in the center of a band and being able to move within (cf. Figure 4) is actually quite different, which could be the reason for the small differences in the ratings.

While multi-track recordings offer great potential, it is still uncommon to release all separate tracks of a song to the public because this would reveal the very core of a music production. As a compromise, Native Instruments' STEMS format includes four distinct tracks for specific parts of a recording [7]. Initially designed to give DJs more creative freedom in mixing two or more songs together, it can also serve as a potential material for NavigaTone. As the file format specifications indicate into which tracks specific instrument or sound groups should be mixed [7], we can pick one with transient sounds.

In both our survey and experiment, we used a vocal pop song as a starting point for our investigation. Other musical genres might have different prerequisites. While the noise beacon somehow fits into the pop-song because of its similarity to a snare drum or hi-hat, it blends less nicely into a piece of classical music. In the future, we intend to further refine the choice of beacon sounds to other musical genres, based on available stems in the recording and mixing qualities with the underlying track.

Participants of the survey also mentioned that they would prefer turn-by-turn style navigation. This could be implemented using a dynamic volume for the navigation cue, dimming it in between waypoints and emphasizing it near changes in direction. For the overlay cues, this reduces their impact on the listening experience, while it might also improve the results of the snare drum cue by making it more present and thus easier do detect if a change in direction is imminent.

We observed that the Invensense MPU-9250 sensor in the Intelligent Headset tends to drift after fast and large head turns. While the fusion algorithms compensate after some time, this can still lead to errors of 45° and more for a brief time. In our experiment, we took great care in controlling the drift and the offset, nevertheless, in a real-world environment the correct alignment of virtual audio source and physical target cannot be guaranteed. Through better sensor fusion algorithms, other IMUs achieve better results in compensating the drift of the gyroscope and therefore show a lower tendency to drift in the overall heading information. However, in practice this behavior might not be of such importance as the paths to choose from are usually well separated, e.g., in city-scale navigation.

CONCLUSION & FUTURE WORK

In this note we presented NavigaTone, a new approach to integrate navigation cues into everyday mobile music listening. Instead of blocking the auditory channel for the single purpose

of presenting an auditory beacon at the target location, we take advantage of multitrack recordings to reduce the impact of the navigation component on the listening experience. In combination with spatial audio rendering, we are able to indicate the direction of the navigation target by moving, e.g., the voice of the singer around the user's head. In a lab study with 16 users, the results of this new approach were on par with the much simpler stereo-panning approach, but found our spatial display to be more natural and less cognitively demanding.

Although increased cognitive load under realistic circumstances might influence the perception, we believe that our approach can provide navigational cues in very different scenarios from pedestrian navigation to navigation in virtual worlds. As we were mostly interested in the ability to localize sources, we performed the lab experiment using a short loop of a vocal track that comes with the Klang:app we used. To be able to work reliably with a multitude of songs, NavigaTone needs to ensure that the orientation cue is audible at the waypoints, i.e., the voice of the singer should be present at an intersection where the user needs to perform a left turn. If we look at the capabilities of modern DJ software that allow us to create remixes on the fly, we can think of incorporating the navigation function even deeper into the playback mechanism. An intelligent algorithm could generate a remix of the original track adapted to the navigation task, with the samples used for localization shifted slightly from their original timing to make sure they are present when needed.

ACKNOWLEDGMENTS

We would like to thank the participants of our study for their time and the people at KLANG:technologies for their feedback and support. This work is part of the SeRGIo project and the Volkswagen Foundation through a Lichtenberg Professorship. SeRGIo is an icon project realized in collaboration with imec, with project support from VLAIO (Flanders Innovation & Entrepreneurship).

REFERENCES

1. Robert Albrecht, Riitta Väänänen, and Tapio Lokki. 2016. Guided by Music: Pedestrian and Cyclist Navigation with Route and Beacon Guidance. *Pers. and Ubiqu. comp.* 20, 1 (2016). DOI: http://dx.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 *CHI '13*. DOI: http://dx.doi.org/10.1145/2470654.2481411
3. Barry Arons. 1992. A Review of The Cocktail Party Effect. *Journal of the American Voice I/O Society* 12 (1992), 35–50. DOI:http://dx.doi.org/10.1.1.30.7556
4. Jens Blauert. 1996. *Spatial Hearing: Psychophysics of Human Sound Localization* (2 ed.). MIT Press.
5. Jonas Braasch and Klaus Hartung. 2002. Localization in the Presence of a Distracter and Reverberation in the Frontal Horizontal Plane. I. Psychoacoustical Data. *Acta Acustica united with Acustica* 88, 6 (2002), 942–955. http://www.ingentaconnect.com/content/dav/aaua/2002/00000088/00000006/art00013
6. Douglas S Brungart, Brian D Simpson, and Alexander J Kordik. 2005. The detectability of headtracker latency in virtual audio displays. In *ICAD '05*. http://hdl.handle.net/1853/50185
7. Chad Carrier and Stewart Walker. 2015. *STEM File Specification*. Technical Report.
8. Richard Etter and Marcus Specht. 2005. Melodious walkabout: Implicit navigation with contextualized personal audio contents. *Pervasive '05 Adjunct Proceedings* (2005).
9. Eve Hoggan, Andrew Crossan, Stephen A Brewster, and Topi Kaaresoja. 2009. Audio or Tactile Feedback: Which Modality when?. In *CHI '09*. DOI: http://dx.doi.org/10.1145/1518701.1519045
10. Simon Holland, David R Morse, and Henrik Gedenryd. 2002. AudioGPS: Spatial Audio Navigation with a Minimal Attention Interface. *Pers. and Ubiqu. comp.* 6, 4 (2002). DOI:http://dx.doi.org/10.1007/s007790200025
11. Andreas Jansson, Eric J Humphrey, Nicola Montecchio, Rachel Bittner, Aparna Kumar, and Tillman Weyde. 2017. Singing voice separation with deep U-Net convolutional networks. In *ISMIR '17*. 323–332.
12. Matt Jones, Steve Jones, Gareth Bradley, Nigel Warren, David Bainbridge, and Geoff Holmes. 2008. ONTRACK: Dynamically Adapting Music Playback to Support Navigation. *Pers. and Ubiqu. comp.* 12, 7 (2008). DOI: http://dx.doi.org/10.1007/s00779-007-0155-2
13. Doreen Kimura. 1964. Left-right differences in the perception of melodies. *Quarterly Journal of Experimental Psychology* 16, 4 (Dec. 1964), 355–358. DOI:http://dx.doi.org/10.1080/17470216408416391
14. Robert W Lindeman, Haruo Noma, and Paulo Goncalves de Barros. 2007. Hear-Through and Mic-Through Augmented Reality: Using Bone Conduction to Display Spatialized Audio. In *ISMAR '07*. DOI: http://dx.doi.org/10.1109/ISMAR.2007.4538843
15. James C Makous and John C Middlebrooks. 1990. Two-dimensional sound localization by human listeners. *J. Acoust. Soc. Am.* 87, 5 (1990). DOI: http://dx.doi.org/10.1121/1.399186
16. Nicholas Mariette. 2009. Navigation Performance Effects of Render Method and Head-Turn Latency in Mobile Audio Augmented Reality. In *ICAD '09*. DOI: http://dx.doi.org/10.1007/978-3-642-12439-6_13
17. T May, S van de Par, and A Kohlrausch. 2013. Binaural Localization and Detection of Speakers in Complex Acoustic Scenes. In *The Technology of Binaural Listening*. DOI: http://dx.doi.org/10.1007/978-3-642-37762-4_15
18. David McGookin and Pablo Priego. 2009. Audio Bubbles: Employing Non-speech Audio to Support Tourist Wayfinding. In *Haptic and Audio Interaction Design*. DOI:http://dx.doi.org/10.1007/978-3-642-04076-4_5

19. Felix Richter. 2014. Infographic: U.S. Teens Love Beats Headphones. (May 2014). https://www.statista.com/chart/2227/preferred-headphone-brands-among-us-teens/

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 *CHI '09*. DOI: http://dx.doi.org/10.1145/1518701.1518722

21. Steven Strachan, Parisa Eslambolchilar, Roderick Murray-Smith, Stephen Hughes, and Sile O'Modhrain. 2005. GpsTunes: Controlling Navigation via Audio Feedback. In *MobileHCI '05*. DOI: http://dx.doi.org/10.1145/1085777.1085831

22. Miikka Tikander, Aki Harma, and Matti Karjalainen. 2003. Binaural positioning system for wearable augmented reality audio. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics.* DOI:http://dx.doi.org/10.1109/ASPAA.2003.1285854

23. Tuyen V Tran, Tomasz Letowski, and Kim S Abouchacra. 2000. Evaluation of acoustic beacon characteristics for navigation tasks. *Ergonomics* 43, 6 (2000). DOI: http://dx.doi.org/10.1080/001401300404760

24. Yolanda Vazquez-Alvarez and Stephen A Brewster. 2011. Eyes-free Multitasking: The Effect of Cognitive Load on Mobile Spatial Audio Interfaces. In *CHI '11*. DOI: http://dx.doi.org/10.1145/1978942.1979258

25. Yolanda Vazquez-Alvarez, Ian Oakley, and Stephen A Brewster. 2012. Auditory display design for exploration in mobile audio-augmented reality. *Pers. and Ubiqu. comp.* 16, 8 (2012). DOI: http://dx.doi.org/10.1007/s00779-011-0459-0

26. Bruce N Walker and Jeffrey Lindsay. 2005. Navigation performance in a virtual environment with bonephones. In *ICAD '05*. http://hdl.handle.net/1853/50173

27. Nigel Warren, Matt Jones, Steve Jones, and David Bainbridge. 2005. Navigation via Continuously Adapted Music. In *CHI EA '05*. DOI: http://dx.doi.org/10.1145/1056808.1057038

28. Dirk Wenig, Johannes Schöning, Brent Hecht, and Rainer Malaka. 2015. StripeMaps: Improving Map-based Pedestrian Navigation for Smartwatches. In *MobileHCI '15*. DOI: http://dx.doi.org/10.1145/2785830.2785862

29. Elizabeth M Wenzel, Marianne Arruda, Doris J Kistler, and Frederic L Wightman. 1993. Localization using nonindividualized head-related transfer functions. *J. Acoust. Soc. Am.* 94, 1 (1993). DOI: http://dx.doi.org/10.1121/1.407089

30. Bob G Witmer and Michael J Singer. 1998. Measuring Presence in Virtual Environments: A Presence Questionnaire. *Presence: Teleoper. Virtual Environ.* 7, 3 (1998). DOI:http://dx.doi.org/10.1162/105474698565686

31. Shingo Yamano, Takamitsu Hamajo, Shunsuke Takahashi, and Keita Higuchi. 2012. EyeSound: Single-modal Mobile Navigation Using Directionally Annotated Music. In *Augmented Human '12*. ACM, New York, NY, USA. DOI:http://dx.doi.org/10.1145/2160125.2160147

32. Matthijs Zwinderman, Tanya Zavialova, Daniel Tetteroo, and Paul Lehouck. 2011. Oh music, where art thou?. In *MobileHCI '11 EA*. DOI: http://dx.doi.org/10.1145/2037373.2037456