Audio Engineering Society

Convention Paper

Presented at the 133rd Convention
2012 October 26–29 San Francisco, USA

Interactive 3D audio: Enhancing awareness of details in immersive soundscapes?

Mikkel N. Schmidt[1], Stephen Schwartz[2], and Jan Larsen[1]

[1] *DTU Informatics, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark*

[2] *SoundTales, Marienlyst Alle 41, 3000 Helsingør, Denmark*

Correspondence should be addressed to Mikkel N. Schmidt (mns@imm.dtu.dk)

ABSTRACT

Spatial audio and the possibility of interacting with the audio environment is thought to increase listeners' attention to details in a soundscape. This work examines if interactive 3D audio enhances listeners' ability to recall details in a soundscape. Nine different soundscapes were constructed and presented in either mono, stereo, 3D, or interactive 3D, and performance was evaluated by asking factual questions about details in the audio. Results show that spatial cues can increase attention to background sounds while reducing attention to narrated text, indicating that spatial audio can be constructed to guide listeners' attention.

1. INTRODUCTION

It is well known that spatial auditory cues such as interaural time and level differences (ITD and ILD) influence the way listeners perceive and group different sound sources in complex audio environments. In a real listening situation in a complex environment, the sound we hear is a convoluted mixture of different sound sources and reverberation arriving from many different directions simultaneously. Despite the complexity of the signal, the human auditory system is capable of identifying and grouping sounds into coherent streams [1].

Through a top-down cognitive process (endogenous attention) we are able to focus on a single sound source, while ignoring other sources which are perceived as noise. However, through a bottom-up cognitive process (exogenous attention) loud or sudden sounds might attract our focus. Thus, what a listener attends to when immersed in a complex soundscape depends both on the listener's concious choices and on properties of the sound itself.

In simple lab experiments studying the process of auditory grouping, spatial cues have been found to be relatively weak compared to other cues such as simultaneous onset or temporal continuity; however, recent findings suggest that "...in more complex conditions, spatial cues are critical for properly parsing the mixture of sound into different objects and

focusing attention on the source of interest." [6] In complex soundscapes, we might thus expect that spatial cues play an important role for attention since "... auditory attention is directed toward objects in subjective locations" [3]. In addition, it has been found that auditory grouping does not happen as a cognitive process isolated from attention, but that attention also influences the process of auditory streaming in an adaptive manner [2].

Another issue that might influence auditory streaming and attention is interaction with the audio environment: For example, it is known that "turning one's head while the sound is being given" aids in localizing sound sources [7]; however, whether head movement plays a substantial role in auditory scene analysis, streaming, and attention or if it is only important for localization is not fully understood. It would be reasonable to believe that moving one's head or more substantially changing listening position might improve the listeners abilty to discern different sound sources since switching position changes the relative signal to noise ratio of the different sources, providing a possible cue for auditory streaming.

1.1. Hypotheses

In this work, we study how the spatial auditory cues combined with the possibility of moving around in an immersive 3D audio soundscape influcences the listeners attention to details in a complex, artificially generated sound environment. Our hypothesis is that realistic spatial cues and interaction improve the listeners ability to discern sound sources leading her to pay more attention to details in the sound environment. If this hypothesis is true, a listener exposed to interactive 3D audio should be able to better recall factual details about the sound environment than a listener who listens passively to monaural audio (without spatial information).

It might, however, also be the case that asking the listener to interactively move around in the soundscape incurs a cognitive load that leads to her paying less attention to the details in the soundscape. For that reason, we compare also to non-interactive 3D audio. Finally, we wish to compare to normal non-interactive stereo audio, to examine whether rudimentary spatial cues (stereo panning) is significantly different from more advanced 3D audio.

Furthermore, we will examine if the hypothesized benefits of 3D and interactive 3D audio differs between groups of relatively experienced (skilled) and inexperienced (unskilled) listeners.

2. METHOD

To examine our main hypothesis, we conducted a set of experiments in which subjects were asked to listen to a number of soundscapes in either interactive 3D audio or non-interactive 3D, stereo, or mono audio. After listening to each soundscape, the subjects were asked a number of factual questions to examine whether certain details had been noticed. Furthermore, the subjects were asked to give a subjective evaluation of a number of aspects related to the listening experience.

2.1. Listening test

First of all, to examine the subjects' listening skills prior to the experiment, we designed a simple listening test. The test consisted of eight sounds, which the subjects were asked to identify or asked a specific question about. The sounds were presented in random order, and the subjects were asked to answer in free text. Each question was anonymously scored, giving one point for each correct answer, one half point for each partially correct answer, and zero points for each incorrect answer. The results of this test was subsequently used to assess differences between skilled and unskilled listeners.

2.2. Soundscapes

We produced nine soundscapes: Six of the soundscapes were narrated, and the remaining three consisted only of environmental sounds. The soundscapes comprised two themes: "The walking street," and "Søren Kierkegaard"; the former set on the walking street of Copenhagen where the listener experienced a number of street performers, and the latter built around some of Søren Kierkegaard's short texts and aphorisms accompanied by suiting background sounds reminiscent of the 19th century.

The soundscapes were created in the following manner: A number of original recordings of background and foreground sounds including narration were con-

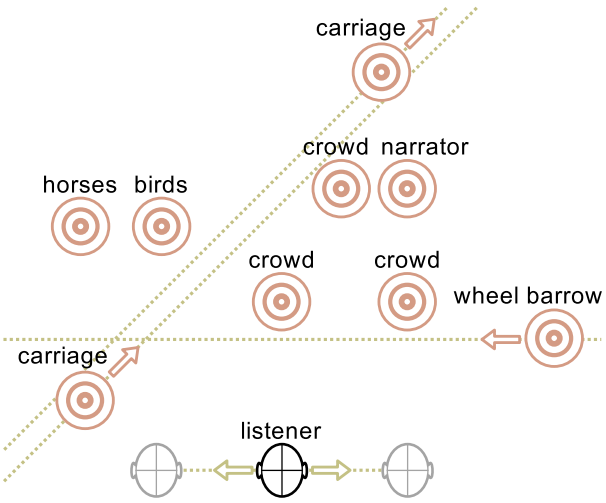

Fig. 1: Example of a narrated soundscape with a number of stationary background sounds (crowd noise, birds, and horses) as well as carriages and a wheel barrow passing by. In the interactive 3D sound setting, the listener is able to move left and right on a prespecified path. In this example, going left would move the listener closer to the horses, birds, and passing carriages, while going right moves the listener closer to the narrator and crowd noise.

ducted.[1] All sound were converted to mono, and each sound was placed at a fixed coordinate in a three-dimensional virtual space or set to move along a path at a given speed. The listening position was placed at a position in the virtual space, and allowed to move left and right on a fixed path by using the arrow keys. See Fig. 1 for an illustration of one of the soundscapes. In the soundscapes that included narration, the narrator was always placed in a front right position.

2.3. Interactive 3D audio

To generate the interactive 3D audio (i-3D) soundscape, we constructed a basic system for computed real time auralization (see e.g. [4]). The sound was filtered and mixed in real time according to the relative positions of the listener and the sound sources. To achieve a simple but convincing 3D audio effect we combined two techniques to simulate lateral and distance cues. For the primary lateral cues (interaural time and level differenecs), we used standard head related transfer functions recorded using a Kemar dummy head. For the primary distance cues (loudness and ratio between direct and reflected sound) we attenuated sounds dependent on the distance and mixed in signal filtered by a diffuse binaural room impulse response at a ratio increasing with the distance to the source. The system parameters were hand tuned to yield what we perceived as a convincing spatial auralization.

To play the interactive soundscape, the sound sources were filtered and mixed in real time and output to the sound device: For each sound source, the angle and distance from the listener to the source was computed. Depending on the angle and distance, the sound source was filtered to generate simulated 3D signals for the left and right ear at the position of the listener. All filtered sound sources were added and played back in real time using the overlap-add method in the frequency domain.

Pressing the left or right arrow key on the computer keyboard allowed the listener to move around in the soundscape. Interaction was limited to moving left and right on a prespecified path, as illustrated in Fig. 1.

2.4. Non-interactive audio

In addition to interactive 3D (i-3D) audio, we produced non-interactive soundscapes in mono, stereo, and 3D audio for comparison. 3D audio was produced by recording the output of the interactive 3D audio engine while keeping the listening position fixed at the initial center position. Thus, the 3D condition was exactly identical to the i-3D condition if the listener had decided not to interact with the soundscape. For the stereo condition, we again recorded the output from the interactive 3D audio engine at the center listening position. The difference between 3D and stereo was that in stereo the lateral and depth binaural cues were limited to interaucal level difference and overall loudness, corresponding to a simple stereo panning. For the mono condition, the left and right channels in the 3D condition were averaged and presented to both ears, thus removing all lateral binaural cues. Audio recordings of the non-interactive soundscapes are available for download [5].

[1] The original recordings were produced by the second author, Stephen Schwartz, SoundTales, and Henrik Olsen, Det Gode Øre, in collaboration with Master interns Wazir Ilyas Abdulrahman, Jung In Jung, and Clive Mitchel at the University of Edinburgh.

2.5. Questions

To inquire into the subjects' listening experience and abilty to recall details, we designed a number of questions centered around three categories: The listening experience, recalling details related to the narration, and recalling details related to the audio environment. All questions were multiple choice with three or four possible answers. The complete list of questions is available for download [5].

The same seven questions related to the *listening experience* were asked after each soundscape. As an example, we asked: *"To what degree were you captivated in the setting's space?"* and gave the options of answering *"Not captivated," "Slightly captivated," "Fairly captivated,"* or *"Strongly captivated."* These questions all asked the subjects to rate a specific aspect of the listening experience on a four point scale. The remaining questions concerned entertainment value, sense of space, clarity of background sounds, overall sound quality, depth perspective, and perceived separation of sound sources.

The questions related to *narration* were designed to probe whether the subject remembered a specific phrase. An example of a question is the following: *"What is mentioned that the woman rescues from the fire?"*, where the possible answers were *"The fire wood," "The fire place,"* or *"The fire tongs."* These questions were all multiple choice with three options, and a total of 23 questions were designed for the six narrated soundscapes.

Finally, questions related to the *audio environment* were designed to examine to what extend subjects had noticed and were able to remember specific sounds heard as part of the soundscape. For example, after one soundscape we asked: *"Which game was played at the end?"* and provided three options: *"Chess," "Backgammon,"* or *"Cards"* to examine if the listener had noticed the characteristic sound of dice in a cup. For the nine soundscapes we designed a total of 32 such questions, all presented as multiple choice with three options.

2.6. Experimental design

The experiment consisted of the initial listening test, a short introduction to interactive 3D audio, followed by presentation of the nine soundscapes, each in one of the four conditions: Mono, stereo, 3D or i-3D. For each subject, the soundscapes within one theme (walking street and Søren Kierkegaard) were presented as a block, and the order of the blocks was randomized between subjects. Wihin each block, the order of the soundscapes was randomized. Each soundscape was presented to each subject in a condition chosen using a Latin square experimental design, such that all subjects experienced each condition 2-3 times thoughout the experiment.

Thirty-one subjects were recruited for the experiment. Around half of the subjects were recruited through the second authors personal network and were mainly people working in the audio and broadcasting business. The rest were recruited through bulleting board postings at the Technical University of Denmark and were mainly university students and staff.

2.7. Instructions

Subjects were given no oral instruction prior to the experiment. Before listening to the first soundscape, the subjects read the following prompt:

> *In a moment you will listen to a number of soundscapes. In some of the soundscapes you can move in the sound scene using the computer's right and left arrow keys. If you hear something on your left side that you would like to move towards, press the left arrow key. Likewise, you can move right by pressing the right arrow key.*

After reading this, the subjects listened to a short demonstration of interactive 3D audio where they were able to move left and right following a speakers voice. Next, they were instructed in the following way:

> *In a moment you will listen to a number of soundscapes. After each soundscape you will be asked a number of questions. If you are in doubt about an answer, please make your best guess.*

Before each i-3D soundscape the following prompt was shown:

> *Please listen carefully to the following sound. NOTE: While you listen, you can move in the soundscape using the RIGHT and LEFT arrow key,*

and before each non-interactive soundscape, the prompt read:

> *Please listen carefully to the following sound. NOTE: You can NOT move in this soundscape.*

2.8. Statistical analysis

To assess the statistical significance of the difference between listeners' subjective evaluation of the listening experience, we pooled the seven questions to-

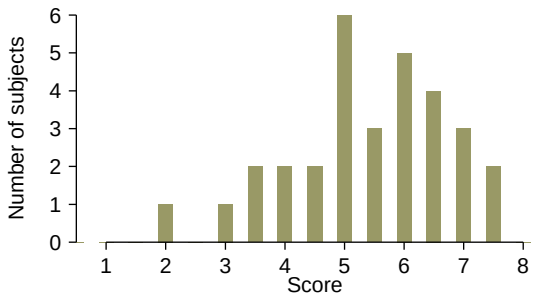

Fig. 2: Distribution of scores on test of listening skills.

gether and computed the average rating on a 0–3 scale as well as the 95 percent confidence interval of the rating. To examine differences between conditions on the questions related to narration and audio environment, we pooled all questions within each category and computed the overall proportion of correct answers as well as the exact 95 percent confidence interval for the proportion. To assess the statistical significance between conditions we compared the proportion of correct answers using Fishers exact test.

3. RESULTS

In the following we go through our main findings. The complete data with the results of the experiments is available for download [5].

3.1. Listening test

The listening test revealed that the listening skills of the subjects prior to the experiment were very diverse, ranging from only two correct answers to an almost perfect score (see Fig. 2.) Based on this, in the following we examine the difference between relatively skilled and unskilled listeners, which we define as listeners scoring respectively above and below the median.

3.2. The listening experience

The results of the analysis of answers to the pool of questions related to the listening experience are shown in Fig. 3. The figure shows the overall average score on a 0–3 point scale as well as separate results for skilled and unskilled listeners. The data show a clear and statistically significant difference between the mono condition and the stereo, 3D, and i-3D

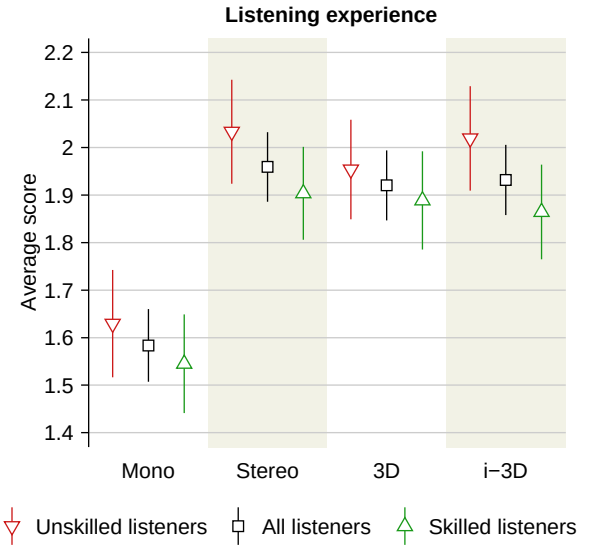

Fig. 3: Average score on a 0–3 point scale on questions related to the listening experience. Mean and 95 percent confidence intervals are indicated.

condition ($p < 0.001$). Differences between the latter three conditions are not statistically significant. For each of the seven questions, the results are similar to the average although with larger confidence intervals. In all conditions, skilled listeners gave a slightly lower score than unskilled listeners.

3.3. Narration and audio environment

The results of the analysis of answers to the two pools of questions related narration and the audio environment are shown in Fig. 4. The figure shows the overall proportion of correct answers in percent as well as separate results for skilled and unskilled listeners.

With respect to the narration, the data show that all subjects were better able to remember specific phrases when listening to the soundscapes in the mono condition compared to stereo, 3D, and i-3D ($p = 0.04$, $p = 0.10$, and $p = 0.08$). Subjects did slightly better in 3D and i-3D conditions compared to stereo, although differences between stereo, 3D, and i-3D are not statistically significant. In all conditions, skilled listeners outperformed unskilled listeners.

For the questions related to the sound environment, subjects performed worse in mono, better in stereo,

better still in 3D, and best in i-3D. Although there seems to be a clear trend, it must be noted that the confidence intervals are too large to make a clear conclusion (for example, comparing mono and i-3D we have $p = 0.18$.) Skilled listeners outperformed unskilled listeners only in the mono, stereo, and 3D conditions, whereas the unskilled listeners did best in the interactive condition.

4. DISCUSSION

When asked to subjectively evaluate the listening experience, subjects express a clear difference in attitude towards mono audio versus stereo, 3D, and i-3D. This is to be expected, since listening to a monaural recording sounds much more flat and dull in comparison. Since the sound reproduction in the 3D and i-3D conditions were identical, it also comes as no surprise that listeners rated these conditions identically. Our initial expectations were, however, that panned stereo would be rated somewhere inbetween mono and 3D; however, the results show that panned stereo is rated equal to 3D, suggesting that the 3D listening experience was not substantially different from the stereo listening experience with only rudimentary spatial cues.

Our main hypothesis was that spatial cues and interaction would improve the listeners ability to discern sound sources leading to better attention to details. Concerning sounds in the audio environment and disregarding narration, our results weakly confirm this hypothesis. Subjects did indeed better remember details in the interactive 3D condition. With respect to remembering details of the narration, the result was that subjects performed significantly better in the mono condition. This could be taken to indicate that the attention of the subjects was more easily distracted by background sounds in the stereo, 3D, and i-3D conditions in accordance with our hypothesis.

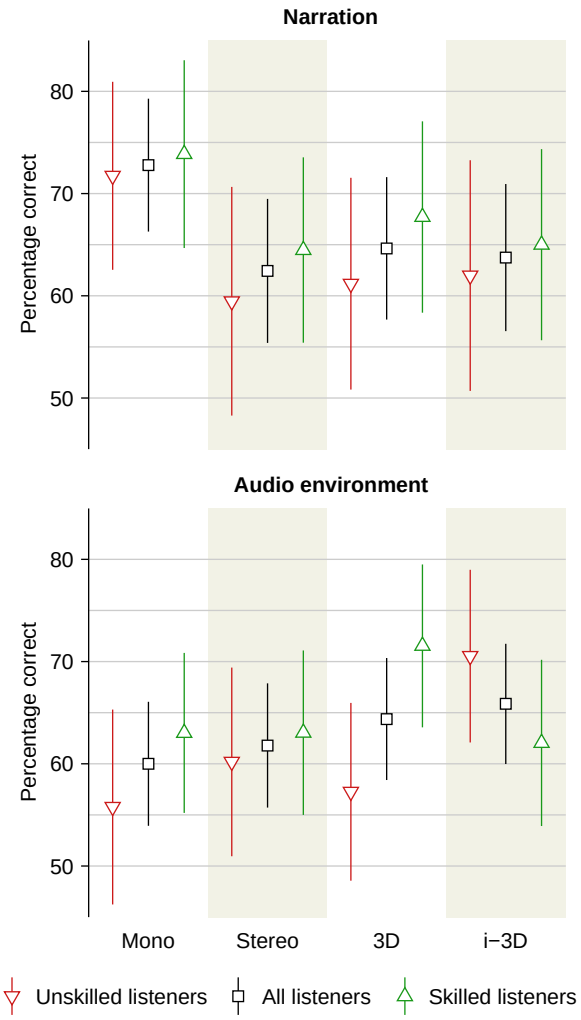

Fig. 4: Fraction of correct answers to questions regarding (a) the narration and (b) the audio environment. Mean and 95 percent confidence interval indicated.

5. ACKNOWLEDGEMENTS

This work was supported by the Network for Danish Sound Technology, `www.soundtechnology.dk`.

6. REFERENCES

[1] A. S. Bregman. *Auditory Scene Analysis*. The MIT Press, 1994.

[2] R. Cusack, J. Deeks, G. Aikman, and R. P. Carlyon. Effects of location, frequency region, and time course of selective attention on auditory scene analysis. *Journal of Experimental Psychology: Human Perception and Performance*, 30(4):643–656, 2004.

[3] J. Darwin and R. W. Hukin. Auditory objects of attention: The role of interaural time differences. *Experimental Psychology: Human Perception and Performance*, 25(3):617–629, 1999.

[4] M. Kleiner, B.-I. Dalenbäck, and P. Svensson. Auralization—An overview. *Journal of the Audio Engineering Society*, 41(11):861–875, November 1993.

[5] M. N. Schmidt, S. Schwartz, and J. Larsen. Extra material for: "Interactive 3D audio: Enhancing awareness of details in immersive soundscapes?". `http://www.imm.dtu.dk/pubdb/p.php?6322`.

[6] B. G. Shinn-Cunningham. Influences of spatial cues on grouping and understanding sound. In *Proceedings of the Forum Acusticum*, 2005.

[7] H. Wallach. On sound localization. *Journal of the Acoustical Society of America*, 10(4):270–274, 1939.