PICTORIZE: TRANSFORMING IMAGE REGION LUMINOSITY TO SOUND BRIGHTNESS FOR MONITORING LOCATION

Georgios Marentakis
Department of Information Technology
Østfold University College
georgios.marentakis@hiof.no

ABSTRACT

A novel algorithm for transforming static and moving images into sound is presented. The algorithm turns image areas into a sound whose brightness depends on image area luminocity. Subjective evaluation investigated the extent to which listeners could identify if a moving target entered an image area based on the timbre of the sound produced from a sonified camera stream, while engaging with a parallel transcribing task. Results show that listeners identified target entries well above threshold, however, false detections increased for image regions of similar luminosity. Although providing visual feedback improved monitoring performance significantly, it increased dual-task cost when both displays were not accessible by peripheral vision.

1. INTRODUCTION

Sonification of images or video streams has been performed for artistic purposes such as the composition of soundscapes [1, 2], however the main body of research in the application of image sonification aims at sensory substitution; to enable people to make inferences with respect to the content of an image through sound with most notable applications in assistive technology [3–9]. However, the applicability of the related techniques should not be constrained to assistive technology or artistic expression. Given that a lot of monitoring activities involve input from cameras they may also be useful for monitoring applications. Multimodal monitoring displays may make it possible to monitor information even when visual contact to the display is temporarily lost and improve dual-task performance [10].

Sound has often been used in multimodal displays. Data sonification and auditory display [11] has been applied successfully in monitoring patient biometrics [12, 13], stock market data, industrial plants, the workplace or software processes [14]. Quite often *spatialized sound* is used to provide spatial cues in air-traffic control tasks e.g. [15, 16]. Benefits due to a multimodal display are often associated with dual-task scenarios. In such cases, multimodal displays can help reduce or even eliminate overlap in the modality used to perform each of the two tasks [17]. This is often associated with a reduced dual-task cost: the deterioration in the performance of each task due to the performance of the parallel task. However, in studies involving audiovisual displays the benefits are often significant only when the displays for each of the two tasks are not within peripheral vision. Often the benefit due to a unimodal auditory display for the one task is mitigated due to reduced task accuracy relative to a visual display [10, 15, 16, 18].

Despite a large body of research in assistive technology, image sonification techniques have not been applied in monitoring applications using camera input. This paper introduces a novel image sonification technique which allows mapping image luminosity to sound brightness. Subsequently, the usability of the technique is evaluated for a location monitoring task. Results indicate that using this technique participants can identify target entries into a monitored area well above threshold while performing a parallel visual task. However, confusions emerge as long as the luminosity of the monitored area is shared by other possible locations. Performance improves significantly in the presence of visual feedback, but dual-task cost increases if the displays to both tasks are not accessible by peripheral vision.

2. BACKGROUND

2.1 Image sonification

Image sonification techniques can be grouped in three classes. The first class maps image pixel location and intensity to sound spectrogram bin frequency, time stamp, amplitude, and sometimes spatial location. The second identifies objects in an image using computer vision and maps them to sounds. The third uses direct mappings, according to which individual pixels instead of objects are directly sonified in the location in which they appear in the visual scene. Most techniques use sound spatialization technology to spatialize sound output according to its location in the image. Most often stereophonic or binaural (HRTF) reproduction is used [19] as the majority of applications are designed to operate using headphones.

Spectrogram mappings: Spectrogram mappings typically map pixel location along the image y-axis to bin frequency, pixel location along the image x-axis to bin time, and the intensity value of each pixel to bin amplitude. Subse-

quently, each image column is synthesized using for example the inverse Short-Time-Fourier-Transform (ISTFT). This yields a wide-band signal whose frequencies are emphasized depending on the intensity of the pixels of each image column e.g. [3]. In certain implementations, the location of each pixel along the x-axis is used to spatialize the resulting sound signal for each image column in azimuth [4]. In another variation, the height of each image pixel is used to determine the pitch of software synthesized instruments [5].

Object-based mappings: In this view, objects of interest in a three-dimensional visual scene are detected using computer vision algorithms and then sonified using pre-defined mappings before being spatialized according to their location in the visual scene [6, 7]. In the ´Depth Scanning' technique proposed by Bujasz et al. [6], a surface moves away from the user along the viewing area and signals the playback of the objects it intersects; sound is spatialized using generalized HRTFs according to object position in the visual field. Distance to object was mapped to sound temporal onset delay and loudness, object size to sound pitch, object type to sound timbre, and object elongation to sound tremolo, vibrato and openness. In another approach [7], objects within the field of view were simply mapped to instrument tones that were spatialized according to object location.

Direct Mappings: Direct mappings map individual pixels in the visual scene to simple abstract sounds, which are often spatialized according to the pixel location in the visual field. A direct implementation of this approach has been done by Gonzalez-Mora et al. [8]. They sonified the contents of pixels within a low-resolution depth map, using clicks spatialized in the locations of pixels of sufficient intensity using HRTFs. The click inter-onset interval was related to the size of the objects.

Associating an elementary sound with each pixel, may however result in unnecessarily complex soundscapes. The rasterization approach of McGee et al [1, 2] is promising in addressing this issue. The method, originally developed for musical purposes, works by scanning the scaled grayscale pixel intensity values within specific image regions into an one-dimensional vector of wavetable samples, reproduced using scanned synthesis [20]. Direct playback of such raw data results in noisy sounds without much variation between separate regions in most images. For this reason, the spectrum of the wavetable data is calculated and processed. Subsequently, the sound signal is obtained using the inverse STFT. Sound output from different regions can be mixed and rendered in stereo or multichannel, whereby the location of the image region determines the azimuth of its sound [2].

The See ColOr project, [9] developed prototypes that sonified HSL color values of individual pixels at the location in which they occurred in the visual field using an HRTF rendering of Ambisonics. Hue was mapped to instrument timbre, saturation to pitch, and luminosity to a bass sound when it was low and to singing voice when it was bright. Hue values were quantized to the 8 basic colours and intermediate values resulted in a mix of the associated timbres. Luminosity was sonified on top of hue, according to an algorithm which emphasized high and low values and deemphasized middle luminosity values. Distance to objects was mapped to into four duration levels.

2.2 Spatial monitoring using sound

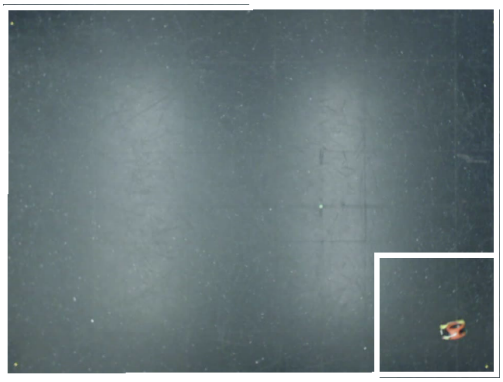
Sound has often been used for monitoring spatial parameters. These are often mapped in a non-spatial auditory dimensions such as fundamental frequency, tempo, loudness, dynamics, or timbre. The distance to objects of interest has also been encoded into changes in the fundamental frequency, modulation, loudness, timbre, or rate of onset (beep rate) of a sound [21–23]. Although spatial Mappings in which the 2D or 3D location of the object is presented using a sound in space are also possible, the relatively low spatial resolution of hearing as well as the necessity for dedicated spatialization hardware and individualized software makes non-spatial auditory Mappings of spatial parameters worth investigating.

Timbre is a particularly appealing parameter because changes in timbre are less annoying than changes in frequency [21], while still resulting in fast processing and high task accuracy [23]. Timbre has been used to encode 1D spatial features (such as distance), but not 2D location. For example, sounds have been given a distinct timbre or were filtered in a different way depending on whether the came from the front or the back of the user to help resolve front-back ambiguity [24]. Timbre is multidimensional. Mappings may be designed using its perceptual dimensions, for example the spectral centroid (which correlates to perceived brightness), the attack time, or the spectral flux [25].

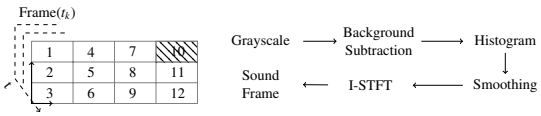
2.3 Summary and Research Questions

Considering image sonification techniques for application to the location monitoring task considered here the following observations can be made based on Section 2.1: 1. a major difficulty with the spectrogram technique is the high cognitive load required to process the generated soundscape into a viable visual scene, 2. object-based techniques rely on sound spatialization so that they can be applied to location monitoring, and may face object tracking and mapping difficulties. Furthermore, due to the abstract mappings involved both techniques require extensive training before they can be used effectively.

The direct mapping techniques target an inherently perceptual mapping and may thus be more effective. With or without spatialization, the combination of simultaneous elementary sounds results in a changing timbre as the contents of sonified image region change. This is potentially interesting for creating low-annoyance auditory Mappings of spatial quantities. However, the output of direct mapping techniques is often unpredictable, hard to associate with image areas, and the resulting soundscapes are difficult to parse. The rasterization approach is interesting as by focusing and processing image areas it addresses the problem to some extent. However, sound output is also unpredictable if further processing is not performed. Therefore, it was not possible to find a straightforward solution



(a)



(b)

Figure 1: (a) A snapshot of the video that was projected on the canvas. Participants were tasked to detect when the car entered the marked area. When Modality was auditory no video was projected but the white frame remained visible. (b) An overview of the sound generation algorithm. The process is illustrated for location number 10 but was repeated for each grid subdivision.

for mapping 2D spatial location into timbre for location monitoring using visual input. It is therefore reasonable to ask: (1) how can timbre be used to encode 2D object location using visual input and, (2) how well can it be used to differentiate between locations.

3. PICTORIZE: THE MAIN CONCEPT

To create the algorithm used here, the rasterization approach is extended [1,2]. A difficulty with the rasterization approach is that scaled intensity values from each region are scanned into a wavetable buffer and played back using scanned synthesis which results in noisy sounds that change little across grid subdivisions. Using filters to yield an aesthetically pleasing sound is a possibility, however, the relationship to the original image content is gradually dissolved as a result, which is problematic for monitoring applications.

This can be overcome if instead of using pixel values as samples fed into scanned synthesis, one calculates a histogram of the pixel grayscale intensity values within each image area. If the histogram is then mapped directly to sound spectrum the resulting sound can be perceptually related to the image content. Specifically, considering a grayscale image a mapping between image region luminosity and sound brightness, typically defined as the instantaneous value of the spectral centroid, is achieved. This approach scales well as image areas of variable dimensions can be defined to suit different applications.

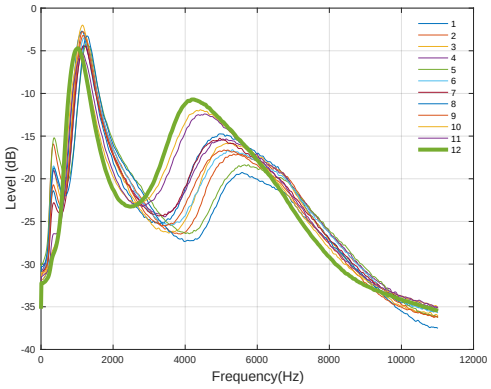


Figure 2: Periodograms of the synthesized sound for each grid subdivision

Application in location monitoring: The application of the algorithm in the location monitoring experiment that follows is based on using a 4×3 grid which exactly covers the image frame. The algorithm operated on the pixels within each grid subdivision. These will be called L1-L12. Numbering starts at the top let corner and proceeds in each column. The monitored area was bottom right (L12). Right above it are L11 and L10 and directly to the left L9.

First, background subtraction was applied, then histograms for each grid subdivision were calculated, normalized, smoothed, and mapped to the sound spectrum for each frame, and finally re-synthesized using the inverse STFT before being normalized by the number of subdivisions and added to the output buffer. Background subtraction resulted in that no sound was produced when the sound did not move and in that no sound was produced from grid subdivisions that did not contain the target as the resulting intensity of all pixels was zero and the histogram and the associated spectrum had a single peak at the DC component. A STFT window of 1024 samples was used. Histograms were linearly interpolated to increase their original 256 bins resolution. The resulting spectra were band-limited between 400 and 12000 Hz and smoothed using a first order low-pass filter to broaden spurious peaks and avoid tonal components.

The application of the algorithm to each video frame results in that the colour patterns of the target car give rise to three peaks in the intensity histogram and the periodogram of the resulting sound (Figure 2). These are relatively broad and give the sound a band-limited noise quality without tonal components. As hypothesized, the peak centre frequency and amplitude changes depend on the grid subdivision the car is currently in (Table 1). Listeners need therefore to be able to correctly identify the target car location based on its timbral signature to perform the location monitoring task.

It follows that the timbre of the sound generated when the car is in L12 should be easy to distinguish from this of L1-L9 because: 1. the difference in both second and third peak centre frequency between the sound generated when the car was in L12 and L1-L9 was consistently above 10% and 2. the first spectral peak was attenuated in L12. However, it should be more difficult to differentiate between

the sound of the car in L12 and this in L10 and L11 because: 1. the difference in the third peak central frequency was ~200Hz (4%, L12-L10) and ~310Hz (7%, L12-L11), 2. the difference in the second peak centre frequency was ~50Hz (4%, L12-L10) and ~100Hz (12%, L12-L10), 3. all locations exhibited a first peak with a similar centre frequency. Changes in the resonant frequency of a second-order filter can be discriminated if they exceed 8% the centre frequency, or even less for Q>1, for centre frequencies between 300 and 2kHz [26].

4. EVALUATION

An experiment was designed in order to answer the research questions. Participants monitored the location of a single object (a remote-controlled toy car) which was moving on a floor surface. The car had been recorded by a video camera placed above the surface while its movement was controlled remotely (Figure 3). Participants reported as soon as the car entered the marked area on the video screen (Figure 1).

Independent Variables: The independent variables were: (monitoring) display Modality (auditory or audiovisual), and (monitoring) display Location. Factor display Modality manipulated whether location monitoring was performed by watching a video of the car moving on a projection screen and hearing a sound that encoded the current car location (audiovisual), or just by hearing the sound (auditory). Factor display Location manipulated whether the location monitoring display was within the participant's visual field (frontal, Figure 3) or in a dorsal Location at the participants' back and outside the visual field.

Hypotheses: The following hypotheses about location monitoring performance with the auditory cue were evaluated: 1. the difference in the frequency of spectral peaks will allow listeners to identify target versus the rest of locations, however, performance will be faster and more accurate with the audiovisual compared to auditory Modality, 2. reversing display Location will incur a location monitoring cost when Modality is audiovisual but not when it is auditory, 3. the likelihood of confusing with L12 will increase for L11 and L12 in comparison to other Locations when Modality is auditory.

Dependent variables: Hypotheses were evaluated by comparing location monitoring speed and accuracy and when relevant transcribing speed and accuracy in the conditions tested in the experiments. To estimate location monitoring accuracy, detections registered while the target (or a part of it) was in the monitored area were classified as hits (or true positives, tp). Detections registered while the target was outside the area were classified as false alarms (or false positives, fp). A missed entry in the target area was classified as a miss (or false negative, fn). Furthermore, hit rate (hr) was calculated by dividing hit count with the total number of target entries in the target area. False alarm rate (fa) was calculated by dividing the number of detections in each Location with the total target entries for each Location for the particular trial. This was done for each possible location yielding the spatial distribution of false alarm rate.

Location monitoring speed was estimated as the fraction of the instance in which a detection was registered relative to the respective duration of each specific target entry in the monitored area in order to counterbalance variability in target entry duration in the monitored area. To this end, the instance in which a detection was registered was normalized within the [0,1] interval (0.5 means detection occurred at half the entry duration). As the number of detections per participant and condition varied, the average of valid detection intervals in each condition for each participant was averaged to yield a balanced data-set.

Transcribing speed and accuracy were quantified as the total number of correctly typed words and the average number of typed characters per second respectively. Both transcribing speed and accuracy were normalized in [0,1] for each participant versus a control typing condition to account for individual differences in transcribing.

Transcribing speed, accuracy, monitoring hit and false alarm rate, monitoring speed, detection frequency were analyzed statistically using a two-way display Location × Modality ANOVA. When analyzing spatial false alarm rate a display Location × Modality × target Location ANOVA was used. ANOVAs were followed by pairwise t-tests at $p<0.05$ with Holm p-value adjustment for multiple comparisons. F-statistic using Type III error terms is reported for ANOVAs. ANOVAs were performed using the ez package of the R language.

	F_1	A_1	F_2	A_2	F_3	A_3	ΔF_2	ΔF_3
L1	0.33	-21	1.14	-3	4.90	-15	11%	15%
L2	0.34	-15	1.16	-3	5.35	-17	13%	26%
L3	0.33	-18	1.15	-2	5.25	-16	12%	24%
L4	0.35	-19	1.18	-4	5.10	-15	15%	20%
L5	0.36	-15	1.26	-4	5.66	-18	23%	33%
L6	0.34	-18	1.20	-3	5.20	-17	17%	22%
L7	0.35	-23	1.17	-4	4.96	-15	14%	17%
L8	0.35	-18	1.28	-3	5.59	-19	25%	32%
L9	0.34	-20	1.17	-3	5.10	-17	14%	20%
L10	-	-	1.07	-6	4.42	-11	4%	4%
L11	-	-	1.15	-5	4.54	-12	12%	7%
L12	-	-	1.02	-4	4.23	-11	-	-

Table 1: The center frequencies (kHz) and amplitude (dB) of the three frequency peaks that characterized the generated sound spectrum depending on the grid subdivision in which the target was located averaged over all five videos used in the experiment. ΔF_2 refers to the difference in frequency to the monitored area.

4.1 Setup

Participants sat in front of a desk which supported a screen and a keyboard. The text to be transcribed was placed on a book rest on the left-hand side of participants (Figure 3), while typed text appeared on the screen.

The video of the target moving was projected onto an acoustically transparent projection screen (white PVC, 390 g/m^2, 7 percent perforation area), which was located at a distance of 2.5m in front of the participant's chair. The video used in each trial was picked randomly out of five videos (about 4:00 minutes duration)in which the car followed different movement paths but entered and stayed in

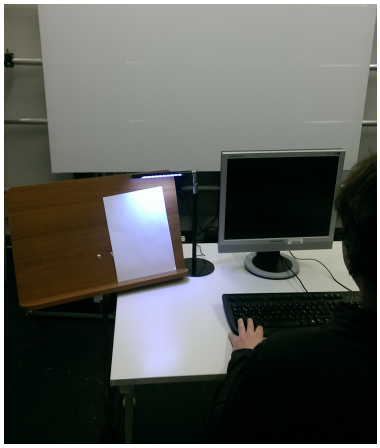


Figure 3: A photo of the experiment setup including a user, the book rest, screen, and keyboard for performing the transcribing task and the projection canvas behind which the loudspeakers were placed. When display Location was reversed the location of the desk and the book rest was reversed and participants looked away from the projection screen.

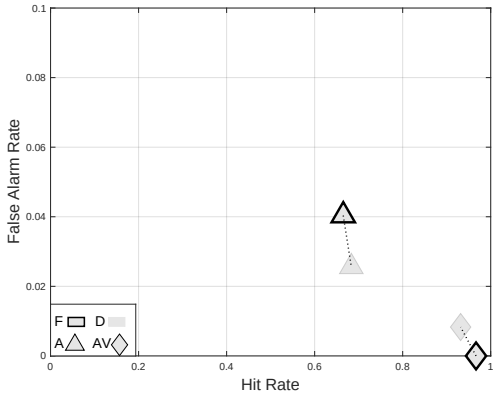


Figure 4: Hit vs. false alarm rate in the conditions in the experiment. [Modality: A = Auditory, AV = Audiovisual, Location: F = Frontal, D = Dorsal]

the monitored area a similar amount of time.

When appropriate sound was played by a single loudspeaker behind the middle of the projection area. When display Location was dorsal, the position of the chair, desk, and book rest were rotated 180° so that the distance to the projection screen and loudspeaker did not change but these were now at the participants' back. Texts used when transcribing contained random word sequences of the same difficulty out of a pool of 72154 lowercase words without accents or special characters. A different text was used in each dual-task condition, which remained the same for all participants in each experiment session.

There were two computers. The first controlled the experiment, video playback and processing, and sound synthesis. The second ran a Pure Data patch, which received Open Sound Control (OSC) messages and automatically opened and saved editor windows for typing and logged key presses when a new condition started. Keyboard input was blocked as soon as a condition was finished. All operations were implemented in real time using Open Frameworks (`https://openframeworks.cc`) and a self-implemented C STFT library built on the FFTW library. Both PCs ran the Debian GNU/Linux distribution. Jack (`http://jackaudio.org`) was used for audio playback running at 44.1 kHz / 24 bit resolution. A Genelec 8020CPM loudspeaker, driven by an RME MADI FX audio interface and Behringer ADA8000 DA converter, was placed 0.5m behind the projection screen to play back sound. Equivalent Continuous Sound Level (Leq) at the listening position was 50 dBA. The experiment was performed in an acoustically treated room, 6m (l) × 4m (w) × 3m (h).

4.2 Procedure

Initially, participants provided written consent and were briefed on the tasks. Subsequently, a single-task transcribing condition was performed to measure typical transcribing speed and accuracy followed by a training singe-task audiovisual location monitoring condition so that participants got used to the relationship between the video and the sound stimulus. Dual-task conditions were performed next in a counter-balanced order as part of a bigger study. Participants received a monetary compensation for their time. None reported hearing problems.

Participants: There were two groups of ten participants (5 female, μ=24 years, σ=3.8 years), which were tested in the frontal and dorsal display Locations, respectively. Listeners had no training in critical listening.

Task: On the single-task training condition, participants pressed a key when the moving target entered the monitored area. The projection and the monitored area boundaries were marked visually using a white frame (see Figure 1). On dual-task conditions, the location-monitoring task did not change but participants performed the task simultaneously with transcribing text.

Instructions: In the single-task transcribing condition, participants were instructed to transcribe as quickly and accurately as possible. In the dual-task conditions, participants were instructed to transcribe as quickly and accurately as possible while they monitor car movement and report as soon as the car entered the monitored area by pressing the ESC key once.

4.3 Monitoring Performance

Location monitoring hit and false alarm rate is shown in Figure 4. Hit rate was highest when Modality was audiovisual followed by auditory; the main effect of Modality was significant $F(1,18)=35.74$, <0.001. The main effect of Location and the Location × Modality interaction were not significant. In pairwise comparisons, hit rate was significantly higher for the audiovisual Modality for both Locations and the effect of Location was not significant for both Modalities. Fewer false alarms were observed when Modality was audiovisual, however, their number increased when display Location became dorsal. The effect of Modality was significant, $F(1,18)=13.68$, $p=0.001$.

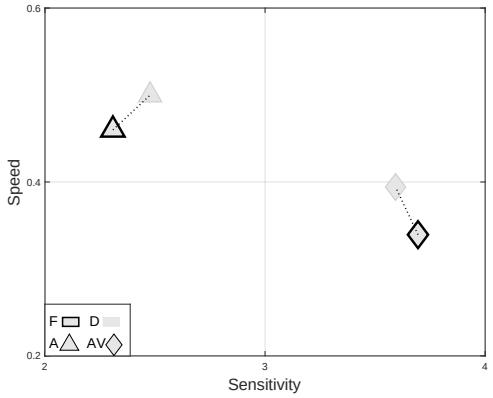


Figure 5: Monitoring sensitivity vs. speed in the conditions in the experiment. [Modality: A = Auditory, AV = Audiovisual, Location: F = Frontal, D = Dorsal]

The effect of Location was not significant. The Modality × Location interaction was marginally significant, F(1,18)= 4.42, p= 0.049. In pairwise comparisons, the difference in false alarm rate due to Modalities was significant when Location was frontal, p<0.001, but not when it was dorsal and false alarm rate deteriorated significantly when Location reversed when Modality was audiovisual, p=0.022, and not when it was auditory.

Sensitivity is shown on Figure 5. Sensitivity was higher when Modality was audiovisual. The effect of Modality was significant, F(1, 18)= 7.47, p<0.001. The effect of Location and the Location × Modality interaction were not significant for sensitivity. Monitoring speed is also shown on Figure 5. Detections were fastest when monitoring display Modality was audiovisual and Location frontal. The effect of Modality, F(1,18)=59.05, p<0.001, and Location, F(1,18)= 14.97 p<0.001 were significant. Pairwise comparisons showed that detections were fastest when Modality was audiovisual for both display Locations (p<0.001); Monitoring speed deteriorated significantly when Location was reversed when Modality was audiovisual but not when it was auditory.

The Spatial false alarm rate is shown on Figure 6. Detection rate was 75% when target was in L12, while false

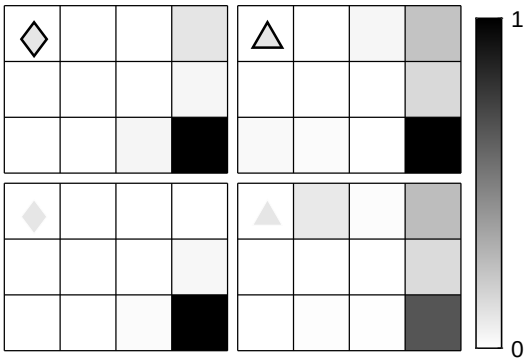


Figure 6: Spatial distribution of detections in the conditions tested in the experiment

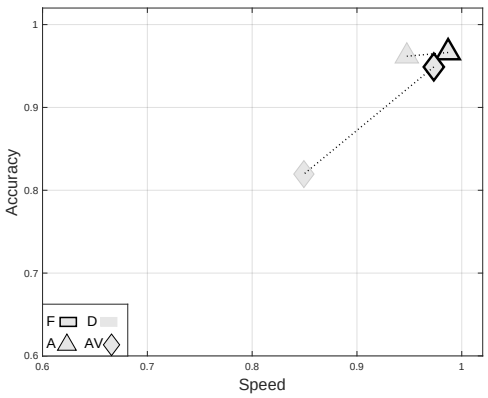


Figure 7: Typing speed vs accuracy in the conditions tested in the experiment. Data have been normalized against their value in the control typing condition for each participant. [Modality: A = Auditory, AV = Audiovisual, Location: F = Frontal, D = Dorsal]

alarm rate for L11 and L10 was 11% and 5% for L9, while false alarm rate was below 1% for the rest of the Locations. In the analysis, only Locations 9-11 are taken into account. The effect of Modality, F(1,18)=34.03, p<0.001, and target Location F(2,36)=7.26, p=0.002, as false alarm rate was higher for the auditory compared to the audiovisual Modality, and false alarm rate in L9 was significantly lower than L10, p=0.034 and L11, p=0.009. False alarm rate between L10 and L11 was not significantly different. The display Location × Modality interaction was significant, F(1,18)= 6.9, p=0.017. While the effect of Modality remained significant for both display Locations, false alarm rate in L9-L11 increased significantly when Location was reversed and Modality was auditory but not when it was audiovisual. The Modality × target Location interaction was also significant F(2,36)=9.62, p<0.001. The differences in false alarm rate among L9-L11 were significant when Modality was auditory, but not when it was audiovisual. False alarm rate varied significantly with Modality for Locations 10 and 11, p<0.001, but not for L9.

4.4 Transcribing Performance

Transcribing speed and accuracy were significantly lower than nominal (1) in both conditions (p<0.05). The deterioration was highest when monitoring display Location was dorsal and for the visual and audiovisual monitoring display Modalities (Figure 7).

The main effect of monitoring display Modality was significant for transcribing accuracy, F(1,18)=5.74, p=0.02 and marginally significant for speed, F(1,18)=3.64, p<0.08. The main effect of Location was not significant. The Modality × Location interaction was significant for both speed, F(1,18)=5.02, p = 0.03, and accuracy F(1,18)=5.74, p=0.02. The difference between modalities in terms of speed and accuracy was not significant when Location was frontal, however, both speed, p=0.018, and accuracy, p=0.003, improved when Modality was auditory when Location was dorsal.

5. DISCUSSION

Hypothesis 1 was confirmed as although participants were relatively successful with identifying target location, location monitoring was faster and more accurate when visual feedback was available. Arguably, this was an easy task considering the high spatial resolution of the visual system. The main limitation of the auditory display was in terms of hit rate, which attained a grand average of 67% and varied between 30% to 100% among participants. This reflects individual differences in the ability of participants to use the timbral cues which may be possible to smooth out with increased exposure to the sounds. Another factor that may have contributed to the low performance of some of the participants is the dynamic nature of the sound stimulus while the target was within the timbre subdivisions. Sound was updated with each new video frame which resulted in that the timbre of the sound changed while it was in each of the grid subdivisions. This was done to help participants associate visual and auditory cues. This could have increased the sonic complexity some participants were prepared to deal with. A variation in which sound is synthesized based on an average histogram could be used to result in a static timbre for each of the possible locations, as in the periodograms in Figure 2. This may improve performance. False alarm rate was 4.2% on average, which resulted in an average sensitivity of 2.28, well above threshold but significantly less than the 3.5 units attained when Modality was audiovisual.

Hypothesis 2 was verified as reversing display Location reduced transcribing speed and accuracy as well as monitoring speed and false alarm rate when Modality was audiovisual but not when it was auditory. This is consistent with other studies reporting improvements in dual-task performance when the displays to each task are not within peripheral vision [15, 27] and predictions of resource allocation theories [10]. Apart from an increased false alarm rate and a reduction in monitoring speed, monitoring with the audiovisual Modality was not affected by reversing Location. This could be attributed to the relatively easy visual monitoring task. As we observed in the experiments, participants did turn to visually verify if the car was in the target Location, when monitoring display Location was reversed. This explains the observed deterioration in transcribing and monitoring performance. The toy-car was, however, moving at a low speed and was thus easy to avoid misses by penalizing transcribing performance. Quite likely participants used auditory feedback to some extent in order to time the points they turned to the visual display. How much did they benefit from auditory feedback cannot be estimated without analyzing data from a condition that does not include auditory feedback.

Hypothesis 3 was also verified as the likelihood of confusing the monitored to another area increased when the target was in Locations 10 ans 11. As discussed in Sections 3, the resulting spectrum from these two locations was most similar to the sound spectrum synthesized using the pixels from the target location. The confusion rate for these locations was therefore higher 10% and 12% respectively. The overall false alarm rate was, however, lower as the target moved in other locations during the trials, for which false alarm rate was well below.

Let us return to our research questions (1) how can timbre be used to encode 2D object location using visual input and (2) how well can it be used to differentiate between locations. The results indicate that the luminosity to brightness timbral Mapping of spatial Location could be a promising way to encode 2D object location to sound. The algorithm we proposed resulted in a distinct sound that varied depending on the luminosity of the current target location. Using the sound timbre users were able to recognize when the target entered a target versus other locations well above threshold. However, the observed limitations in hit rate and false alarm rate for locations of similar luminosity may restrict the applicability of the technique to unimodal auditory displays when increased sensitivity is required. Furthermore, the evaluation presented here involved a single target, so it is not clear how the algorithm will perform had more than one object moved in the display. Future work could concentrate on improving user performance perhaps by using a constant sound timbre for each possible location, enrich Mapping by involving other parameters, such as colour, to increase the variation of sound timbre in the different locations, but also use spatialization techniques to investigate potential improvements due to compound timbral and spatial Mappings.

6. CONCLUSION

A novel algorithm for sonifying the location of an object in an image or a video stream using a luminosity to sound brightness mapping was presented. Using sound output, users were able to monitor the entry of a moving object in an area of interest while performing a transcribing task with an accuracy that was well above threshold. However, monitoring accuracy and speed remained lower in comparison to the performance obtained when visual feedback was provided in addition to auditory. Despite these limitations, monitoring based on auditory feedback resulted in lower cost due to the performance of a parallel visual task when the displays to both tasks were not within peripheral vision and thus provides increased flexibility with respect to the positioning of the user relative to the monitoring display.

Acknowledgments

The author would like to acknowledge the assistance of Marian Weger in setting up and running the experiment.

7. REFERENCES

[1] R. McGee, "Vocis: a multi-touch image sonification interface," in *Proceedings of the New Interfaces for Musical Expression Conference (NIME 2013)*. ACM, 2013, pp. 460–463.

[2] R. M. McGee, J. Dickinson, and G. Legrady, "Voice of sisyphus: An image sonification multimedia installation," in *The 18th Internatiaonal Conference on Audi-*

tory Display (ICAD 2012). Georgia Institute of Technology, 2012.

[3] P. Meijer, "An experimental system for auditory image representations," *Biomedical Engineering, IEEE Transactions on*, vol. 39, no. 2, pp. 112–121, Feb 1992.

[4] C. Capelle, C. Trullemans, P. Arno, and C. Veraart, "A real-time experimental prototype for enhancement of vision rehabilitation using auditory substitution," *Biomedical Engineering, IEEE Transactions on*, vol. 45, no. 10, pp. 1279–1293, Oct 1998.

[5] G. Balakrishnan, G. and. Sainarayanan, R. Nagarajan, and S. Yaacob, "Wearable real-time stereo vision for the visually impaired," 2006.

[6] M. Bujacz, P. Skulimowski, and P. Strumillo, "Sonification of 3d scenes using personalized spatial audio to aid visually impaired persons," in *International Conference on Auditory Display*. International Community for Auditory Display, 2011.

[7] Y. Kawai and F. Tomita, "A support system for visually impaired persons to understand three-dimensional visual information using acoustic interface," in *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, vol. 3, 2002, pp. 974–977 vol.3.

[8] A. F. Rodríguez-Hernández, C. Merino, O. Casanova, C. Modrono, M. A. Torres, R. Montserrat, G. Navarrete, E. Burunat, and J. L. González-Mora, "Sensory substitution for visually disabled people: Computer solutions," *WSEAS Transact Biol Biomed*, pp. 1–10, 2010.

[9] G. Bologna, B. Deville, and T. Pun, "Sonification of color and depth in a mobility aid for blind people," in *Proceedings of the 16th International Conference on Auditory Display (ICAD 2010). Washington, DC, USA*, 2010, pp. 9–13.

[10] C. D. Wickens, "Multiple resources and performance prediction," *Theoretical issues in ergonomics science*, vol. 3, no. 2, pp. 159–177, 2002.

[11] T. Hermann, A. Hunt, and J. G. Neuhoff, *The sonification handbook*. Logos Verlag Berlin, GE, 2011.

[12] Á. Csapó and G. Wersényi, "Overview of auditory representations in human-machine interfaces," *ACM Computing Surveys (CSUR)*, vol. 46, no. 2, p. 19, 2013.

[13] T. Hildebrandt, T. Hermann, and S. Rinderle-Ma, "Continuous sonification enhances adequacy of interactions in peripheral process monitoring," *International Journal of Human-Computer Studies*, 2016.

[14] P. Vickers, "Sonification for process monitoring," in *The Sonification Handbook*. Logos Verlag, 2011, pp. 455–492.

[15] A. J. Hornof, Y. Zhang, and T. Halverson, "Knowing where and when to look in a time-critical multimodal dual task," in *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 2010, pp. 2103–2112.

[16] D. R. Begault, "Head-up auditory displays for traffic collision avoidance system advisories: A preliminary investigation," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 35, no. 4, pp. 707–717, 1993.

[17] C. D. Wickens, "Multiple resources and mental workload," *Human factors*, vol. 50, no. 3, pp. 449–455, 2008.

[18] P. M. Sanderson, D. Liu, and S. A. Jenkins, "Auditory displays in anesthesiology," *Current Opinion in Anesthesiology*, vol. 22, no. 6, pp. 788–795, 2009.

[19] E. Wenzel, M. Arruda, D. Kistler, and F. Wightman, "Localization using nonindividualized head-related transfer functions," *The Journal of the Acoustical Society of America*, vol. 94, no. 1, pp. 111–123, 1993.

[20] R. Boulanger, P. Smaragdis, and J. Fitch, "Scanned synthesis: An introduction and demonstration of a new synthesis and signal processing technique." ICMA, 2000.

[21] R. H. Lorenz, A. Berndt, and R. Groh, "Designing auditory pointers," in *Proceedings of the 8th Audio Mostly Conference*. ACM, 2013, p. 6.

[22] I. Hussain, L. Chen, H. T. Mirza, K. Xing, and G. Chen, "A comparative study of sonification methods to represent distance and forward-direction in pedestrian navigation," *International Journal of Human-Computer Interaction*, vol. 30, no. 9, pp. 740–751, 2014.

[23] P. Bazilinskyy, W. van Haarlem, H. Quraishi, C. Berssenbrugge, J. Binda, and J. de Winter, "Sonifying the location of an object: A comparison of three methods," 2016.

[24] S. Holland, D. R. Morse, and H. Gedenryd, "Audiogps: Spatial audio navigation with a minimal attention interface," *Personal and Ubiquitous computing*, vol. 6, no. 4, pp. 253–259, 2002.

[25] S. McAdams, "Musical timbre perception," in *The Psychology of Music (Third Edition)*, third edition ed., D. Deutsch, Ed. Academic Press, 2013, pp. 35 – 67.

[26] J.-P. Gagné and P. Zurek, "Resonance-frequency discrimination," *The Journal of the Acoustical Society of America*, vol. 83, no. 6, pp. 2293–2299, 1988.

[27] D. Brock, B. McClimens, and M. McCurry, "Virtual auditory cueing revisited," in *proceedings of the 16th international conference on auditory display, Washington, DC*, 2010.