Sound sample detection and numerosity estimation using auditory display

HANNES GAMPER, Department of Media Technology, Aalto University, Finland
CHRISTINA DICKE, Department of Computer Science, University of Canterbury, New Zealand
MARK BILLINGHURST, Department of Computer Science, University of Canterbury, New Zealand
KAI PUOLAMÄKI, Department of Information and Computer Science, Aalto University, Finland

This paper investigates the effect of various design parameters of auditory information display on user performance in two basic information retrieval tasks. We conducted a user test with 22 participants in which sets of sound samples were presented. In the first task, the test participants were asked to detect a given sample among a set of samples. In the second task, the test participants were asked to estimate the relative number of instances of a given sample in two sets of samples. We found that the stimulus onset asynchrony (SOA) of the sound samples had a significant effect on user performance in both tasks. For the sample detection task, the average error rate was about 10% with an SOA of 100 ms. For the numerosity estimation task, an SOA of at least 200 ms was necessary to yield average error rates lower than 30%. Other parameters, including the samples' sound type (synthesised speech or earcons) and spatial quality (multichannel loudspeaker or diotic headphone playback), had no substantial effect on user performance. These results suggest that diotic or indeed monophonic playback with appropriately chosen SOA may be sufficient in practical applications for users to perform the given information retrieval tasks, if information about the sample location is not relevant. If location information was provided through spatial playback of the samples, test subjects were able to simultaneously detect and localise a sample with reasonable accuracy.

Categories and Subject Descriptors: H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems—*Artificial, augmented, and virtual realities*; *Audio input/output*; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*Auditory (non-speech) feedback*; H.1.2 [**Information Systems**]: User/Machine Systems—*Human factors*

General Terms: Experimentation, Human Factors

Additional Key Words and Phrases: Spatial sound, diotic headphone playback, speech synthesis, earcons, SOA

ACM Reference Format:

Gamper, H., Dicke, C., Billinghurst, M. and Puolamäki, K. 2013. Sound sample detection and numerosity estimation using auditory display. ACM Trans. Appl. Percept. 10, 1, Article 4 (February 2013), 18 pages.
DOI = 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

This work was supported by the Helsinki Graduate School in Computer Science and Engineering (HeCSE), the [MIDE program] of Aalto University, the Nokia Research Foundation, the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement no. [203636], Tekniikan edistämissäätiö (TES), and the Finnish Centre of Excellence for Algorithmic Data Analysis Research (ALGODAN). Authors' addresses: H. Gamper, Dept. of Media Technology, Aalto University School of Science, FI-00076 Aalto, Finland; C. Dicke and M. Billinghurst, Dept. of Computer Science, University of Canterbury, New Zealand; K. Puolamäki, Dept. of Information and Computer Science, Aalto University, Finland.

DOI 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

1. INTRODUCTION

Auditory display is the audio analogue of graphical display. McGookin and Brewster [2004b] define it as "the use of sound to communicate information about the state of a computing device to a user". It successfully enhances the usability of mobile devices with a small visual display [Brewster 2002]. Hornof et al. [2010] report on the benefit of auditory feedback in a multimodal task involving a graphical display device with varying peripheral visibility. Unlike a graphical display, an auditory display does not require a stable line of sight and is not hindered by a limited field of view; an auditory display is perceivable regardless of the user's head orientation. Channelling some information to the ears leaves the eyes free to observe the environment and reduces the visual and cognitive loads [Peres et al. 2008]. The auditory system is particularly well-suited for alerting and monitoring because it can ignore expected sounds and rapidly detect unexpected sounds [Shinn-Cunningham et al. 1997]. Response times to auditory stimuli are often lower than to visual stimuli [Nees and Walker 2009]. Furthermore, the auditory system can monitor multiple data streams in parallel. For example, a listener can selectively attend to one speaker in a group of concurring speakers, a phenomenon known as the "cocktail party effect" [Cherry 1953]. In his work on auditory scene analysis, Bregman [1990] reviews how the auditory system processes audio streams.

In this paper we investigate the use of an auditory display in information retrieval tasks. We conducted a user study to assess the effect of various display parameters on user performance in two tasks adapted from information visualisation and human–computer interaction research: I) The detection of a certain element in the presence of distractor elements, and II) the estimation of the percentage of elements with a certain characteristic [Treisman 1986; Julesz and Bergen 1987; Healey et al. 1996; Michalski and Grobelny 2008]. Here we investigate how effective various configurations of an auditory display are when performing these basic tasks. In our study, the elements consisted of short sound samples, presented to the user as a concatenated list of samples.

In analogy to a detection task used in information visualisation research, task I of the user experiment consisted in detecting a specific sound sample, referred to in this paper as the <key> sample, among distractor samples. Sagi and Julesz [1985] argue that in human vision, detection and localisation of target elements can be done in parallel. To test whether the same holds for auditory display, i.e., whether users are able to determine the position of the <key> sample once detected without listening to the stimulus again, we asked the test participants to indicate the perceived direction of the <key> sample. Task II of the user study is an adaptation of the percentage estimation task used in information visualisation research. The task tested the ability to estimate the relative number of <key> sample instances among distractor samples. The display of a large number of samples with fixed stimulus onset asynchrony (SOA) results in a long trial duration and increases the effect of participant memory and fatigue. To keep the trial duration low, we tested a small number of samples and asked test subjects to give a relative estimate of the numerosity of <key> samples by comparing two sets of samples and indicating which set contained more instances of the <key> sample. By varying the relative difference between the number of instances in both sets, we could determine response thresholds (RT) for the detection of a relative difference. These two tasks provide valuable data that can be used to improve the design of audio interfaces for information display. Factors affecting the detection rate of a <key> sample need to be taken into account when presenting lists or menu items to a user via audio, for example for hands-free audio interfaces or screen readers. When presenting data as audio samples it is interesting to know whether users will be able to estimate numerosity, and which factors affect this ability.

The parameters studied here are derived from the steps necessary to design an auditory display. First, the data or information to be displayed needs to be represented as audio. In our test, we compared

the effectiveness of two types of audio representations that are well-researched for their usage in auditory displays: non-speech sounds called earcons, i.e., "abstract, synthetic tones that can be used in structured combinations" [Brewster et al. 1993], and synthesised speech. Next, the information needs to be arranged in time for playback. We studied the impact of the stimulus onset asynchrony (SOA), i.e., the delay between subsequent sound sample onsets, on user performance. Finally, the information needs to be displayed aurally. In our research we were interested in the effect of spatial separation in an auditory display on user performance. To assess this effect we compared two playback systems: a multichannel loudspeaker system allowing spatial auditory display, and a headphones setup providing nonspatial auditory display. Hence, we cover all the steps necessary for an audio interface design; information representation, arrangement in time, and audio playback.

2. RELATED WORK

In studying the display of auditory information our work builds on prior research on auditory display. To convey information to a user via auditory display, the information must be made audible. This process is straightforward if the original information source is acoustic, including speech or music. To display nonacoustic information, including text, interface elements, or other data, the information must be encoded as acoustic signals. Researchers have proposed various approaches for this process.

An obvious choice to represent information on an auditory display is *speech*. Speech is used in public announcement systems, or to communicate absolute data values [McGookin and Brewster 2004a]. A screen-reader is an application that verbalises information displayed on a computer screen. Screen readers allow blind and visually impaired people to access verbal information effectively [Nees and Walker 2009]. The generation of speech output can be fully automised through text-to-speech synthesis, and its meaning can be readily interpreted by users. However, using speech to display information may be quite time-consuming due to its sequential and transient nature [Sawhney and Schmandt 2000]. *Sonification* describes the process of rendering audible numerical data by mapping it to acoustic parameters [Peres et al. 2008]. This method has been successfully applied to the exploration of complex quantitative data via auditory graphs [Nees and Walker 2009; Brown et al. 2002]. Further examples of auditory information displays are *alarms and alerts*, which are short, unsubtle sounds designed to capture the user's attention and indicate that action is required [Nees and Walker 2009]. Icons are a popular way of displaying information in user interfaces. An icon establishes a metaphorical mapping between information and its representation. A classic example is the virtual trashcan icon in graphical user interfaces holding deleted items, in an analogy to a real paper bin. *Auditory icons* employ metaphors to map sounds to their virtual referents [Gaver 1986]. For example, the sound of crinkling paper is associated with emptying a (virtual) trash bin [Peres et al. 2008]. If the mapping between information and sound in an auditory icon is intuitive, it should be relatively easy for the user to learn and remember [Gaver 1986]. On the downside, such a mapping is not always available, leading to limitations in the use of auditory icons. *Earcons* are non-speech sounds designed to offer more flexibility than auditory icons or simple alarms, and were originally introduced by Blattner et al. [1989]. Instead of synthetic tones, musical instruments often serve as the basis of earcons due to their rich harmonic structure [Brewster et al. 1995]. Unlike auditory icons, the abstract nature of earcons allows them to be assigned to any item or process [Nees and Walker 2009]. A major advantage of earcons is the ability to represent hierarchical structures by mapping a node's position to sound parameters, including rhythm, timbre, or pitch [Brewster et al. 1995]. Because the semantic association between earcons and the information they represent is arbitrarily attributed, they must be learnt by the user [Garzonis et al. 2009].

For the information retrieval tasks in our study, we chose synthesised speech and earcons as audio encoding strategies. Both sound types are established and actively researched for their use in audi-

tory display and user interfaces. Prior work studied the combination of speech and earcons in a user interface. Karshmer et al. [1994] investigated how to speed up a speech-based interface via the use of earcons. Ramloll et al. [2001] found that combining speech output and non-speech feedback in the form of simple musical tones was beneficial for the accessibility of tabular numeric information. Vargas and Anderson [2003] found that adding earcons to a speech-only menu improved navigation performance of users. Walker et al. [2006] compared menu navigation performance of several auditory representations, including speech-only and the combination of speech and earcons. In their study, users were able to navigate faster and more accurately with the speech-only representation than with the combination of speech and earcons.

Unlike the aforementioned studies, we compared the effectiveness of speech and earcons separately. Previously, Tran et al. [2000] compared speech and non-speech beacons in a navigation task. In their study, users found non-speech sounds easier to localise and more pleasant than speech. Dingler et al. [2008] compared a variety of sound cues for environmental features, including speech and earcons, in terms of their learnability. They found that, on average, earcons required between eight and nine training cycles before they could be correctly identified by test participants, whereas very few participants needed more than one training cycle to correctly map all speech cues to their visual representations. Bonebright and Nees [2009] compared speech and earcons in a dual attention task. In their test, participants had to map a sound sample to its visual representation on a screen, whilst listening to a speech passage. Different audio encoding strategies were compared in terms of their performance, including speech and earcons. Speech playback led to the highest mean number of correct responses, and to the lowest mean response time. However, after some practice time, earcons led to relatively comparable accuracy. In our study, earcons and speech were compared in two simple information retrieval tasks. As shown in the works by Dingler et al. [2008] and Bonebright and Nees [2009], speech playback outperforms non-speech playback in tasks involving mapping sound cues to their textual representations. In our study, no visual display was used, hence test participants were not required to map sound cues to their visual representations. Furthermore, each test participant was required to concentrate on only one <key> sample throughout the whole test, to minimise the effects of learnability or memory on earcon performance identified in prior studies. Therefore, the earcons used in this study serve mainly as a non-speech alternative to speech output. The specific characteristics of earcons as an encoding strategy for auditory display have been studied extensively elsewhere [Brewster et al. 1993; 1995; McGookin and Brewster 2004a], and are not considered here.

Earlier work on the presentation of concurrent earcons and speech samples suggests that user performance deteriorates as the number of distractors or maskers increases [Brungart et al. 2001; Brungart et al. 2002; McGookin and Brewster 2004b]. In these studies, the number of maskers used was between one and three. In our work, the number of concurrent samples was determined by the samples' temporal overlap and parametrised via the stimulus onset asynchrony (SOA). Here, up to eleven samples were played back simultaneously. In studies by Brungart et al. [2001] and Brungart et al. [2002], users were required to extract information by attending an audio stream. Rather than attending a stream, our test participants listened to short sound samples to perform basic information retrieval tasks. McGookin and Brewster [2004b] found that using a stimulus onset asynchrony (SOA) of 300 ms increased earcon identification performance. Based on a pilot study and the findings of McGookin and Brewster [2004b], we hypothesised that a large SOA would have a positive effect on user performance in both tasks of our user study. To study the effect in more detail, rather than testing a single SOA, we tested values spanning a range of SOAs that we found critical for user performance in a pilot study. Figure 1 illustrates two examples of audio sample playback with SOA Δt. An SOA of 50 ms leads to a dense presentation of short, partially overlapping sound samples, with up to eleven samples played

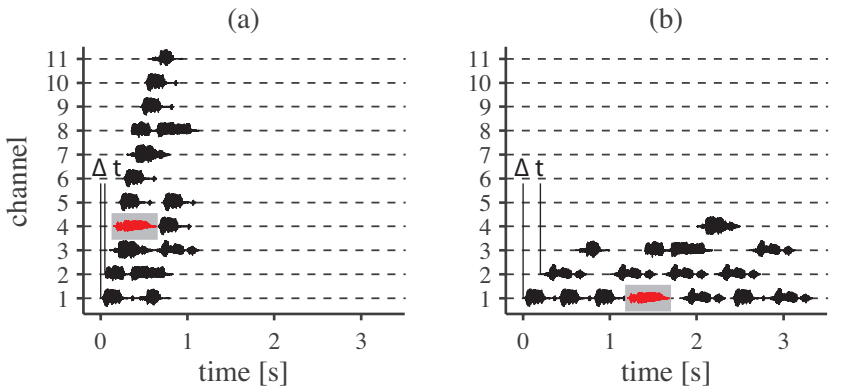

Fig. 1. Audio sample playback with (a) 50 ms and (b) 200 ms stimulus onset asynchrony (SOA) Δt. The two displayed sets contain 15 samples, with the <key> sample highlighted and marked red. The samples are distributed over the minimum number of channels such that no temporal overlap occurs within any channel, to illustrate the temporal density of the samples. The channels do not correspond to playback channels.

back concurrently. With an SOA of 200 ms, the maximum number of concurrently presented samples is reduced to four.

The samples displayed in Figure 1 were played back diotically over headphones or from randomised directions via a multichannel loudspeaker system. Research on spatial release from masking (SRM) and the cocktail party effect indicates that spatial separation of concurrent sounds improves user performance in various tasks [Ihlefeld and Shinn-Cunningham 2008a; 2008b; Brungart and Simpson 2002; McGookin and Brewster 2004a; Bronkhorst 2000]. For speech signals, however, the SRM depends on a variety of factors, including the masker type and the spatial configuration of the target and the maskers [Kidd et al. 2010], in addition to a priori information on the target direction ("knowing where to listen") [Brungart et al. 2002; Kidd et al. 2005]. Furthermore, monaural cues, including vocal characteristics, prosodic features, and level differences between target and maskers, affect speech perception in the presence of maskers [Bregman 1990; Darwin and Hukin 2000; Brungart et al. 2001]. Therefore, predicting the impact of spatial separation on user performance in the information retrieval tasks in our user study is difficult. However, we expected a priori that spatial separation of the samples would result in improved user performance in both tasks of our study.

3. EXPERIMENTAL DESIGN AND PROCEDURE

We conducted a listening test to determine the impact of various design parameters of auditory display on user performance in two basic information retrieval tasks: I) detecting a <key> sample among a set of distractor samples and II) estimating the relative quantity of <key> sample instances in two sets. The sample sets were presented to the users as a list of prerecorded sound samples staggered with a stimulus onset asynchrony (SOA). Both tasks were performed under a number of test conditions, differing in their combination of the levels of the four independent variables: the playback setup and sound type used for the auditory display, the SOA, and the relative difference between the <key> sample instances (task II). We studied the effect of these parameters on user error rates. Figure 2 gives a schematic overview of the experimental design.

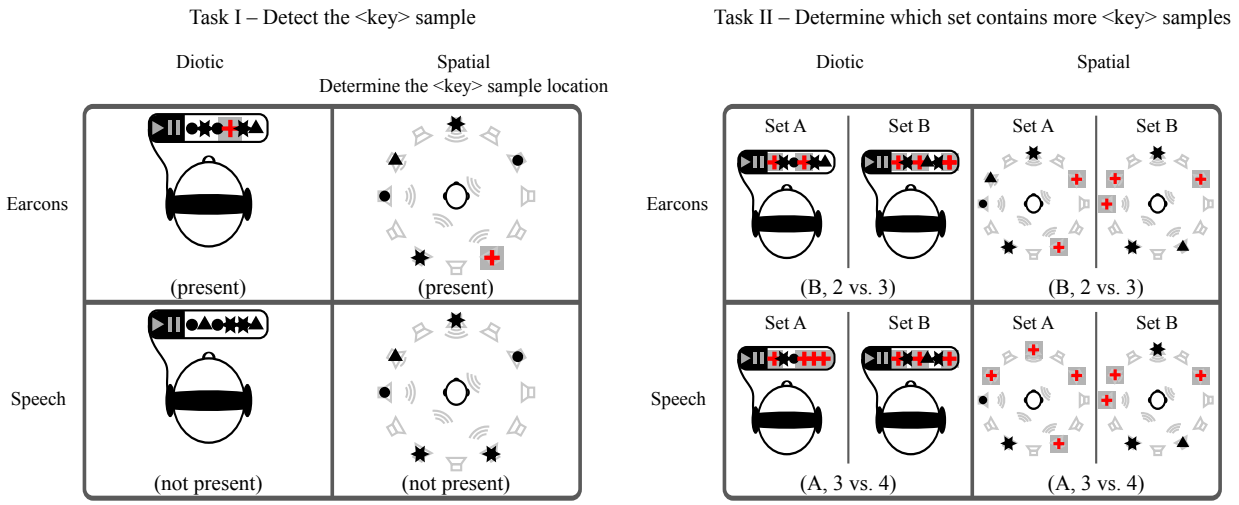

Fig. 2. Experimental setup with examples of sample sets used in the study. The highlighted red cross represents the <key> sample; all other symbols represent distractors. In task I, each set contained a total of 15 samples staggered with stimulus onset asynchrony (SOA) of 50, 100, 200, or 400 ms. In task II, each set contained two to seven instances of the <key> sample, totalling 10 or 20 samples staggered with an SOA of 100, 200, or 400 ms.

3.1 Test conditions

During the listening test, each test subject completed two tasks. In task I, test subjects were presented with a set of 15 sound samples in each trial. The test subjects were asked to detect a specific <key> sample in a set of sound samples. For the spatial loudspeaker playback, test subjects were asked to state which loudspeaker they thought the <key> sample was being played from. We expected users to be able to recall the direction of a <key> sample, once detected, in analogy with findings related to human vision [Sagi and Julesz 1985]. For the diotic headphone playback, this subtask was omitted, as the samples were not separated spatially. In task II, two sets of samples were presented to the test subjects. Each set contained a total of 10 or 20 samples. Between two and seven of the samples in each scene were instances of the <key> sample. We tested a small subset of all possible combinations of the number of instances in each set, to keep the test duration low. Users were asked to determine whether the number of <key> sample instances in the two sets was the same, or, if it differed, which set contained more instances. We expected to observe an effect of the relative difference between the number of <key> sample instances in the two sets on user performance.

Each set consisted of short synthesised speech or earcon samples, concatenated to staggered sequences with a fixed SOA. A small SOA speeds up playback but increases the overlap between sound samples. In the experiment, we compared the effect of 50 (task I), 100, 200, and 400 ms SOA on user performance. This critical SOA range was determined for the given tasks and sound samples in a pilot test.

The audio playback of two different setups was compared: spatialised audio using a multichannel loudspeaker system, and nonspatial audio via headphone playback. For the spatial audio playback, each sample in a set was played from a randomly selected loudspeaker, thus spreading the set elements evenly along a circle around the listener. The physical separation of the loudspeakers ensured accurate reproduction of localisation cues for each sample and thus maximised the possible benefit of spatial separation of the samples along a circle in the horizontal plane. Nonspatial playback was implemented

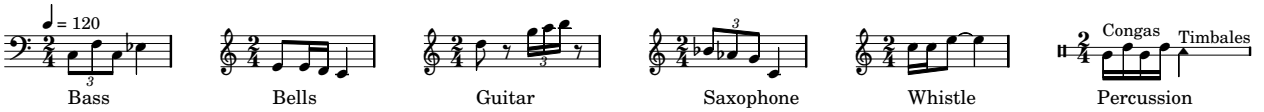

Fig. 3. Earcons used in the user study. The timbres were produced via OSX's GarageBand inbuilt MIDI instruments.

by presenting the sound samples diotically to both ears, i.e., feeding the same signal to the left and right headphone. This setup ensured a controlled playback of all samples and minimised the influence of head movements on the sample perception.

3.2 Apparatus and sound samples

The study was conducted in a multi-purpose research space with a wideband reverberation time of about 0.3 s and a direct-to-reverberant ratio of over 30 dB, calculated as the energy ratio between the first three milliseconds of the impulse response to the reverberation tail [Larsen et al. 2008]. Given these acoustic specifications, we are confident that the effect of room reverberation on user performance is negligible. The multichannel loudspeaker system used in the test consisted of 12 Genelec 1029A loudspeakers, arranged in a circle of 5 m radius at 30 degree intervals. For headphone playback, we used Sennheiser HD 212 Pro and AKG K66 headphones. The sound samples were played back diotically, at equal loudness levels, and without artificial reverberation or spatialisation.

Two different sound types were compared in the study: synthesised speech and earcons. For the speech samples, we selected a set of short, common english words representing objects that test subjects would be familiar with. The speech samples were obtained by synthesising the words "book", "chair", "keys", "microwave", "couch", and "cup" via the Mac OS X's inbuilt speech synthesiser using the male voice "Alex". The durations of the resulting sound samples were 0.4 to 0.8 s. Earcons were chosen as a non-speech alternative to speech output. A total of six earcons was used. The earcons differed in timbre, melody, and rhythm, and were designed to be recognisably different, as suggested by Brewster et al. [1995]. The earcons were generated via Mac OS X's GarageBand using the inbuilt MIDI instruments with the following timbres: "Bass", "Bells", "Guitar", "Saxophone", "Whistle", and "Percussion". The duration of the earcons was one second each. A transcription of the earcons is shown in Figure 3. To ensure equal loudness, the signal levels of the speech and earcon samples were normalised using A-weighting.

3.3 Test procedure

The study was laid out in a fully randomised within-subject design. Twenty-two subjects (six female), aged 19 to 43, participated in the study. No participant reported any hearing impairments. The duration of the experiment was about 90 minutes per subject. Upon completion, each test subject was compensated with a movie ticket. The test subjects were seated in the centre of the circle of loudspeakers during the entire test. User data were collected via a questionnaire that the test participants filled during the experiment. All test participants have given their written consent to use the data collected during the experiment for scientific research. The experiment was organised in two sets of four rounds, each round having a different combination of task, playback setup, and sound type. To minimise learning effects, the order of the test rounds was randomised. At the beginning of each round, the test subjects were introduced to the task and to the <key> sample. One earcon and one speech sample were assigned to each test participant and used as the <key> sample throughout the test, hence each participant had to concentrate on only two samples (i.e., one of each sound type) during the entire experiment. In task I, test subjects were asked to detect the presence of the <key> sample among a set of distractor samples. In task II, tests subjects were asked to determine which of two sets

of samples contained more instances of the <key> sample. One test round consisted of 20 (task I) or 30 (task II) trials. Each set was played as a sequence of speech or earcon samples staggered in time, spatially separated through a multichannel loudspeaker setup or diotically via headphones, depending on the test condition. The range of the test parameters was chosen based on a pilot study.

3.4 Analysis of the results

Each response of the test participants was categorised as either "correct" or "incorrect". For each tested condition, the total error count was calculated. The error count data were summarised in a contingency table, with columns representing different test categories. Pairwise Pearson's chi-squared tests were performed on adjacent columns of the table to test the null hypothesis that there is no association between the test category and the error count. Although the layout of our user study in a within-subject design would favour analysing the data via McNemar's chi-squared test [Sheskin 2000], we found reporting the data as error counts per category more illustrative and therefore chose Pearson's chi-squared test. From here on, we refer to Pearson's chi-squared test simply as chi-squared test. If more than two columns were compared, the Holm-Bonferroni correction was applied to p-values [Holm 1979; Dudoit et al. 2003]. A repeated-measures analysis of variance was performed to check for second-order interaction effects between the independent variables. In task I, the localisation performance of test participants was calculated as the root mean squared error (RMSE) of the estimated lateral angle in each trial. A one-way analysis of variance (ANOVA) was performed to test the null hypothesis that the RMSE is equal across lateral angles, stimulus onset asynchronies (SOAs), and sound type of the samples.

4. TASK I: DETECT THE <KEY> SAMPLE

4.1 Results

The <key> sample was present in 80% of all trial sets. Because test subjects were unaware of the correct answer distribution, guessing whether or not the sample was present would result in an error rate of 50%.

The results show that error rates decrease as the stimulus onset asynchrony (SOA) is increased. A chi-squared contingency table test of error count data reveals the effect to be statistically significant at an alpha level of 0.05 ($X^2(3, N = 1760) = 119.35$, $p < 0.001$, $V = 0.26$). Pairwise comparisons of each SOA with the next highest SOA show a statistically significant error rate decrease for each SOA increase (Table I). The Holm-Bonferroni correction adjusts the p-values of multiple comparisons [Holm 1979; Dudoit et al. 2003].

The average error rate as a function of playback condition differs by 3%, with the diotic headphone playback yielding slightly better results (Table I). Although the difference is statistically significant according to a chi-squared test of error count data ($X^2(1, N = 1760) = 4.04$, $p = 0.044$, $V = 0.050$), the association between the two independent variables is weak according to Cramér's V.

The average error rate as a function of sound type differs by 5%, with earcons performing slightly better (Table I). A chi-squared test indicates that the difference is statistically significant ($X^2(1, N = 1760) = 9.57$, $p = 0.002$, $V = 0.076$). The "bass" earcon, a short melody with a bass timbre (see Figure 3), was correctly identified in all trials.

To check for second-order interaction effects between the independent variables SOA, playback condition, and sound type, a repeated-measures analysis of variance is performed. Figure 4a depicts the average error rates of both playback conditions across the tested SOAs. A repeated-measures analysis of variance with Greenhouse-Geisser correction for departure from sphericity does not indicate a statistically significant interaction between the playback condition and the SOA ($F(2.53, 275.22) =$

Table I. Error count table for Task I. Chi-squared tests indicate statistically significant differences between all adjacent columns of the independent variables. V denotes Cramér's V.

	Stimulus onset asynchrony (SOA) [ms]				**Playback**		**Sound type**	
	50	100	200	400	Diotic	Spatial	Earcon	Speech
Trials	440	440	440	440	880	880	880	880
Errors	104	46	26	11	80	107	73	114
Errors [%]	24	10	6	2	9	12	8	13
p-value	<0.001	0.037	0.037		0.044		0.002	
V	0.175	0.083	0.085		0.050		0.076	

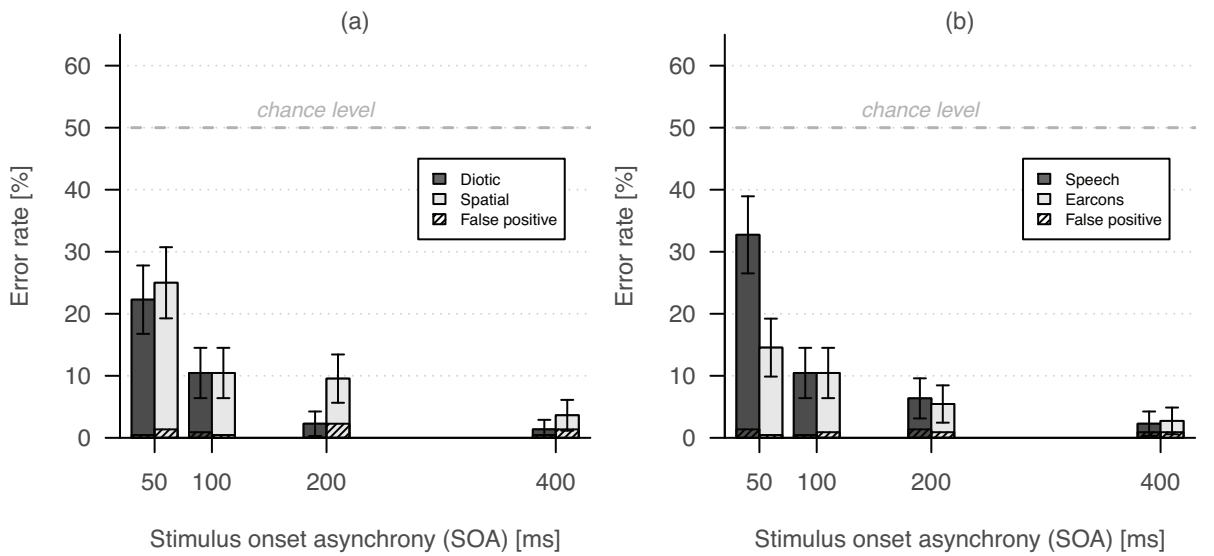

Fig. 4. Average error rates as a function of (a) playback condition and (b) sound type for different stimulus onset asynchronies (SOAs). The hatched area indicates false positive errors. Error bars indicate 95% confidence intervals for the means.

$1.91, p = 0.1389$). Figure 4b depicts the average error rates of both sound types across the tested SOAs. A repeated-measures analysis of variance with Greenhouse-Geisser correction for departure from sphericity indicates a statistically significant interaction between the sound type and the SOA ($F(2.44, 265.83) = 10.53, p < 0.001$). A repeated-measures analysis of variance does not indicate a statistically significant interaction between the playback condition and the sound type ($F(1, 109) = 3.74, p = 0.0558$).

False positive errors, occurring when a user indicated that the <key> sample was present in the set when it was not, were below 5% on average and committed only by 7 out of 22 participants (Figure 4, hatched area).

For the spatial loudspeaker playback, the test subjects were asked to state from which loudspeaker they thought the <key> sample was being played. To calculate the angle mismatch of the answers, both actual and perceived directions were mapped to lateral angles between -90 and 90 degrees. This mapping eliminates front–back reversals occurring when a subject perceives a sample played from the back to originate from the front and vice versa. Figure 5 shows the root mean squared error (RMSE) of the direction estimates, calculated from the difference between the actual and the perceived lateral angle of a correctly detected <key> sample. For false positives no RMSE was calculated. The RMSE is

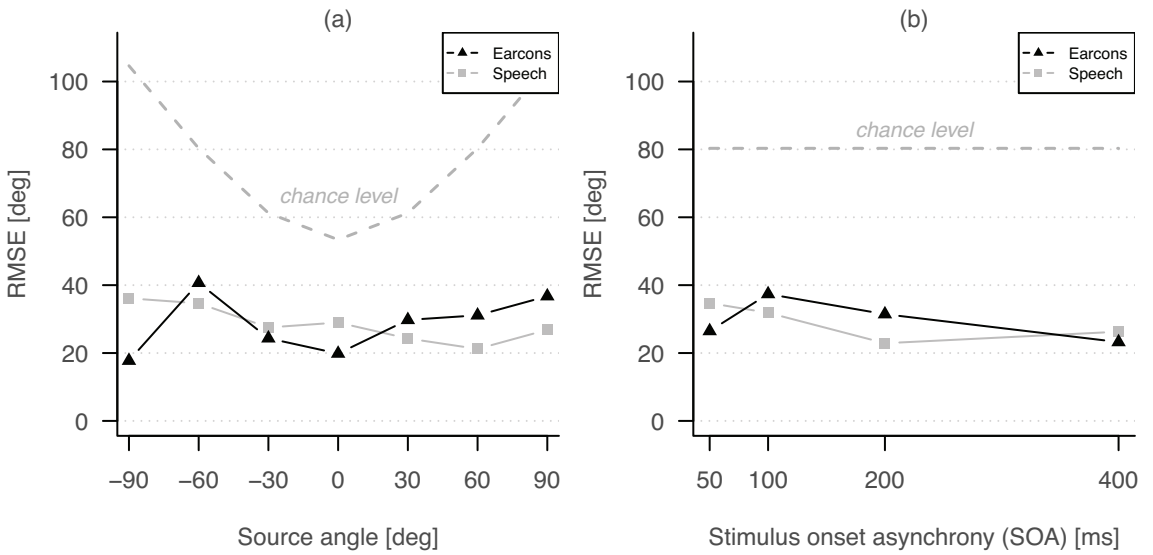

Fig. 5. Root mean squared error (RMSE) of speech and earcon playback as a function of (a) the lateral angle of the <key> sample and (b) the stimulus onset asynchrony (SOA).

well below chance level. A visual inspection of the graphs in Figure 5 indicates that the RMSE does not differ substantially between earcon and speech playback, and that there is no visible dependency from the lateral angle of the <key> sample (Figure 5a) and the SOA (Figure 5b). A one-way ANOVA with RMSE as the dependent variable indicates that there is neither a statistically significant difference between earcons and speech across lateral angles ($F(1, 12) = 0.001, p = 0.982$, cf. Figure 5a) or across SOAs ($F(1, 6) = 0.029, p = 0.87$, cf. Figure 5b), nor between lateral angles ($F(6, 7) = 0.862, p = 0.564$, cf. Figure 5a) or SOAs ($F(3, 4) = 1.629, p = 0.317$, cf. Figure 5b).

4.2 Discussion

The error rates decrease significantly with each SOA increase. This finding follows our hypothesis: increasing the SOA reduces the temporal overlap between consecutive samples and thus the total number of concurrently presented samples, improving user performance at the cost of overall playback duration. With a minimum SOA of 100 ms, the average error rate for determining the presence of the <key> sample drops to about 10% for all test conditions (cf. Figure 4). This finding implies that detection rates are relatively high even with a dense temporal presentation of the samples.

Contrary to our hypothesis that spatial separation of samples would improve their detectability, we found no substantial difference in the user performance between spatial loudspeaker and diotic headphone playback. Earlier work has shown that spatial separation enhances the perception of concurrent sound sources. However, in our test the separation did not have a substantial effect on user performance for detecting the <key> sample. Unlike the classical cocktail party situation, where the listener attends to sound coming from a certain direction, the direction of the <key> sample was randomised in our experiment and thus not known by the user beforehand. The lack of a priori information on the target direction seems to have cancelled the advantage of spatial separation, which follows the findings of Brungart et al. [2002] and Kidd et al. [2005]. The result that detectability rates were not affected by spatial separation may be important for practical applications, where spatial separation of sound samples can be difficult or costly to implement.

Table II. Error count table for Task II. Chi-squared tests indicate statistically significant differences between all adjacent columns of the independent variables, except for "playback type", i.e., diotic headphone and spatial loudspeaker playback. V denotes Cramér's V.

	SOA [ms]			**Difference [%]**					**Playback**		**Sound type**	
	100	200	400	33 (3 vs. 4)	50 (2 vs. 3)	67 (3 vs. 5)	100 (3 vs. 6)	133 (3 vs. 7)	Diotic	Spatial	Earcon	Speech
Trials	880	880	880	528	528	528	528	528	1320	1320	1320	1320
Errors	433	238	103	254	215	131	103	71	385	389	412	362
Errors [%]	49	27	12	48	41	25	20	13	29	29	31	27
p-value	<0.001	<0.001		0.037	<0.001	0.045	0.030		0.898		0.036	
V	0.228	0.194		0.074	0.169	0.064	0.082		0.003		0.042	

The user performance did not differ substantially between synthesised speech and earcon playback. The results indicate that the presence of one short speech sample among other samples can be determined with similar accuracy as the presence of a certain earcon among other earcons, with the exception of the "bass" earcon, for which the error rate was 0%, presumably due to low frequency content not present in other earcons. A repeated-measures analysis of variance indicates a statistically significant interaction between the sound type and the SOA. The interaction effect is visible in Figure 4b, where at an SOA of 50 ms earcons seemed to outperform speech playback. The superior performance of earcons might indicate that the users' auditory system could not cope with the multitude of overlapping speech samples as well as with overlapping instrumental sounds. However, more data would be required for a thorough investigation of this effect. Another factor affecting performance may be that the earcons varied in terms of timbre, rhythm, and melody (cf. Figure 3), whereas all speech samples were generated using the same synthesiser parameters. Further improvement of the detectability rates for speech playback may be achieved by varying vocal characteristics, including gender, pitch, vocal tract size, accent, or speaking style of the speech samples. For practical applications, our results indicate that with a minimum SOA of 100 ms, synthesised speech achieves similar detectability rates as non-speech audio, whilst offering the advantages that it does not require users to learn or memorise samples and that it can be generated automatically with standard software tools.

With a false positive error rate of less than 5%, test subjects were more likely to answer "not present" than "present" when doubting the presence of the <key> sample. Conversely, if test subjects indicated the <key> sample to be present, they were correct in over 95% of the cases.

Neither the lateral angle of the <key> sample, nor the sound type, nor the SOA had a substantial effect on the localisation performance of test subjects. After the <key> sample had been detected by the test subjects, its approximate location would be implicitly known, with an average RMSE of about 30 degrees, which may be sufficient for certain practical applications where sample location needs to be communicated to the user at low additional cognitive cost.

5. TASK II: DETERMINE WHICH OF TWO SETS CONTAINS MORE INSTANCES OF THE <KEY> SAMPLE

5.1 Results

The distribution of <key> samples was randomised, with set A containing more instances of the <key> sample in 50% of the cases and set B containing more instances in the remaining 50% of the cases. Test subjects were not aware of the distribution. Therefore, guessing would result in a 67% error rate.

The error rates decrease with increasing SOA. A chi-squared contingency table test of error count data reveals the effect to be statistically significant at an alpha level of 0.05 ($X^2(2, N = 2640) = 301.88$, $p < 0.001$, $V = 0.338$). Pairwise comparisons with Holm-Bonferroni correction show a statistically

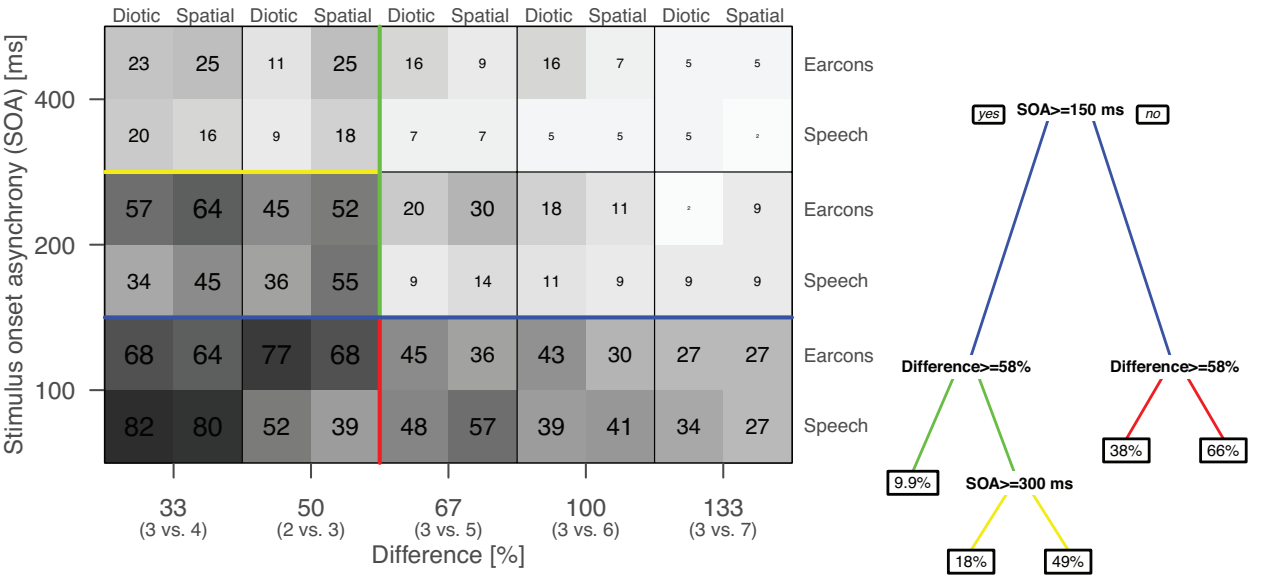

Fig. 6. On the left, error rates (in per cent) as a function of stimulus onset asynchrony (SOA), difference between the number of <key> samples, playback condition (diotic headphone or spatial loudspeaker playback), and sound type (speech or earcons). The <key> sample numerosity is expressed as a relative difference in per cent. Pure guessing would give an error rate of 67%. Table III gives the covariates used to train the regression tree on the right. The leaf nodes indicate average error rates.

significant error rate decrease for an increase in SOA from 100 ms to 200 ms and from 200 ms to 400 ms (Table II and Figure 6).

A similar relationship holds between error rates and the numerosity of the <key> samples. For the analysis, the numerosity is expressed as a relative difference in per cent. For example, 3 vs. 4 <key> samples in two sets to be compared corresponds to a relative difference of 33%. The effect of the relative difference on user performance is statistically significant ($X^2(4, N = 2640) = 216.94$, $p < 0.001$, $V = 0.287$). Pairwise comparisons with Holm-Bonferroni correction reveal a statistically significant error rate decrease for each increase of the relative difference between the number of <key> samples (Table II).

The average error rates under both playback conditions, i.e., diotic headphone and spatial loudspeaker playback, are equal (Table II). A chi-squared error count test does not indicate a statistically significant difference ($X^2(1, N = 2640) = 0.02$, $p = 0.898$, $V = 0.003$).

The average error rates as a function of sound type differ by 4%, with speech performing slightly better (Table II). Although a chi-squared test indicates that the difference is statistically significant ($X^2(1, N = 2640) = 4.39$, $p = 0.036$, $V = 0.042$), the association between the two independent variables is weak according to Cramér's V.

A repeated-measures analysis of variance with Greenhouse-Geisser correction for departure from sphericity indicates a statistically significant interaction between the SOA and the relative difference ($F(6.5, 280.1) = 5.77, p < 0.0001$), and between the SOA and the playback setup ($F(1.6, 68.6) = 5.37, p = 0.0111$). The repeated-measures analysis of variance does not indicate any of the other second-order interactions between the independent variables SOA, relative difference, playback setup, and sound type to be statistically significant.

For further insight into the data, we estimated the error rates using the regression tree analysis and the R rpart library [Therneau et al. 2011] with default settings and a minimum of 200 observations

Table III. Covariates of regression tree analysis.

Test setup	**Test condition**	**Demographic data**	**Musical background**
Group	Playback setup	Gender	Plays instrument
User	Sound type	Age	Can read musical notation
	Stimulus onset asynchrony (SOA)	Native English speaker	Plays in band
	Relative # of <key> objects	Hearing Impairments	Uses 3-D sound
	Total # of objects		Has Perfect Pitch

per terminal leaf node. We used the variables shown in Table III as covariates. Besides test conditions, the covariates contain demographic data and information about the musical background of the test participants. The regression tree analysis confirms that the SOA and relative difference are the main determinants that affect user performance; the effect of demographic data, musical background, sound type, and playback condition was pruned out. Figure 6 displays the resulting regression tree and highlights the areas in the covariate space identified by the regression tree. For small relative differences (33% and 50%) and an SOA of 100 ms, the average error rate is 66%, which corresponds to random chance (Figure 6, bottom left). A larger relative difference or SOA improves performance substantially. With an SOA of at least 200 ms and a relative difference of at least 67%, the average error rate drops below 10% (Figure 6, top right).

To illustrate the main determinants of user performance discussed above, we collapsed the data across dimensions pruned out by the regression tree, including spatial quality and sound type of the samples. Figure 7 relates the remaining variables, SOA and relative difference, to the user performance. The user performance is given as the percentage of correct answers in each test condition. Due to the statistically significant differences for different SOAs described above, we grouped the data by SOA. Borrowing a concept from psychophysics, we fitted a psychometric function to the data for each SOA. In a detection or discrimination task, a psychometric function relates the subject's response to a physical stimulus [Klein 2001], and allows deriving a response threshold (RT), i.e., the stimulus strength required to achieve a certain performance level [Wichmann and Hill 2001]. Figure 7 depicts psychometric functions for the perception of relative differences of the number of <key> samples in two sets, for the three tested SOAs. The functions are fitted via the Matlab psignifit toolbox [Wichmann and Hill 2001]. The lower asymptote of the psychometric function is defined by the "guessing rate", i.e., the performance rate achieved with pure guessing [Treutwein and Strasburger 1999]. Here, it lies at 33%, because test subjects were presented with three possible answers ("Set A contains more <key> samples", "Set B contains more <key> samples", "Both sets contain the same amount of <key> samples"). The upper asymptote of the psychometric function is defined by the "lapsing rate", i.e., the rate at which test subjects lapse even at stimulus levels so high that the sensory mechanism is assumed to be perfect [Treutwein and Strasburger 1999]. The lapsing rate is constrained within the range [0, 6] per cent [Wichmann and Hill 2001] and estimated separately for each psychometric function via a maximum-likelihood search using the psignifit toolbox. Goodness-of-fit tests of each fitted psychometric function via the psignifit toolbox do not indicate a lack of fit, thus the best-fitting cumulative Gaussian fit, shown in Figure 7, can not be rejected as underlying function F for the data. A commonly used definition for RT is the stimulus strength required to produce a probability correct halfway up the psychometric function [Klein 2001]. Using this definition, RT can be calculated as RT$= F_{0.5}^{-1}$, i.e., the inverse of the psychometric function F at a performance level of 50% [Wichmann and Hill 2001]. Table IV shows the values of $F_{0.5}$ and RT, as well as guessing and lapsing rates, for each fitted psychometric function. No RT could be determined for 400 ms SOA, as the whole psychometric function lies above $F_{0.5}$.

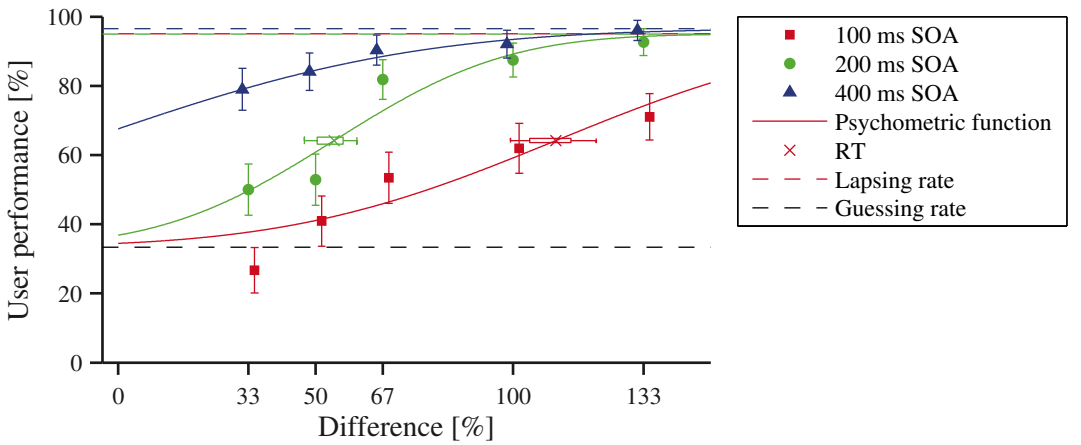

Fig. 7. Task II: Psychometric functions for the SOAs tested. The abscissa represents the relative difference in the number of <key> samples in per cent; the ordinate represents the percentage of users' correct answers. Markers and vertical error bars indicate the mean user performance and 95% confidence intervals of the mean. The psychometric functions are approximated via a cumulative Gaussian fit, using the Matlab psignifit toolbox. Horizontal boxes and error bars indicate bootstrap confidence limits for RT (box = 1 standard deviation, bars = 2 standard deviations), obtained from 1999 bootstrap simulations.

Table IV. Psychometric function parameters (cf. Figure 7).

SOA [ms]	guessing rate [%]	$F_{0.5}$ [%]	(100−lapsing rate) [%]	RT [%]
100	33	64	95	111
200	33	64	95	54
400	33	65	97	-*

*RT lies outside the test range.

5.2 Discussion

The effect of the test conditions on user performance in task II is similar to the findings of task I. As expected, each SOA increase significantly improved performance for comparing the number of <key> samples in two sets. Increasing the SOA reduced the temporal overlap of samples and thus made the recognition of individual samples easier, which improved the listeners' ability to understand the total number of <key> samples in one set relative to another.

Another factor affecting user performance was the numerosity of <key> samples in each set. Expressed as a relative difference in per cent, each increase in relative difference led to a significant user performance improvement. This result is expected, as we hypothesised that large numerosity differences would be easier to detect than small differences. Although only one of infinitely many combinations of <key> sample numerosity was tested for each relative difference, we believe that similar results may be obtained for other combinations of <key> sample numerosity (e.g., 2 vs. 4, 2 vs. 5, etc.).

Figure 7 summarises the effect of the SOA and the relative difference between the number of <key> sample instances on user performance. With an SOA of 100 ms, user performance was poor for most tested relative differences. The 50% performance level $F_{0.5}$ is exceeded only at a relative difference of 133%, i.e., with one set containing more than twice as many instances of the <key> sample than the other. Therefore, for the given task, an SOA of 100 ms seems to be at the lower end of SOAs useful in practice. With an SOA of 400 ms, on the other hand, user performance was at or above 79%, i.e., above $F_{0.5}$, for all tested relative differences. This indicates that the task was too easy to observe an

RT, probably because the sound samples were sparse enough for test subjects to be able to count the number of <key> sample instances in both sets. In practical applications, an SOA larger than 400 ms might only marginally improve user performance, whilst increasing overall playback duration. With an SOA of 200 ms, user performance ranges from 50% to 93%, and the RT is estimated at 54% relative difference of the number of <key> sample instances. Therefore, for an SOA of 200 ms, the tested relative differences cover the perceptually most critical range.

We found no difference in the average user performance between spatially separated sound samples played via the multichannel loudspeaker system and diotic headphone playback. The users' ability to obtain an overview of the amount of <key> samples present in the sets was not affected by the spatial quality of the sound samples. As in Task I, the directions of the <key> and distractor samples were randomised and unknown to the user a priori, which cancelled the advantage of spatial separation.

The user performance did not differ substantially between earcons and speech playback. In analogy to the results of Task I, this indicates that the user can obtain an overview of the amount of <key> samples in a set from synthesised speech samples with a similar accuracy as with earcons. As in Task I, varying the vocal characteristics of the speech samples or the earcon design might further improve performance.

The statistical analysis indicates a statistically significant interaction between the SOA and the relative difference, and between the SOA and the playback setup. A visual inspection of Figure 6 confirms a strong interaction between the SOA and the relative difference, and indicates that this interaction clearly dominates other possible interactions, including an interaction between the SOA and the playback setup.

The regression tree analysis (Figure 6, right) verified that the SOA and relative difference are indeed the main determinants of the users' performance in task II. The regression tree analysis identifies a threshold near 60% relative difference (i.e., 3 versus 5 <key> samples): Differences above this threshold (67–133%, i.e., 3 <key> samples in one set versus 5, 6 or 7 <key> samples in the other) were detected substantially and significantly better than smaller differences (33–50%, i.e., 2 versus 3, and 3 versus 4 <key> samples). For the SOA, there are two main thresholds: 150 ms and 59%, and a third at 300 ms for small differences (less than 59%). The results indicate that neither sound type nor playback condition, nor demographic data or musical background had a substantial effect on user performance. The thresholds are in line with the response thresholds identified via the psychometric function fitting, which indicates that a relative difference of 60% is a perceptually critical detection threshold.

6. CONCLUSION

This paper presents results from an experiment on using auditory display in basic information retrieval tasks. The experiment tested the effect of auditory display parameters on user performance for sound sample detection (task I) and numerosity estimation (task II). The parameters tested are derived from the design stages involved in creating an auditory display: representing the information or data via audio; arranging the audio in time; displaying the audio via a playback system. The effect of these design parameters on user performance was studied by comparing speech and earcon sounds, a range of stimulus onset asynchronies (SOAs), and diotic headphone and spatial loudspeaker playback, in terms of user error rates.

Our results suggest that for both the detection and the numerosity estimation task, earcons and synthesised speech were similarly effective. As the earcons employed in this study were stripped of their specific properties, including their mapping to a referent and their integration into a hierarchical structure, we expect other types of non-speech audio to perform similarly for the given tasks. The samples used in the study could be further optimised for both sound types to improve user performance.

When arranging the sound samples in time for playback, the stimulus onset asynchrony (SOA) had a significant effect on user performance in both tasks. A longer SOA decreases the temporal overlap of sound samples and thus the number of concurrent sound samples, which in turn improves user performance. As a trade-off, a longer SOA comes with a longer overall playback duration. For the detection task (task I), we found that error rates dropped to about 10% with an SOA of 100 ms. This indicates that auditory display is effective for sample detection tasks even with a dense temporal arrangement of the samples. Task II proved to be more difficult, and an SOA of at least 200 ms was required for the user error rates to drop below 30% on average. However, even with an SOA of 200 ms, users found it challenging to discriminate 2 vs. 3 or 3 vs. 4 samples. For a difference of 3 vs. 5 samples, i.e., a relative difference of 67%, performance improved substantially. With an SOA of 400 ms average user performance was 79% or better for all tested relative differences. This might indicate that test subjects were able to count sample instances. Therefore, an SOA of 400 ms may be required in practical contexts if accurate perception of relative differences is required.

Somewhat surprisingly, the spatial arrangement of samples did not have a substantial effect on user performance. In both tasks, there was no substantial difference between diotic headphone and spatial loudspeaker playback. It seems that the lack of a priori information about the sample direction cancelled the advantage of spatial separation between samples. As a consequence, in practical applications, diotic playback, or indeed monophonic playback, may be sufficient to convey the presence or numerosity of a target sample, at least if explicit spatial information is not required. However, if spatial information was available in our test through the use of a multichannel loudspeaker system, users were able to estimate the sample direction relatively accurately, once they detected the sample. Interestingly, this ability was neither affected by the lateral angle of the sample nor by the SOA or sound type. Therefore, even with a dense temporal presentation of samples, if the test subjects were able to detect a target sample, they implicitly knew its approximate direction.

To conclude, auditory display was effective in both information retrieval tasks presented in this study. Test participants were able to determine presence and direction of a sound sample with relatively high accuracy, even with a dense temporal presentation of the samples. Conveying numerosity to users proved to be more challenging, and required a sparser temporal presentation for acceptable performance rates. The results of these experiments may be applicable in practical contexts requiring to convey the presence or numerosity of a target sample. In human–computer interfaces, menu items or lists (e.g., mobile phone contacts or calendar entries) can be presented via auditory display. Augmented reality applications may present the user with a list of objects in the user's environment via auditory display. The user may want to discover whether a certain object of interest is present in the surroundings (e.g., "is there a cash machine nearby?"), or gain an overview of the environment via the numerosity of certain types of objects or services (e.g., "are there many bars/restaurants around here?"). The study of practical applications of our findings is left for future work.

ACKNOWLEDGMENTS

We thank Tapio Lokki and Lauri Savioja for comments and discussions.

REFERENCES

BLATTNER, M. M., SUMIKAWA, D. A., AND GREENBERG, R. M. 1989. Earcons and icons: their structure and common design principles. *Hum.-Comput. Interact. 4*, 11–44.

BONEBRIGHT, T. AND NEES, M. 2009. Most earcons do not interfere with spoken passage comprehension. *Applied Cognitive Psychology 23,* 3, 431–445.

BREGMAN, A. S. 1990. *Auditory Scene Analysis: The Perceptual Organization of Sound*. The MIT Press, Cambridge, USA.

BREWSTER, S. A. 2002. Overcoming the Lack of Screen Space on Mobile Computers. *Personal Ubiquitous Comput. 6*, 188–205.

BREWSTER, S. A., RATY, V.-P., AND KORTEKANGAS, A. 1995. Representing Complex Hierarchies with Earcons. Tech. rep., ERCIM.

BREWSTER, S. A., WRIGHT, P. C., AND EDWARDS, A. D. N. 1993. An evaluation of earcons for use in auditory human-computer interfaces. In *Proc. of the ACM CHI 93 Human Factors in Computing Systems Conference*, S. Ashlund, K. Mullet, A. Henderson, E. Hollnagel, and T. White, Eds. ACM Press, New York, USA, 222–227.

BREWSTER, S. A., WRIGHT, P. C., AND EDWARDS, A. D. N. 1995. Experimentally Derived Guidelines for the Creation of Earcons. In *Proc. of BCS HCI*, M. Kirby, A. Dix, and J. Finlay, Eds. Cambridge University Press, Cambridge, UK, 155–159.

BRONKHORST, A. W. 2000. The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker condition. *Acoustica 86*, 117–128.

BROWN, L., BREWSTER, S. A., RAMLOLL, R., YU, W., AND RIEDEL, B. 2002. Browsing Modes For Exploring Sonified Line Graphs. In *Proc. of BCS HCI*, X. Faulkner, J. Finlay, and F. Détienne, Eds. Springer, London, UK, 6–9.

BRUNGART, D. S., ERICSON, M., AND SIMPSON, B. D. 2002. Design considerations for improving the effectiveness of multitalker speech displays. In *Proc. of the 8th Int. Conf. on Auditory Display (ICAD2002)*, R. Nakatsu and H. Kawahara, Eds. Advanced Telecommunications Research Institute (ATR), Kyoto, Japan.

BRUNGART, D. S. AND SIMPSON, B. D. 2002. The effects of spatial separation in distance on the informational and energetic masking of a nearby speech signal. *J. Acoust. Soc. Am. 112,* 2, 664–676.

BRUNGART, D. S., SIMPSON, B. D., ERICSON, M. A., AND SCOTT, K. R. 2001. Informational and energetic masking effects in the perception of multiple simultaneous talkers. *J. Acoust. Soc. Am. 110,* 5, 2527–2538.

CHERRY, E. C. 1953. Some Experiments on the Recognition of Speech, with One and with Two Ears. *J. Acoust. Soc. Am. 25,* 5, 975–979.

DARWIN, C. J. AND HUKIN, R. W. 2000. Effectiveness of spatial cues, prosody, and talker characteristics in selective attention. *J. Acoust. Soc. Am. 107,* 2, 970–977.

DINGLER, T., LINDSAY, J., AND WALKER, B. N. 2008. Learnabiltiy of sound cues for environmental features: Auditory icons, earcons, spearcons, and speech. In *Proc. of the 14th Int. Conf. on Auditory Display*. Paris, France.

DUDOIT, S., SHAFFER, J. P., AND BOLDRICK, J. C. 2003. Multiple hypothesis testing in microarray experiments. *Statist. Sci. 18,* 1, 71–103.

GARZONIS, S., JONES, S., JAY, T., AND O'NEILL, E. 2009. Auditory icon and earcon mobile service notifications: intuitiveness, learnability, memorability and preference. In *Proc. of the 27th Int. Conf. on Human factors in computing systems (CHI '09)*, S. Greenberg, S. E. Hudson, K. Hinckley, M. R. Morris, and D. R. Olsen Jr., Eds. ACM Press, New York, USA, 1513–1522.

GAVER, W. W. 1986. Auditory icons: using sound in computer interfaces. *Hum.-Comput. Interact. 2,* 2, 167–177.

HEALEY, C. G., BOOTH, K. S., AND ENNS, J. T. 1996. High-speed visual estimation using preattentive processing. *ACM Trans. Comput.-Hum. Interact. 3,* 2, 107–135.

HILL, J. psignifit. `http://www.bootstrap-software.org/psignifit/`.

HOLM, S. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics 6*, 65–70.

HORNOF, A. J., ZHANG, Y., AND HALVERSON, T. 2010. Knowing where and when to look in a time-critical multimodal dual task. In *Proc. of the 28th Int. Conf. on Human factors in computing systems (CHI '10)*. ACM Press, New York, USA, 2103–2112.

IHLEFELD, A. AND SHINN-CUNNINGHAM, B. 2008a. Spatial release from energetic and informational masking in a divided speech identification task. *J. Acoust. Soc. Am. 123,* 6, 4380–4392.

IHLEFELD, A. AND SHINN-CUNNINGHAM, B. 2008b. Spatial release from energetic and informational masking in a selective speech identification task. *J. Acoust. Soc. Am. 123,* 6, 4369–4379.

JULESZ, B. AND BERGEN, J. R. 1987. Textons, the fundamental elements in preattentive vision and perception of textures. In *Readings in computer vision: issues, problems, principles, and paradigms*, M. A. Fischler and O. Firschein, Eds. Morgan Kaufmann Publishers Inc., San Francisco, USA, 243–256.

KARSHMER, A. I., BRAWNER, P., AND REISWIG, G. 1994. An experimental sound-based hierarchical menu navigation system for visually handicapped use of graphical user interfaces. In *Proc. of the first annual ACM Conf. on Assistive technologies*. Assets '94. ACM Press, New York, USA, 123–128.

KIDD, G., ARBOGAST, T. L., MASON, C. R., AND GALLUN, F. J. 2005. The advantage of knowing where to listen. *J. Acoust. Soc. Am. 118,* 6, 3804–3815.

KIDD, J. G., MASON, C. R., BEST, V., AND MARRONE, N. 2010. Stimulus factors influencing spatial release from speech-on-speech masking. *J. Acoust. Soc. Am. 128,* 4, 1965–1978.

KLEIN, S. A. 2001. Measuring, estimating, and understanding the psychometric function: a commentary. *Perception And Psychophysics 63,* 8, 1421–1455.

LARSEN, E., IYER, N., LANSING, C. R., AND FENG, A. S. 2008. On the minimum audible difference in direct-to-reverberant energy ratio. *J. Acoust. Soc. Am. 124,* 1, 450–461.

McGookin, D. and Brewster, S. A. 2004a. Space, the final frontearcon: The identification of concurrently presented earcons in a synthetic spatialized auditory environment. In *Proc. of the 10th Int. Conf. on Auditory Display (ICAD2004)*, S. Barrass and P. Vickers, Eds. Sydney, Australia.

McGookin, D. K. and Brewster, S. A. 2004b. Understanding concurrent earcons: Applying auditory scene analysis principles to concurrent earcon recognition. *ACM Trans. Appl. Percept. 1,* 2, 130–155.

Michalski, R. and Grobelny, J. 2008. The role of colour preattentive processing in human–computer interaction task efficiency: A preliminary study. *Int. J. of Industrial Ergonomics 38,* 3-4, 321–332.

Nees, M. and Walker, B. 2009. Auditory Interfaces and Sonification. In *The Universal Access Handbook*, C. Stephanidis, Ed. L. Erlbaum Associates, New York, USA, 507–522.

Peres, S. C., Best, V., Brock, D., Frauenberger, C., Hermann, T., Neuhoff, J. G., Valgerdaur, L., Shinn-Cunningham, B., and Stockman, T. 2008. Auditory Interfaces. In *HCI Beyond the GUI: Design for Haptic, Speech, Olfactory, and Other Nontraditional Interfaces*, D. Penrose and M. James, Eds. Morgan Kaufmann, Waltham, USA, 147–195.

Ramloll, R., Yu, W., Riedel, B., and Brewster, S. 2001. Using non-speech sounds to improve access to 2D tabular numerical information for visually impaired users. In *15th Annual Conf. of the British HCI Group*. Springer, London, UK, 515–529.

Sagi, D. and Julesz, B. 1985. "where" and "what" in vision. *Science 228,* 4704, 1217–1219.

Sawhney, N. and Schmandt, C. 2000. Nomadic radio: speech and audio interaction for contextual messaging in nomadic environments. *ACM Trans. Comput.-Hum. Interact. 7,* 3, 353–383.

Sheskin, D. 2000. *Handbook of Parametric and Nonparametric Statistical Procedures*. Chapman&Hall/CRC, USA.

Shinn-Cunningham, B., Lehnert, H., Kramer, G., Wenzel, E., and Durlach, N. 1997. Auditory Displays. In *Binaural and Spatial Hearing in Real and Virtual Environments*, R. H. Gilkey and T. R. Anderson, Eds. Lawrence Erlbaum Associates, Mahwah, USA, 611–663.

Therneau, T. M., Atkinson, B., and Ripley, B. 2011. *rpart: Recursive Partitioning*. http://cran.r-project.org/package=rpart.

Tran, T., Letowski, T., and Abouchacra, K. 2000. Evaluation of Acoustic Beacon Characteristics for Navigation Tasks. *Ergonomics 43,* 6, 807–827.

Treisman, A. 1986. Preattentive processing in vision. In *Papers from the second workshop Vol. 13 on Human and Machine Vision II*. Vol. 3. Academic Press Professional, Inc., San Diego, USA, 313–334.

Treutwein, B. and Strasburger, H. 1999. Fitting the psychometric function. *Perception and psychophysics 61,* 1, 87–106.

Vargas, M. L. M. and Anderson, S. 2003. Combining speech and earcons to assist menu navigation. In *Proc. of the 9th Int. Conf. on Auditory Display (ICAD2003)*, E. Brazil and B. Shinn-Cunningham, Eds. Boston University Publications Production Department, Boston, USA, 38–41.

Walker, B. N., Nance, A., and Lindsay, J. 2006. Spearcons: speech-based earcons improve navigation performance in auditory menus. In *Proc. of the 12th Int. Conf. on Auditory Display (ICAD2006)*, T. Stockman, L. V. Nickerson, C. Frauenberger, A. D. N. Edwards, and D. Brock, Eds. 63–68.

Wichmann, F. A. and Hill, N. J. 2001. The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception and Psychophysics 63,* 8, 1293–1313.

Received Month YYYY; revised Month YYYY; accepted Month YYYY