Guiding Visual Search Tasks Using Gaze-Contingent Auditory Feedback

Viktor Losing  
Faculty of Technology,  
University of Bielefeld  
33615 Bielefeld, Germany  
vlosing  
@techfak.uni-bielefeld.de

Thies Pfeiffer  
Center of Excellence Cognitive Interaction Technology  
Inspiration 1  
33615 Bielefeld, Germany  
tpfeiff @techfak.uni-bielefeld.de

Lukas Rottkamp  
Faculty of Technology,  
University of Bielefeld  
33615 Bielefeld, Germany  
lrottkam  
@techfak.uni-bielefeld.de

Michael Zeunert  
Faculty of Technology,  
University of Bielefeld  
33615 Bielefeld, Germany  
mzeunert  
@techfak.uni-bielefeld.de

Abstract
In many applications it is necessary to guide humans’ visual attention towards certain points in the environment. This can be to highlight certain attractions in a touristic application for smart glasses, to signal important events to the driver of a car or to draw the attention of a user of a desktop system to an important message of the user interface. The question we are addressing here is: How can we guide visual attention if we are not able to do it visually? In the presented approach we use gaze-contingent auditory feedback (sonification) to guide visual attention and show that people are able to make use of this guidance to speed up visual search tasks significantly.

Author Keywords
sonification; auditory feedback; eye-tracking; gaze-contingent, visual search, guidance

ACM Classification Keywords
H.5.2. [Information Interfaces and Presentation (e.g. HCI)]: Multimedia Information Systems

Introduction
Guiding someone’s visual attention to a specific point in the environment is crucial for many communication tasks. Two humans will naturally make use of pointing gestures
to direct the visual attention of their interlocutor to objects in the environment. Screen-based computer interfaces use flashing effects to attract attention. All the described strategies are purely visual and only work if the recipient is able to see the action, if only in the periphery. One shortcoming with current technical approaches is that the systems require an explicit user action to acknowledge the perception of the signal. In our approach we address this problem by monitoring the visual attention of the user with an eye-tracking system. A more difficult problem arises if the technical system that is trying to guide the attention of a user cannot easily be integrated into the visual path between the recipient and the target. Examples are assistance systems during driving (without HUDs), applications for certain smart glasses or for example systems that guide a visitor’s gaze towards interesting areas of paintings in a museum.

In such cases a cross-modal approach, for example using auditory or tactile feedback, could be a viable alternative. Following this idea, there are several questions that need to be answered: How can such cross-modal feedback be realized? Can people integrate this new kind of feedback and make active use of it? What precision can be achieved with such a cross-modal approach? How does the feedback affect the performance of the user?

In this paper we present a gaze-contingent approach for cross-modal visual guidance using acoustic signals. The aim is to guide the recipient’s visual attention towards a specific point in the environment by providing auditory feedback (see example in Figure 1).

A naive approach would be transforming the position of the target into a 3D sound, presenting that to the user and relying on the 3D acoustic orientation to get the direction. While this might work for coarse basic left-right orientations, it would be difficult to get more fine-grained orientation changes or changes along the vertical axis. In addition to that, there would be only a direct feedback loop while the user turns the head towards the target (as the ears will then move and the computed 3D sound will have to be updated), but not if the general direction towards the target has been reached.

In our approach we instead use a direct feedback loop to signal the error between the current direction of gaze as measured by an eye-tracking system and the target direction. Following eye movements provides a much higher granularity and the error measurement works in any direction (horizontal and vertical) equally well. After a short summary of related work, we present the details of our approach and then report on a study using desktop-based visual search tasks in which we show that users are able to make use of this gaze-contingent sonification to speed up visual search tasks.

Related work
Auditory feedback
Sonifications are often used if the primary modality, most of the time the visual, is occupied with a different task, as for example to indicate the state of anesthetized patients [14]. There are also a few examples in which the sonification is embedding spatial information. Simple collision warnings based on the distance to the target have been shown to be successful for pilots [1] or car drivers [13], and these systems are already broadly available, e.g. as parking aids. These systems are including a human into the control loop (hearing the warning sound, reacting by reducing speed or stopping), but the control task is only one-dimensional and the feedback is about the state of a system (car, plane) not the user herself.
In contrast, auditory feedback can also be the dominant or sole feedback channel, as in the case of gaze-triggered musical instruments [8]. There are also some interesting examples in which the auditory feedback is used to compensate for the visual channel, for example in the AcouMotion System [6], which enables blind people to play Badminton with a virtual ball. The motion of the ball is simulated and real rackets using motion tracking are used to interact with the system. In another scenario, areas of interest are detected in camera images taken from ego-perspective and transformed into acoustic signals to aid visually impaired people [2]. The task here is to provide a quick and rough estimation of the environment, but obviously not the guidance of a visual search.

Figure 2: The presented approach has been tested in a desktop-based set-up using the commercially available remote eye-tracking system MyGaze. The bright circle shows the current fixation of the participant for demonstration purposes, in the study condition the plain image is presented without visual augmentations.

Eye tracking and auditory feedback
An evident use of converting the user’s eye movements to sound is the purpose of creating music for entertainment, as for example through an visually controlled interface [7]. Away from defined interfaces, gaze data can also be related to the user’s environment as for example when making depth audible for visually impaired people [11]. In these cases the user’s gaze triggers the system, but the system does not observe and guide the user in a goal directed way.

Eye tracking and visual search
Visual search has been studied “with text or textlike material, with pictorial stimuli, with complex arrays such as X-rays, and with randomly arranged arrays of alphanumerical characters or objects” [3]. Related work contains studies regarding qualitative and quantitative aspects of visual processing in search tasks, which however—to the best of our knowledge—never consider the use of an active guiding system. Passive uni-modal guidance encoded in the visual stimuli has been successfully used to investigate a process-monitoring model of visual search [5]. A related example of active support proposes aiding visual search by occluding already searched areas [10], which is a way to support search but it is not guided by the system.

Our approach
In our approach, the system aims at guiding the visual attention of the user towards a specific target. By monitoring the eye movements of the user, a direct feedback loop is generated and the feedback is given using acoustic signals. To generate appropriate auditory feedback in a given situation, basically three steps have to be taken: Firstly, the difference between the current direction of gaze and the target direction has to be measured. Secondly, this difference has to be transformed into an appropriate feedback value, which is then made audible in a third step.

Gaze-contingent measurement of distance to target
As a precondition, the direction of the target relative to the user has to be known and the system has to have access to the current direction of gaze. In desktop-computer scenarios, such as the one we use for our study, typically a remote eye-tracking system is mounted below the screen at a distance of approximately 70 cm (see Figure 2). Recent developments such as the EyeTribe system, the Tobii Rex or the MyGaze system provide means to realize this setup at a small investment. Such systems will provide the screen coordinates at which the user’s gaze is directed. If the target is displayed on the screen, the angular difference can be approximated by the euclidean distance between the two screen coordinates, the current gaze coordinates and the coordinates of the target. In a mobile set-up, the angular difference between the 3D vector pointing from the user towards the target...
and the 3D vector representing the current direction of
gaze needs to be approximated.

**Determining feedback value**
The naive approach to converting this difference to an
user feedback would be to scale it linearly. However,
auditory displays have a restricted feedback interval, as for
example a too low or too high pitch or extreme loudness is
not appropriate. We noticed in a preliminary study, that a
linear mapping to a chosen interval covering the whole
screen real estate leads to only subtle perceived
differences between two similar angles because of the low
slope of the mapping. This proved to be troubling when
the target was already approximately located and the user
was merely fine-tuning towards its exact location. Using a
capped exponential or bell-curve mapping on the other
hand appeared to be rather unnatural for the user. Our
compromise consists of a linear mapping when being
"near the target" covering only a quarter of the screen
real estate, and a constant value anywhere else. The user
then typically probes a few regions and starts searching
locally as soon as a difference is discovered. Of course,
the area of constant feedback has to be reasonably sized.
In our setting, every gaze direction having a higher
angular difference than 20° fell in this area, approximately
half of the left-to-right viewing angle. This is visualized in
Figure 3.

**Auditory feedback**
When it comes to presenting data in the form of audio
signals, different mappings are possible. Often chosen
dimensions include pitch, loudness and tempo, which are
easy to implement and intuitive [12]. Literature often
advises to handle loudness with care, as it interferes with
other dimensions such as pitch [9], or may be especially
sensitive to equipment changes [4]. Still, during an early
evaluation of our system, we found that mapping silence
to the concept of no error, has a distinct advantage: The
user at once knows whether he has indeed reached the
target. This appeared desirable and convinced us to
indeed use loudness to represent our single dimension by a
linear mapping to the amplitude of a constant, rather low
tone of 196 Hz. We further make use of the *Synthesis
ToolKit in C++*¹ to generate a flute sound, which is
hoped to be more pleasant than for example a sine tone.

**User study**

**Method**
To evaluate our approach, we conducted a small study
with ten members of our faculty. Some of them used
eye-tracking systems before, but nobody had prior
exposure to our system. The participants were pursuing a
master’s or doctor’s degree, and therefore had previous
experience with user studies in general. After a short
explanation of our system itself, we asked each participant
to sit down at a typical desktop workstation, roughly
70 cm in front of a 22" display at which our eye tracker, a
MyGaze system² (30 Hz; 0.5° gaze position accuracy
according to manufacturer), was mounted. We then
calibrated the eye-tracking system with its 5-point
calibration tool and verified proper functionality by asking
the user to for example target the screen’s corners.
Finally, the user was equipped with headphones which also
were adjusted so that the user confirmed appropriate
volume settings.

The study itself was a collection of visual search tasks
presented in a slideshow-like fashion: on each slide the
user had to click with the mouse on a specific search

¹Perry R. Cook and Gary P. Scavone: “The Synthesis ToolKit in
C++”, https://ccrma.stanford.edu/software/stk/
²http://www.mygaze.com/
target. After the click the next slide started, there was no explicit feedback on whether they had been successful in identifying the target or not. Half of the tasks were supported by our auditory feedback, the other half were pure visual search tasks. We had two groups of participants, one were supported using acoustic guidance on even slides the other group on odd slides. This way we recorded interpersonal data of the performance with and without guidance for each task. After a few brief slides explaining the procedure, the training phase started. Then, with breaks in-between, three different kinds of tasks followed, as described below. After finishing these tasks, which took approximately 20 minutes, we further asked the user to give us feedback regarding her or his experiences with the system.

**Tasks**

The training phase consisted of 10 trials, each of which presented the same 8x5 dot grid on the screen. Distance between two horizontally or vertically neighboring dots was 210 px or 5.91 cm, which translates to an angular difference of 4.84° from the participant’s point of view. In each trial, the coordinates of one dot acted as the search target. The user’s task was to select this target point by clicking on it. However, the user was not able to distinguish this target from other points visually. As the targets were chosen in a non-predictable pattern\(^3\), he or she had to rely on the auditory feedback given through the headphones to locate the target. After selecting a point, a blank screen was displayed for half a second, after which the next trial started. As soon as the tenth trial was completed, a slide stated the end of the training phase and we went on to ask if the user experienced any problems and then explained the next task.

The following tasks, here addressed as shapes-, numbers- and picture-task, differed from the training phase by having a visually detectable target—this was however not immediately obvious: The shapes-task consisted of a 40x25 grid of squares, which included but one circle, acting as the target in the fashion described above. An excerpt of the resulting screen image can be seen in Figure 4. The distance between two horizontally or vertically neighboring symbols was 42 px or 1.18 cm, implying an angular difference of 0.96° from the participant’s point of view. Items alternated between having acoustic guidance as experienced in the training phase and having no acoustic guidance, requiring the user to perform the search task purely visually. To obtain measurements five times with and five times without auditory feedback for each item, we split participants into two groups. 20 items were presented in the shapes- and numbers-task respectively, while the picture-task consisted of but nine different items. To prevent user frustration, we implemented a per-item timeout, which automatically switched to the next item if the user was not able to locate the target before the 90th second. For each item, the duration between start (search task is displayed) and end (user submits a solution per mouse click or timeout after 90 seconds, whichever occurred first), was recorded.

After the user completed the shapes-task, we explained the numbers-task, which only differed by the target being one 9 in a grid of 8s while the grid resolution and structure remained unchanged. This task was chosen to be more time-consuming than the shapes-task, hopefully inspiring the users to make heavy use of the acoustic guidance again available with every other item.

\(^3\)The target of every item was randomly placed before the start of the study and fixed for all participants, so that every user was presented the same, albeit non-predictable, sequence of items. This applies not for the picture-task, whose items were not systematically generated by us.
When the user had completed this task, we explained the picture-task, which consisted of images depicting cartoons of “real-life” scenes. Each item now had a different target, a person with certain characteristics, which was explained in a brief task on the left side of the picture. Again, auditory feedback was only activated in every other item, depending of the participant’s group membership.

Results

Table 1: Average time in seconds spent per item for each of the 10 participants. OFF and ON describe the state of the sonification module. Each cell describes 10 measurements, except for the picture-task, which only describes 4 each. */x marks significant speedups at 1%/5% significance level.

<table>
<thead>
<tr>
<th>User #</th>
<th>Shapes OFF</th>
<th>Shapes ON</th>
<th>Numbers OFF</th>
<th>Numbers ON</th>
<th>Picture OFF</th>
<th>Picture ON</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.0</td>
<td>4.0</td>
<td>17.6</td>
<td>*5.9</td>
<td>18.0</td>
<td>15.6</td>
</tr>
<tr>
<td>2</td>
<td>10.2</td>
<td>*4.5</td>
<td>61.7</td>
<td>*7.8</td>
<td>38.9</td>
<td>9.0</td>
</tr>
<tr>
<td>3</td>
<td>7.9</td>
<td>5.9</td>
<td>36.1</td>
<td>16.6</td>
<td>31.4</td>
<td>24.9</td>
</tr>
<tr>
<td>4</td>
<td>5.1</td>
<td>4.1</td>
<td>39.0</td>
<td>*9.2</td>
<td>22.9</td>
<td>11.4</td>
</tr>
<tr>
<td>5</td>
<td>9.2</td>
<td>8.9</td>
<td>45.1</td>
<td>*8.0</td>
<td>30.7</td>
<td>27.3</td>
</tr>
<tr>
<td>6</td>
<td>4.9</td>
<td>4.1</td>
<td>53.3</td>
<td>*9.0</td>
<td>22.9</td>
<td>*6.2</td>
</tr>
<tr>
<td>7</td>
<td>6.0</td>
<td>4.5</td>
<td>31.7</td>
<td>*7.4</td>
<td>12.9</td>
<td>19.0</td>
</tr>
<tr>
<td>8</td>
<td>7.1</td>
<td>5.9</td>
<td>32.5</td>
<td>*6.5</td>
<td>29.2</td>
<td>8.6</td>
</tr>
<tr>
<td>9</td>
<td>5.1</td>
<td>5.1</td>
<td>19.7</td>
<td>*6.4</td>
<td>10.3</td>
<td>16.9</td>
</tr>
<tr>
<td>10</td>
<td>7.4</td>
<td>4.9</td>
<td>33.2</td>
<td>*7.2</td>
<td>46.7</td>
<td>8.6</td>
</tr>
</tbody>
</table>

MEAN 6.7 5.2 37.0 8.4 26.4 14.7

Figure 5: Heatmap of gaze points of two numbers-items. The top image shows the distribution of attention without acoustic guidance, the bottom image that with acoustic guidance. The white circle marks the target. The heatmap shows that irrelevant areas distant to the target are receiving less attention when using the acoustic guidance.

Figure 6: Comparison of search time per item in seconds for the three tasks shapes, numbers and picture, each with and without acoustic guidance. Plot shows quartiles (box), median and outliers. The whiskers mark the most extreme data points not more than two times the interquartile range away from the nearest quartile. Points outside are plotted as outliers.

The values obtained during the picture-task shown in Table 1 should be approached with special caution. The task consisted of but nine items, of which the first was excluded as some users did not at once understand the new task. As a consequence, each average is only calculated over four measurements. The items further differed in difficulty. While it appears that participants benefited from using the acoustic guidance, the data of only one person shows a significant decrease of search time.

visual search and thus a significant decrease of time needed per item was achieved using acoustic guidance for almost all participants (see Figure 5 for examples of typical attention distributions for this task). Participant 2, showing sub-average performance in the non-guided case, profited most drastically by using the guidance, then achieving average performance.
Significance was evaluated using the unpaired one-sided Wilcoxon signed-rank test between measurements with and without acoustic guidance. The data can not be assumed to be normally distributed, as for example users sometimes were very fast detecting the target.

Tables 2 and 3 present a per-item view of the measured data. The shapes-task is not presented, as only small improvements were recorded (see Table 1). Looking at numbers-task (Table 2), a much stronger difference can be seen. Here, 16 timeouts were recorded without acoustic guidance, which were evaluated as 90 seconds. Most of these timeouts occurred in the second half of the task, maybe indicating a certain loss of concentration, which does not manifest when these items were presented in a guided fashion (remember that conditions were presented interleaved). Here, item 4 contained one comparably very slow answer (58.9 s) which was caused by the participant experimenting with the system, as he told us later. Despite some curious effects, for example item 20 having the lowest mean with guidance compared to timing out two times without, which can be explained with the target being in the last row and consequently found very late in a row-by-row search pattern, the data strongly indicates the acoustic guidance as being helpful.

The picture-task (Table 3) also shows a decrease of mean search time for every item. The deviation between users was also often greatly decreased, owing to the fact that a target may be overlooked without acoustic guidance. In this situation, the user typically searches the complete "untouched" area, which causes a high measurement. With acoustic guidance, these long searches, possibly even ending in a timeout, can be avoided. Items 1 and 3 seem to be more difficult, while especially items 2, 5, 7 and 8 seem to be easily understood and promptly found.

<table>
<thead>
<tr>
<th>Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item</td>
</tr>
<tr>
<td>#</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>11</td>
</tr>
<tr>
<td>12</td>
</tr>
<tr>
<td>13</td>
</tr>
<tr>
<td>14</td>
</tr>
<tr>
<td>15</td>
</tr>
<tr>
<td>16</td>
</tr>
<tr>
<td>17</td>
</tr>
<tr>
<td>18</td>
</tr>
<tr>
<td>19</td>
</tr>
<tr>
<td>20</td>
</tr>
</tbody>
</table>

Table 2: Measured times in seconds and number of timeouts (more than 90 s) spent per item for each of the 20 NUMBER items. Each cell describes 5 measurements.
<table>
<thead>
<tr>
<th>Item #</th>
<th>Sound OFF</th>
<th>Sound ON</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>1</td>
<td>37.3</td>
<td>26.1</td>
</tr>
<tr>
<td>2</td>
<td>17.9</td>
<td>15.3</td>
</tr>
<tr>
<td>3</td>
<td>57.5</td>
<td>28.9</td>
</tr>
<tr>
<td>4</td>
<td>26.8</td>
<td>36.4</td>
</tr>
<tr>
<td>5</td>
<td>17.0</td>
<td>13.8</td>
</tr>
<tr>
<td>6</td>
<td>24.2</td>
<td>21.0</td>
</tr>
<tr>
<td>7</td>
<td>16.6</td>
<td>6.9</td>
</tr>
<tr>
<td>8</td>
<td>13.7</td>
<td>4.5</td>
</tr>
<tr>
<td><strong>MEAN</strong></td>
<td><strong>26.4</strong></td>
<td><strong>19.1</strong></td>
</tr>
</tbody>
</table>

Table 3: Measured times in seconds and number of timeouts (more than 90s) spent per item for each of the 8 PICTURE items. Each cell describes 5 measurements.

**Participant feedback**

We asked every participant for feedback regarding the performance of the system, applied search strategies and overall experience. Questions asking if any delay of the auditory response was perceived were always negated, which was also the most common answer when we asked if the system appeared inaccurate. One participant however stated that he perceived the numbers-task as very accurate, the picture-task less so. The next participant then described the numbers-task as being inaccurate but the picture-task very accurate.

However these, and most other participants, stated that they used the acoustic guidance mainly for finding the right region, and then locally using it much less or even only for verification (see Figure 7, bottom), especially in case of the picture-task. In contrast, when having no guidance, most participants applied a row-wise, or column-wise search pattern for the numbers-task (see Figure 7, top), often complaining about it’s cumbersomeness but also mentioning that other patterns often overlook the target. Some refined it by ”risking” a brief scan before and only falling back to a structured pattern if this scan was unlucky. Regarding the shapes-task, it was often said that the task was easy enough so that no special strategy, or auditory feedback, was necessary. When asked if the auditory feedback was then unwelcome, the typical response was that it would not be unhelpful or indicated that it ”was not doing any harm”. Although every participant assessed the system as well-functioning and useful, some agreed that the sonification tone itself could need some improvement.

One participant proposed sounds ”as they occur in natural environment”, however no other ideas could be gathered.

**Conclusion**

The significant increase of search performance in the tasks with the numbers is a promising first result of our approach. A striking result is the strategy developed by the participants to incorporate the acoustic guidance into the search process: From our small sample it seems that participants first locate the right region almost only by hearing, and then, as if as a second step, locally search the target almost only by seeing. When looking at tasks with more “hidden” targets, we believe that the acoustic guidance will be more important in the local search, as we have seen in our picture-task. Another interesting result is that even easy tasks, which would not require special guidance, can benefit from such a guidance.

The mapping of the distance to the target to a sonification is still rather ad-hoc and could benefit from optimizations. This will be done systematically in follow-up studies. Besides different mapping functions that are better adjusted to human’s perceived loudness, other dimensions of acoustic feedback will be evaluated.

In a larger mobile set-up, we are also planning to use 3D
spatial sound and thus the location of sound around the user, which was not practical for the small search domain presented on the computer screen.

It would be interesting to compare this cross-modal guidance with visual guidance, e.g. using pointers or blinking effects. However, as delineated above, we are interested in approaches that do not interfere with the modality that is to be guided. While it is very reasonable to expect that any form of feedback would improve the search process, this might depend on many aspects. For example, while we did expect a clear advantage for the numbers-task, the shapes-task seemed to be simple enough to be solved equally well without guidance. But even there we find a small yet not significant improvement. Providing feedback in a cross-modal approach could also help to reduce fatigue, not only because the overall time needed for the search process is significantly reduced, but also because one modality could recover while the other modality is made use of. This has to be investigated in longer sessions in which the level of fatigue is monitored over time, e.g. by measuring the blink rate.

It would be interesting to incorporate gaze-contingent auditory feedback into a desktop application to facilitate the perception of alerts. We are also planning to migrate to mobile settings, e.g. the already mentioned scenarios at the point-of-sale, to extend the approach to fully surrounding 360° scenarios. In such settings the guidance will have to be more sophisticated as the screen real estate to be covered by the mapping is much larger and, for example, complete 180° turns have to be covered.

Acknowledgments
We would like to thank the participants of our study for their time and feedback. We also thank the reviewers for their contribution and tips. This work has partly been supported by the Deutsche Forschungsgemeinschaft in the SFB 673 Alignment in Communication and the Center of Excellence Cognitive Interaction Technology.

References


