Comparison of Auditory and Visual Short-Term Memory Capabilities using a Serious Game Application

György Wersényi*, and Ádám Csapó†

***Abstract*—A comprehensive serious game application has been designed and implemented to examine the capacity and effective- ness of short-term auditory and visual memory, otherwise known as working memory in human subjects. Participants engaged in an adaptation of the well-known paired association game that entails turning over cards and recalling their placement within a 2D matrix structure of various resolutions. Each trial introduced either visual icons (vision-only condition) or auditory objects (audio-only condition). User performance was evaluated through a detailed statistical analysis focusing only on the highest 6x8 resolution condition in the application. Findings suggest that visual memory did not conclusively outperform auditory memory in the context of this game. However, within the scope of auditory stimuli, familiar iconic sounds, such as excerpts of speech and commonplace sounds, were recalled more effectively than unfamiliar, synthetic sounds like parametric waveforms. Furthermore, performance appeared to be influenced by demographic factors, with male and younger subjects yielding superior results.**

***Index Terms*—Auditory memory; visual memory, virtual simulation; gamification; memory game**

I. INTRODUCTION

In this paper, we present a serious game application that we have developed to assess users' capabilities towards retaining different kinds of auditory and visual stimuli in short-term memory. Based on results from the application, we draw conclusions that are relevant to the design of audiovisual user interfaces in a wider technological context.

A. The role of working memory in human cognition and perception

In psychology, working memory is defined as the part of short-term memory that is concerned with immediate conscious perceptual and linguistic processing. It is the cognitive system involved in the temporary storage and processing of a limited amount of information as a given task is being carried out [1]. The most important modalities in this context are the visual and the auditory modalities, with each being characterized by different capabilities [2]–[5].

Many researchers view working memory and short-term memory as significantly overlapping concepts. The difference between the two in the context of this study is that working memory is regarded as an active mechanism (process) for the manipulation and application of memory objects over a short period of time, while short-term memory simply refers to the temporary storage (capacity) of the brain, making information readily available for a short period of time [6]. Apart from this, the two terms may be used interchangeably.

Most of the information kept in short-term memory will be stored for approximately 20 to 30 seconds, or even less. Some information, however, can last in short-term memory for up to minutes, but most information spontaneously decays quite quickly [7], [8].

In contrast with working memory and short-term memory, long-term memory refers to a vast store of knowledge pertaining to prior events. It differs from short-term memory both in terms of duration and capacity. The question of decay / forgetting is still an actively researched area, although some works have argued that there may be a chunk capacity limit even in the case of long-term memory [2].

B. Investigating visual and auditory performance in a modern technological context

A well-known model of working memory is the Baddeley-Hitch model, which differentiates between two components of working memory: a place where visual and spatial information is stored, and another for recording auditory information. According to the model, a central executive part controls and mediates these components [9], [10].

Although a wide range of experiments have already been conducted to test human capabilities and to compare performance in different modalities, recent technical developments have paved the way towards new applications in the digital world with an increasingly extensive use of audiovisual information.

On the one hand, the concepts of Digital Reality and Internet of Digital Reality encompass various kinds of developments that benefit from the broader use of such kinds of enhanced user experience and immersive 3D scenarios in a functionally driven, networked artificial intelligence context [11], [12].

* Széchenyi István University, Győr, Hungary (e-mail: wersenyi@sze.hu)
† Corvinus Institute for Advanced Studies & Institute of Data Analytics and Information Systems Corvinus University of Budapest, Hungary (e-mail: adambalazs.csapo@uni-corvinus.hu)

DOI: 10.36244/ICJ.2024.2.7

Healthcare applications, combat and military simulators, gaming, and various other AR and VR application domains rely on visual and/or auditory information during feedback. Virtual audio displays (VADs) in a wider sense can be applied in simulators, virtual and embedded environments, making extensive use of auditory stimuli.

On the other hand, assistive technologies targeting elderly or disabled individuals using wearables, e.g. for reading, navigation (Electronic Travel Aids) or other use cases, also strongly rely on well-designed auditory stimuli. More generally, cognitive rehabilitation procedures can involve testing and training of the memory capabilities of patients. In all of these applications, the auditory modality and auditory memory play a significant role. Therefore, understanding what type of simulated sound sources are best suited for a given environment – in terms of, e.g. spectral content, duration, amplitude profile – can be crucial.

Serious gaming offers possibilities for testing, training, and scientific data collection using entertaining gaming environments where user involvement and motivation are maintained [1], [13]–[18]. In the era of mobile devices and virtual reality headsets, serious games can reach a large group of subjects both in online and offline scenarios. Using gamification, experiments can be designed in order to collect scientifically relevant data in an entertaining and motivating process. It is an especially useful method if the gamified scenario resembles the original environments and use cases. Gamified applications are useful for both children and elderly users, where maintaining focus and motivation is a key factor. 3D immersive virtual reality scenarios also open new areas for serious gaming where (immersive) audiovisual experience may be extended by haptic/tactile feedback. State-of-the-art gaming applications can be close to reality and virtual training scenarios, and can be suited for testing human cognition and perception performance.

There are different definitions of serious gaming. While some definitions focus almost exclusively on the "learning element" of these games (e.g., [19]), a broader and also very common conception is reflected in multiple definitions, such as "[A serious game is] any form of interactive computer-based game software for one or multiple players to be used on any platform and that has been developed with the intention to be more than entertainment" [20], or "[Serious games are] video games aimed toward problem-solving rather than entertainment" [21]. It is also reported in the literature as being widely accepted that serious games are built using novel technologies, includes at least some game-like features (such as competition, desire to win, or strategy, among others), and are created with a "serious" intention to achieve a concrete objective [22].

On the one hand, all of these areas highlight the importance of the optimization of the presentation of audiovisual information. On the other hand, a gamified environment built for the purposes of evaluating human audio-visual capabilities can in itself be regarded as a serious game. One key aspect in this context is users' ability to recall and remember various properties of audiovisual items, i.e., presence or absence, meaning, semantic connections, temporal or spectral variations, spatial locations, etc. Although the currently presented experiment did not include an evaluation of any possible learning effects, the game design allows for further experiments targeting the learning effect as well, with dedicated experimental design (repeated controlled measurements, inclusion of lower resolutions, etc.). The following subsections provide a brief overview of human visual and auditory working memory.

C. Visual memory

Most research on visual memory in the past has focused on humans' ability to remember visual stimuli over either shorter or longer periods of time, i.e. time was taken to be the key parameter [23]. Thus, it was shown that humans can remember objects seen even for brief exposures or after very long time. The visual working memory is considered as a system that retains and manipulates information over the short term, whilst visual long-term memory is defined as a passive storage of information for longer time periods [24], [25].

The most important property of working memory is its limited capacity, however, it is a core cognitive process supporting human behaviour that relies on temporarily stored visual information [25]. To this end, easily interpretable iconic representations may be most useful. In fact, although object identification and recognition is generally associated with long-term memory, it nevertheless plays a significant role in short-term memory processes as well.

Memory limitation experiments usually apply the so-called change-detection task, where objects are displayed in an array, and after a short break, another array is displayed with changes that have to be identified by the test subjects. Alternatively, single-item presentation instead of an array of items can be used, but in this case the relationship between visual objects in terms of their joint effects on memory cannot be assessed.

Luck and Vogel have suggested that working memory can store only a limited, discrete number of objects [26]. In their experiment, subjects were asked to recall an array of items characterized by a single or multiple features (i.e., color, orientation). After a short delay, another array was shown that was identical or different, and the task was to identify what changes had been applied, if any. Based on the study, Luck and Vogel found that subjects could store only 3-4 objects in working memory, and they could detect changes in both single and across multiple features. Other studies reported that capacity is reduced as feature load increases [27]. Thus, visual memory experiments have to consider both feature load (number of features) and object load (number of objects). This model assumes that subjects can remember all features of the objects within the 3-4 item limit, or will fail completely.

Another model of the memory sees the capacity as information based and limited by a finite resource that can vary unevenly across different items in a display [28]. Here, subjects reported on a continuous scale. For example, a color have to be recalled in form of selecting from a continuous color palette. It was observed that with increasing set size, the precision of representations decreased. They concluded that subjects could store a continuous amount of information with varying precision: even with set size greater than four items, subjects are able to store more than four items in memory.

All models of the working memory suggest a capacity of 3–4 representations, but this capacity may also be limited by the amount of information load in the display (stimulus complexity) [29]–[32]. Nevertheless, it is possible that items with higher complexity also have higher similarity leading to greater errors.

Former results also showed capacity estimates to be individually different from 1 to 5 objects, furthermore, there is also a large variability within subject in repeated trials [33]. Interestingly, better memory performance was reported when the display consisted of more meaningful stimuli than meaningless images [34]. Moreover, it was demonstrated that training in action video games and even in cognitive training games can contribute to better memory capabilities [35]–[37].

D. Auditory memory

The number of objects or “chunks” humans can remember is limited and depends on the used modality (audio only, visual only, or mixed), presentation method, former training, etc. Working memory refers to the ability to retain stimuli in mind that are no longer physically present and to perform mental operations on them. It allows the temporary storage of relevant information and its task-dependent manipulation.

Regarding the subject of auditory memory, Kaiser summarized the results of several relevant experiments from a neuroscience point of view. Most of these studies have focused on the short-term retention of acoustic information [38]. Auditory memory has to do with the ability to remember words and sounds and to recall information that was received verbally [39], [40]. Zimmermann et al give a good overview of short-term and long-term auditory memory capabilities in different scenarios [41].

Memory capacity limits have been suggested to be “around seven plus or minus two” under various circumstances not limited to auditory tasks indicating a relatively low number of items humans can recall using their memory [42]. Studies show that recognition memory for sounds is usually inferior to memory for visuals [43]–[48]. In [43] four experiments were conducted to examine the nature of auditory and visual memory, including an evaluation of the role of experience in auditory and visual memory, supporting this finding. On the other hand, Lehnert and Zimmer tested the short-term memory of object locations in the auditory and visual modalities pure and mixed, and found the same memory performance in mixed and pure conditions with a very similar decline in performance to the memory load manipulation [49]. They concluded that locations of auditory and visual input are stored in common memory.

Setti et al. compared blind subjects’ and sighted subjects’ performance using semantic and non-semantic sounds to verify if semantic rather than non-semantic sounds could be better recalled, as well as to see whether exposure to an auditory

if semantic rather than non-semantic sounds could be better recalled, as well as to see whether exposure to an auditory

scene could lead to enhanced memorization skills [50]. In the study, semantic sounds were spatialized in order to reproduce an audio scene. Results showed on the one hand that sighted subjects performed better than blind participants following the exploration of the semantic scene. More generally, although both blind and sighted individuals showed similar audio spatial memory skills, blind participants were found to focus more on the perceived sound positions and less on the information that they gathered on the location of individual items during their initial exploration on of the scene. These findings suggest that whereas visual experience allows the simultaneous processing of multiple stimuli, auditory processing is much more sequential.

In a 2007 experiment, 100 students participated in a learning task, where visual icons had to be associated and learned together with their auditory counterparts [51]. The visual stimuli appeared in two sets of 15 icons arranged in 3 columns and 5 rows, sound stimuli were selected from a set of auditory icons (having a semantic relationship with the visual icon) and earcons. Results showed that participants made faster and more correct matches between visual icons and auditory icons than between visual icons and earcons. This suggests the superiority of auditory icons over (non-familiar) earcons. It was also suggested that localization may be a useful cue for learning the associations between them, however, it was not conclusive.

Current studies reported experiments conducted with musicians [52], [53]. Human communicative sounds could be detected better than other sounds, especially in the case of speech and human-generated vocal sounds. Similarly, song-like vocal phrases can be remembered better, and musical training plays a significant role. If rapid pip-tones were presented to subjects, the auditory memory was found to be sensitive to repeating audio structures [54].

The number of sound events and sonification methods is a central problem of the user interface and the audio modeling as well [55]–[61]. VR environments can influence auditory memory performance based on the context-dependent representation [62].

Connected to the “Sound of Vision” research project, a serious game-based application has been developed for testing memory capabilities [63], [64]. The game was the auditory-only version of the memory game, where players have to pair cards in a matrix arrangement (e.g. Matching Pairs, Find the Pair, etc.). Preliminary results showed that users made fewer pairing errors with familiar than with unfamiliar sounds. However, the number of pairs can have a significant impact on the results.

Generally, memory games test and train visual memory in an entertaining way, using only visual modality. Figure 1 shows a real-world audio version of the game in a museum for children. Different fillings result in different noises by shaking the boxes. The same idea is behind the (non-action-based) serious game application being developed.

Fig. 1. A "noisy game" in an exploratory installation for children.

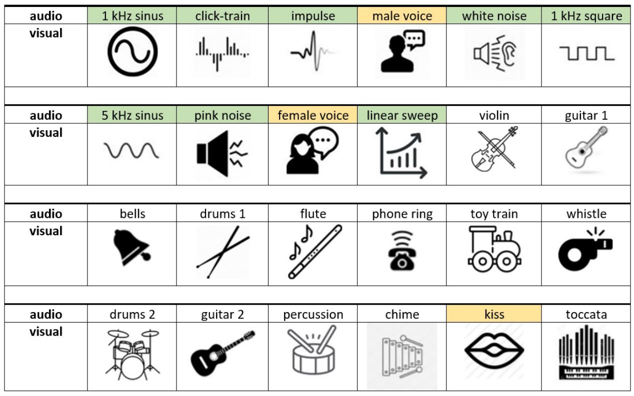

Fig. 2. Summary of the visual and auditory representations on the 6x8 resolution. Yellow marked items are labeled as "human sounds", green marked items are "measurement signals", and all others are everyday sounds called "auditory icons".

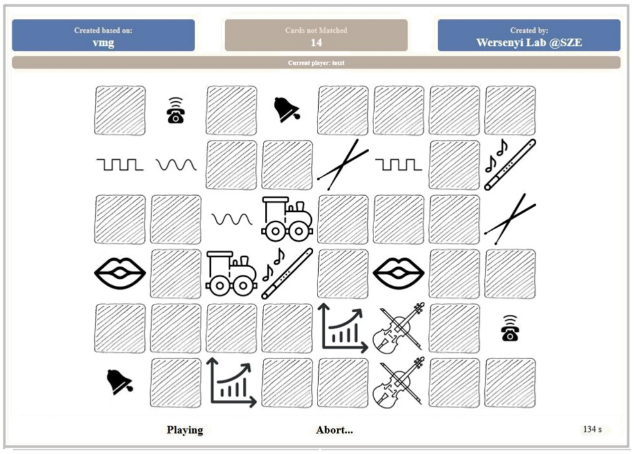

Fig. 3. Screen shot of the user interface during visual gameplay in 6x8 resolution (24 pairs).

The aim of the investigation presented in the next sections targets the evaluation and comparison of

- the memory capabilities of the visual and auditory modality;
- sound stimuli having different frequency content;
- different groups of users (male-female, young-old).

Section 2 introduces the methodology and the measurement setup. In section 3, results will be presented, and section 4 discusses the findings. After the concluding remarks, the main directions for future work will be highlighted.

II. Measurement Setup

A. Game design

The memory game we have developed for the purposes of this experiment relies on a scenario of matrix-arranged cards, with the back of each card initially facing the user. In visual mode, simple black-and-white icons are temporarily shown to users as they flip two cards of their choosing in each round, before the cards return to their original face-down position. In audio-only mode, instead of visual images, short auditory events are played back as the cards are flipped. These iconic sound samples are about 2-4 seconds long each.

The goal of the game is to find matching pairs in each round, by remembering the positions of previously overturned cards, and to do this with the least number of errors (flips). Given that the audio-only mode did not include visual stimuli, a red dot was used to indicate the back of the card that was overturned in each round until both auditory stimuli finished playing. Note that although digital adaptations of this game for single and multiplayer modes have been developed for different platforms, all of them are based on visual representations only. Our application is extended by a modality selector, a set of auditory representations, and the automated logging of results.

Figure 2 shows all visual icons with the corresponding audio sample names as well as their allocation into subgroups. Figure 3 shows an example screenshot of the game in visual mode. In the case of a full HD resolution screen (1080x1920), all cards have a resolution of 80x80 pixels. This size is fixed for all gameplays, independent of the number of cards. The game offers different resolutions from 5x2 up to 6x8, in order to theoretically allow for the comparison of performance as the number of pairs and scope of stimuli types increases. In this experiment, however, only the highest resolution of 6x8 (24 pairs) was used so as to remove the effect of potential confounding factors, like learning effects and/or under- or overrepresentation of stimuli types appearing in multiple resolutions.

After each completed game, a log file was saved on the device with the following records:

- User ID data (nickname, gender, and age);
- Date and time (for controlling later training sessions);
- Time in seconds needed for completing the game;
- Number of total flips (error rate over all pairs);
- Number of flips needed for each sound/icon to be paired.

```
{
    "subjectId": 002,
    "subjectName": "wgy",
    "subjectAge": 47,
    "subjectSex": "Male",
    "timeCost": 134,
    "totalFlips": 52,
    "modality": "audio",
    "rows": 6,
    "cols": 8,
    "timestamp": "Fri Apr 07 2023 12:40:08 CET",
    "numberOfFlipsPerCard": {
     "onekhzsquare": 3,
     "female": 3,
     "impulse1": 7,
     "clicktrain": 5,
     "whitenoise": 4,
     "pinknoise": 3,
     … },
}
```


Fig. 4. Example of the .json file for logging results. This user needed 52 flips and 134 seconds to complete the game in audio only mode. The cards covering the impulse sounds were paired after 7 flips. Data can be easily imported into Excel for statistical analysis.

Figure 4 shows an example log file with recorded data.

B. Participants

All participants were untrained in listening tests, had no formal musical training or any significant gaming experience, but were familiar with the basic idea of the game. No audiometric screening was applied. 40 subjects were included, 20 males and 20 females (mean age 28.85). Furthermore, age groups were created and subjects were allocated to subgroups ”young” (20 subjects below 25) and ”old” (20 subjects above 25). Upon finishing the games, participants were asked informally to give feedback about the experiment (motivation, difficulty, possible changes in the procedure, etc.).

C. Visual and auditory representations

Both visual icons and sound events are included in the levels hierarchically. Every new level (resolution) contains icons/signals from the previous level as well. Because the highest level (6x8) contains all the icons/sounds, it was selected as the basic experiment. Sound events were recorded/generated by the authors, or downloaded from public databases, followed by post-processing. They were selected to represent a variety of sound types, including human sounds, meaningless sounds and everyday sounds. After compiling the sound data base, visual icons were designed with semantic correlation where possible.

D. Methodology

Listening tests were carried out in a silent but non-anechoic room with supervision using the same mobile device (12.4 inch tablet). Users first were informed about the goal of the experiment, but neither the icons nor the sound samples were presented prior the game. All subjects played two games under the exact same circumstances.The first round was always a game in vision-only mode, followed by the audio-only mode. Randomizing the order of modalities was considered, but having the same conditions was considered to be more important given the sample size of the experiment. Moreover, we believe that since the auditory case included no visual stimuli (other than the red dots mentioned previously), there could have been no learning effects or other cross-effects from the visual only mode which was presented to users first. Subjects were instructed to do their best (minimizing the error rate, thus, avoiding wrong flipping of the cards), but otherwise, they were free to choose their gaming strategy and speed. In case of 10 seconds of inactivity, the game would be aborted automatically without logging the results. For maintaining motivation, subjects achieving optimal performance (completing the game without errors) would get a ”perfect game” feedback.

E. Implementation details

In this subsection, we provide a brief summary of some of the technical details of the implementation of the software, both from a technology and algorithmic perspective.

1) Software environment: From a software perspective, the memory game was implemented as a Progressive Web Application (PWA) and shared on Amazon AWS S3. The software was developed based on an application called “vue-memory-game”, shared on Github through the MIT license (https://github.com/leftstick/vue-memory-game). We extended this application with the following features:

- support for pre-defined wave files besides images;
- logging as detailed earlier;
- a “luck management” mechanism detailed in the next subsection;
- a refined user interface.

2) Minimizing the impact of “lucky” initial flips: The game consists of different levels (resolutions). The smallest and easiest is a 5x2 grid with ten cards and 5 pairs to be found. The smaller the resolution, the higher the likelihood that pairs would be found based on pure luck. Finding pairs just by clicking them by chance does not help in evaluating memory capabilities. Although this probability significantly decreases as the number of cards increases (and indeed, in this experiment we used only the highest resolution), the developed software included a correction of this effect of luck.

The algorithm works as follows. Let 2N be the number of cards, thus N is the number of pairs. The main idea is that upon initialization, the cards were not directly connected to the icons/sounds yet – instead, in the first N/2 flips, an icon/sound became associated with the card that was overturned upon demand. In each of these first N/2 flips, the method ensured that no two icons/sounds would be the same.

This method was continued until N/2 if N was even, or (N/2)+1 if N was odd, ensuring that no pairing would be possible until at least half of the icons/sounds were revealed. After this point, the rest of the stimuli were randomly allocated to the remaining cards.

It should be noted that this mechanism operated in a way that was invisible to the users (who could still flip any card

Group	Count	Sum	Mean	Variance	SS
1 kHz sinus	40	280	7,00	8,72	340,00
click-train	40	302	7,55	8,97	349,90
impulse	40	307	7,68	7,56	294,78
male voice	40	276	6,90	5,84	227,60
white noise	40	323	8,08	9,35	364,78
1 kHz square	40	298	7,45	11,43	445,90
5 kHz sinus	40	304	7,60	8,66	337,60
pink noise	40	303	7,58	7,69	299,78
female voice	40	282	7,05	8,92	347,90
linear sweep	40	312	7,80	9,24	360,40
violin	40	291	7,28	6,77	263,98
guitar 1	40	291	7,28	6,82	265,98
bells	40	297	7,43	9,89	385,78
drums 1	40	287	7,18	6,76	263,78
flute	40	293	7,33	11,30	440,78
phone ring	40	303	7,58	9,79	381,78
toy train	40	267	6,68	6,79	264,78
whistle	40	279	6,98	6,85	266,98
drums 2	40	299	7,48	10,41	405,98
guitar 2	40	296	7,40	10,91	425,60
percussion	40	312	7,80	7,86	306,40
chime	40	302	7,55	9,64	375,90
kiss	40	281	7,03	8,59	334,98
toccata	40	300	7,50	9,54	372,00

Fig. 5. Summarized results showing the mean flip values and variances for every visual icon based on 40 participants' gameplays.

Group	Count	Sum	Mean	Variance	SS
1 kHz sinus	40	309	7,72	7,48	291,97
click-train	40	301	7,52	6,66	259,97
impulse	40	311	7,77	8,17	318,97
male voice	40	222	5,55	2,87	111,90
white noise	40	303	7,57	7,17	279,77
1 kHz square	40	305	7,62	9,93	387,37
5 kHz sinus	40	292	7,30	5,75	224,40
pink noise	40	300	7,50	7,28	284,00
female voice	40	218	5,45	3,84	149,90
linear sweep	40	307	7,67	7,60	296,77
violin	40	294	7,35	6,18	241,10
guitar 1	40	319	7,97	5,10	198,97
bells	40	291	7,27	4,97	193,97
drums 1	40	318	7,95	7,99	311,90
flute	40	273	6,82	7,17	279,77
phone ring	40	255	6,37	6,75	263,37
toy train	40	252	6,30	3,44	134,40
whistle	40	277	6,92	9,04	352,77
drums 2	40	303	7,57	6,81	265,77
guitar 2	40	313	7,82	7,27	283,77
percussion	40	282	7,05	6,04	235,90
chime	40	286	7,15	7,26	283,10
kiss	40	226	5,65	2,38	93,10
toccata	40	288	7,20	6,16	240,40

Fig. 6. Summarized results showing the mean flip values and variances for every sound based on 40 participants' gameplays. Red numbers indicate statistically significant difference within the group. The male, female voice samples and the kiss sound could be recalled better than other sounds.

in any location), but it guaranteed that for the first N/2 flips, subjects could not accidentally stumble upon a pair without actually relying on their short-term memory capabilities. Given that the mechanism was also applied to each game in the same way, this introduced no comparative bias into the game at any point.

III. Results

Evaluation of results was based on the number of flips for the total game and for each individual icon/sound using ANOVA and post-hoc analysis using Tukey HSD test. Tukey's HSDF tests all pairwise differences while controlling the probability of making one or more Type I errors. Significance level of 0.05 was set in all paired t-tests and corrected for during the Tukey test.

A. Visual memory

Every participant played the visual-only game first. In order to avoid feature load representations, object load solution was selected. The visual icons (see Figure 2) were similar in size, color and iconic representation. Furthermore, this approach is a modified version of the usual change-detection tasks, where an array of images were shown, but consecutively instead of simultaneously. The statistical analysis showed no difference among the 24 icons (F=0.49; p=0.978). Figure 5 shows descriptive statistics for each icon based on 40 measurements.

Among men, there was a large variability in individual results. 8 out of 10 young and 1 out of 10 old subjects were significantly better than the others. Among females, only one young subject was better. For both genders, younger subjects performed better: males have a mean of 137 (young) and a mean of 193 (old) (F=25.37; p=8.58E-05); females have a mean of 174 (young) and a mean of 193 (old) (F=6.62; p=0.019). Comparing all males and all females, the mean of males (169) is significantly better than the mean of females (189) (F=4.99; p=0.031).

Although mean values of completion time seemed to be quite different, the ANOVA did not support significant difference (246 sec. for males and 275 sec. for females).

B. Auditory memory

Every participant played the audio-only game after completing the visual modality. The audio samples were played back in their entirety after subjects clicked on a given card. Comparing the 24 audio samples, we found three sound samples with the lowest mean values: male voice, female voice, and kiss sound with mean flip numbers of 5.55; 5.45; and 5.65, respectively. All other sounds have means in the range of 6.30 to 7.98 (marked red in Figure 6).

There is, however, no significant difference among these three audio samples. Based on the Tukey-test there was no significant difference among the everyday auditory icons either. It is more important that five of the measurement signals, namely the 1 and 5 kHz sine, click-train, sweep and the 1 kHz square samples were outperformed by the other signals in many (but not all) of the pairwise t-tests. There were significantly higher error rates (number of flips) compared to auditory icons in 6-8 cases for each of these signals.

Among men, only 2 out of 20, and among females, only 1 out of 20 were significantly better than others. In case of males, younger subjects made less errors than older subjects (F= 16.15; p=0.0008), but there was no difference among females. Comparing all males and all females, the mean of males (161) was significantly better than the mean of females (181) (F=7.61; p=0.0087).

Based on the mean completion times, males were significantly faster (mean 440 sec.) than females (mean 491 sec.) based on ANOVA (F=7.30; p=0,010).

Comparing results of all 40 participants between the audio and visual modality, the mean number of flips was equal to 171 for audio-only, and 177 for vision-only. The statistical analysis supported that there was no significant difference between the two modalities (p=0.4).

IV. DISCUSSION

Testing the visual modality supported a-priori assumptions. As expected, there was no difference among the different visual icons, due to the similarity in size and color. The average number of flips needed to find any of the pairs is around 7. The main findings here are that males outperformed females, and younger subjects outperformed older subjects.

There is a vast literature about comparing memory capabilities of different modalities. When viewed together, most of the prior studies focusing on the comparison of visual and auditory memory under various circumstances offer no conclusive results as to one modality being superior to the other. Thus, while a majority of previous experiments showed visual memory to outperform memory in the auditory modality, e.g. [46], [65]–[68]; several other papers reported no difference between the two [69], [70]. In special cases, it has been shown that memory scores could be even higher when processed through the auditory modality in special cases, e.g., for children [71], [72].

In the case of the study reported in this paper, results did not support former findings that visual memory was more reliable than the auditory modality. Comparing the means in Figure 5 and Figure 6, the mean number of flips were around 7 in the case of audio samples as well. Significant improvements were found only in the case of human sounds.

Although not supported by statistical evidence during the paired t-tests in every case, measurement signals also tended to be worse than other audio samples. We can speculate that natural occurring sounds can be recalled and applied better, which supports former findings reported in [51], [55]. It was also expected that similarity would play a significant role. Thus, there were two guitar samples (one distorted, the other not) and three kinds of drum sounds. Furthermore, white and pink noise, sinusoidal and square signals with the same base frequency may have sounded similar, thus, they might have been confused more often according to our expectation. However, this was not supported by the results.

Comparing genders and age groups, males outperformed females. Younger participants delivered better results, but only for men. If we look at the individuals among men and women, there are subjects who are significantly better than others. Interestingly, the variability is greater in case of visual icons: almost half of the males are better than the rest of the group. For females, only 1-2 subjects performed better. This difference between the genders is almost gone in the auditory modality.

If we analyse the factor of age, note that the limit was set to 25 years. Although half of the participants were above this limit, the mean of the age is still low (most of them were below 35). We can assume that a better selection of participants, including more elderly and partitioning them into more age groups (i.e., below 25, between 25-45, above 45) would result in a different outcome. A dedicated experiment is needed for more conclusive results designed for testing the effect of age.

Another factor that was measured is completion time. The two modalities could not be compared, as the mean time is higher for the audio modality due to the playback times of the audio samples. For a conclusive comparison, the visual game should have been delayed after clicking and all sound samples should have been exactly of the same length. Nevertheless, if checking the differences between males and females in the mean completion time in the audio modality, men were significantly faster (F= 7.30; p=0.010).

As mentioned earlier, players are motivated to achieve a ”perfect game”. This occurs if the number of total flips is minimal, all flipped cards could be recalled after the first appearance. According to the informal feedback from the subjects, it is a too difficult task in both modalities in the case of 24 icons/samples. As a matter of fact, even the best players were unable to complete perfect games if the number of pairs exceeds 10 (4x5 resolution). Subjects reported that the game is relatively easy up to 5-6 pairs, but becomes difficult if it has more than 8 pairs.

A follow-up investigation was conducted using all available resolutions. The investigation highlighted the importance of the participant selection process. Specifically, choosing participants from a broader range of age groups and those with diverse experience and training backgrounds can help gather individuals with varying cognitive abilities. In addition, motivation and performance may vary depending on the serious gaming scenario and the user interface used. Comparing gamification methods with traditional memory assessment techniques could reveal their respective advantages and disadvantages in terms of efficiency. A detailed analysis of the follow-up study is planned as future future work.

V. Conclusions

This paper presented the results of an experiment on the auditory and visual short-term memory in a serious game application. 40 subjects played the memory game of ”finding pairs” in a 6x8 resolution with 24 pairs of visual icons, followed by 24 pairs of auditory representations. Based on the number of total flips, results showed no significant difference between the visual and auditory modalities. Nevertheless, in audio-only mode, human sounds could be recalled the best, followed by familiar everyday auditory icons and unfamiliar measurement signals. Completion time could not be associated with the results. Male and younger subjects delivered better results, however, the age limit of 25-years must be increased, and/or a more detailed set of participants is needed for a conclusive outcome.

Future works include the involvement of different (lower) resolutions, the use of mixed modality (audio and vision together), dedicated sessions to test the effect of training and the age of participants, as well as developing a crowdsourcing module for unsupervised data collection.

References

[1] S. Deterding, S. L. Björk, L. E. Nacke, D. Dixon, and E. Lawley, “Designing gamification: creating gameful and playful experiences,” in CHI’13 Extended Abstracts on Human Factors in *Computing Systems*, 2013, pp. 3263–3266.

[2] N. Cowan, “What are the differences between long-term, short-term, and working memory?” *Progress in brain research*, vol. 169, pp. 323–338, 2008. DOI: 10.1016/S0079-6123(07)00020-9

[3] D. Norris, “Short-term memory and long-term memory are still different,” *Psychological bulletin*, vol. 143, no. 9, p. 992, 2017. DOI: 10.1037/bul0000108

[4] D. Burr and D. Alais, “Combining visual and auditory information,” *Progress in brain research*, vol. 155, pp. 243–258, 2006. DOI: 10.1016/S0079-6123(06)55014-9

[5] M. Kubovy and D. Van Valkenburg, “Auditory and visual objects,” *Cognition*, vol. 80, no. 1-2, pp. 97–126, 2001. DOI: 10.1016/S0010-0277(00)00155-4

[6] W. J. Chai, A. I. Abd Hamid, and J. M. Abdullah, “Working mem- ory from the psychological and neurosciences perspectives: a review,” *Frontiers in psychology*, vol. 9, p. 401, 2018. DOI: 10.3389/fp-syg.2018.00401

[7] R. C. Atkinson and R. M. Shiffrin, “Human memory: A proposed system and its control processes,” in *Psychology of learning and motivation*. Elsevier, 1968, vol. 2, pp. 89–195.

[8] P. Kelley, M. D. Evans, and J. Kelley, “Making memories: why time matters,” *Frontiers in human neuroscience*, vol. 12, p. 400, 2018. DOI: 10.3389/fnhum.2018.00400

[9] A. Baddeley, “Working memory,” *Science*, vol. 255, no. 5044, pp. 556–559, 1992. DOI: 10.1126/science.17363

[10] T. Shallice and C. Papagno, “Impairments of auditory-verbal short-term memory: Do selective deficits of the input phonological buffer exist?” *Cortex*, vol. 112, pp. 107–121, 2019. DOI: 10.1016/j.cortex.2018.10.004

[11] P. Baranyi, A. Csapo, T. Budai, and G. Wersényi, “Introducing the concept of internet of digital reality-part I,” *Acta Polytechnica Hungarica*, vol. 18, no. 7, pp. 225–240, 2021. DOI: 10.12700/APH.18.7.2021.7.12

[12] G. Wersényi, A. Csapó, T. Budai, and P. Baranyi, “Internet of digital reality: Infrastructural background–part II,” *Acta Polytechnica Hungarica*, vol. 18, no. 8, pp. 91–104, 2021. DOI: 10.12700/APH.18.8.2021.8.5

[13] F. Bellotti, B. Kapralos, K. Lee, P. Moreno-Ger, and R. Berta, “Assessment in and of serious games: an overview,” *Advances in human-computer interaction*, vol. 2013, pp. 1–1, 2013. DOI: 10.1155/2013/136864

[14] J. B. Hauge, I. A. Stanescu, A. Stefan, M. B. Carvalho, T. Lim, S. Louchart, and S. Arnab, “Serious game mechanics and opportunities for reuse,” *eLearning & Software for Education*, vol. 2, pp. 19–27, 2015. DOI: 10.12753/2066-026X-15-000

[15] A. Dimitriadou, N. Djafarova, O. Turetken, M. Verkuyl, and A. Fer- worn, “Challenges in serious game design and development: Educators’ experiences,” *Simulation & Gaming*, vol. 52, no. 2, pp. 132–152, 2021. DOI: 10.1177/1046878120944197

[16] A. C. T. Klock, I. Gasparini, M. S. Pimenta, and J. Hamari, “Tailored gamification: A review of literature,” *International Journal of Human-Computer Studies*, vol. 144, p. 102495, 2020. DOI: 10.1016/j.ijhcs.2020.102495

[17] A. Rapp, F. Hopfgartner, J. Hamari, C. Linehan, and F. Cena, “Strengthening gamification studies: Current trends and future opportunities of gamification research,” *International Journal of Human-Computer Studies*, vol. 127, pp. 1–6, 2019. DOI: 10.1016/j.ijhcs.2018.11.007

[18] D. Djaouti, J. Alvarez, and J.-P. Jessel, “Classifying serious games: the G/P/S model,” in *Handbook of research on improving learning and motivation through educational games: Multidisciplinary approaches*. IGI global, 2011, pp. 118–136.

[19] D. Ralf, S. Göbel, W. Effelsberg, and J. Wiemers, “Serious games: Foundations, concepts and practice,” 2016.

[20] D. Johnson, E. Horton, R. Mulcahy, and M. Foth, “Gamification and serious games within the domain of domestic energy consumption: A systematic review,” *Renewable and Sustainable Energy Reviews*, vol. 73, pp. 249–264, 2017. DOI: 10.1016/j.rser.2017.01.134

[21] B. Pilote and G. Chiniara, “Chapter 2 - the many faces of simulation,” in *Clinical Simulation (Second Edition)*, G. Chiniara, Ed. Academic Press, 2019, pp. 17–32. ISBN 978-0-12-815657-5

[22] P. M. Staccini and J.-P. Fournier, “Chapter 4 - virtual patients and serious games,” in *Clinical Simulation (Second Edition)*, G. Chiniara, Ed. Academic Press, 2019, pp. 41–51. ISBN 978-0-12-815657-5

[23] M. W. Schurgin, “Visual memory, the long and the short of it: A review of visual working memory and long-term memory,” *Attention, Perception, & Psychophysics*, vol. 80, pp. 1035–1056, 2018. DOI: 10.3758/s13414-018-1522-y

[24] T. F. Brady, T. Konkle, J. Gill, A. Oliva, and G. A. Alvarez, “Visual long-term memory has the same limit on fidelity as visual working memory,” *Psychological science*, vol. 24, no. 6, pp. 981–990, 2013. DOI: 10.1177/0956797612465439

[25] W. J. Ma, M. Husain, and P. M. Bays, “Changing concepts of working memory,” *Nature neuroscience*, vol. 17, no. 3, pp. 347–356, 2014. DOI: 10.1038/nn.3655

[26] S. J. Luck and E. K. Vogel, “The capacity of visual working memory for features and conjunctions,” *Nature*, vol. 390, no. 6657, pp. 279–281, 1997. DOI: 10.1038/36846

[27] D. Fougnie, J. W. Suchow, and G. A. Alvarez, “Variability in the quality of visual working memory,” *Nature communications*, vol. 3, no. 1, p. 1229, 2012. DOI: 10.1038/ncomms2237

[28] P. Wilken and W. J. Ma, “A detection theory account of change detection,” *Journal of vision*, vol. 4, no. 12, pp. 11–11, 2004. DOI: 10.1167/4.12.11

[29] G. A. Alvarez and P. Cavanagh, “The capacity of visual short-term memory is set both by visual information load and by number of objects,” *Psychological science*, vol. 15, no. 2, pp. 106–111, 2004. DOI: 10.1111/j.0963-7214.2004.01502006.x

[30] T. F. Brady and G. A. Alvarez, “No evidence for a fixed object limit in working memory: Spatial ensemble representations inflate estimates of working memory capacity for complex objects,” *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 41, no. 3, p. 921, 2015. DOI: 10.1037/xlm0000075

[31] K. O. Hardman and N. Cowan, "Remembering complex objects in visual working memory: Do capacity limits restrict objects or features?" *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 41, no. 2, p. 325, 2015. DOI: 10.1037/xlm0000031

[32] K. Fukuda, E. Awh, and E. K. Vogel, "Discrete capacity limits in visual working memory," *Current opinion in neurobiology*, vol. 20, no. 2, pp. 177–182, 2010. DOI: 10.1016/j.conb.2010.03.005

[33] T. F. Brady, T. Konkle, and G. A. Alvarez, "A review of visual memory capacity: Beyond individual items and toward structured representations," *Journal of vision,* vol. 11, no. 5, pp. 4–4, 2011. DOI: 10.1167/11.5.4

[34] I. E. Asp, V. S. Störmer, and T. F. Brady, "Greater visual working memory capacity for visually matched stimuli when they are perceived as meaningful," *Journal of cognitive neuroscience*, vol. 33, no. 5, pp. 902–918, 2021. DOI: 10.1162/jocn a 01693

[35] K. J. Blacker, K. M. Curby, E. Klobusicky, and J. M. Chein, "Effects of action video game training on visual working memory," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 40, no. 5, p. 1992, 2014. DOI: 10.1037/a0037556

[36] Y. Yao, R. Cui, Y. Li, L. Zeng, J. Jiang, N. Qiu, L. Dong, D. Gong, G. Yan, W. Ma et al., "Action real-time strategy gaming experience related to enhanced capacity of visual working memory," *Frontiers in Human Neuroscience*, vol. 14, p. 333, 2020. DOI: 10.3389/fn-hum.2020.00333

[37] J. Mishra, D. Bavelier, and A. Gazzaley, "How to assess gaming-induced benefits on attention and working memory," *Games for Health: Research, Development, and Clinical Applications*, vol. 1, no. 3, pp. 192–198, 2012. DOI: 10.1089/g4h.2011.0033

[38] J. Kaiser, "Dynamics of auditory working memory," *Frontiers in psychology*, vol. 6, p. 613, 2015. DOI: 10.3389/fpsyg.2015.00613

[39] S. E. McAdams and E. E. Bigand, "Thinking in sound: The cognitive psychology of human audition," in *Based on the fourth workshop in the Tutorial Workshop series organized by the Hearing Group of the French Acoustical Society*. Clarendon Press/Oxford University Press, 1993. DOI: 10.1093/acprof:oso/9780198522577.001.0001

[40] W. Ritter, D. Deacon, H. Gomes, D. C. Javitt, and H. G. Vaughan Jr, "The mismatch negativity of event-related potentials as a probe of transient auditory memory: A review," *Ear and hearing*, vol. 16, no. 1, pp. 52–67, 1995. DOI: 10.1097/00003446-199502000-00005

[41] J. F. Zimmermann, M. Moscovitch, and C. Alain, "Attending to auditory memory," *Brain research*, vol. 1640, pp. 208–221, 2016. DOI: 10.1016/j.brainres.2015.11.032

[42] N. Cowan, "The magical number 4 in short-term memory: A reconsideration of mental storage capacity," *Behavioral and brain sciences*, vol. 24, no. 1, pp. 87–114, 2001. DOI: 10.1017/S0140525X01003922

[43] M. E. Gloede, E. E. Paulauskas, and M. K. Gregg, "Experience and information loss in auditory and visual memory," *Quarterly Journal of Experimental Psychology*, vol. 70, no. 7, pp. 1344–1352, 2017. DOI: 10.1080/17470218.2016.1183686

[44] K. C. Backer and C. Alain, "Attention to memory: orienting attention to sound object representations," Psychological research, vol. 78, no. 3, pp. 439–452, 2014. https://doi.org/10.1007/s00426-013-0531-7

[45] J. L. Burt, D. S. Bartolome, D. W. Burdette, and J. R. Comstock Jr, "A psychophysiological evaluation of the perceived urgency of auditory warning signals," *Ergonomics*, vol. 38, no. 11, pp. 2327–2340, 1995. DOI: 10.1080/00140139508925271

[46] M. A. Cohen, T. S. Horowitz, and J. M. Wolfe, "Auditory recognition memory is inferior to visual recognition memory," *Proceedings of the National Academy of Sciences*, vol. 106, no. 14, pp. 6008–6010, 2009. DOI: 10.1073/pnas.0811884106

[47] N. Cowan, "Visual and auditory working memory capacity," *Trends in cognitive sciences*, vol. 2, no. 3, p. 77, 1998. DOI: 10.1016/S1364-6613(98)01144-9

[48] B. G. Shinn-Cunningham, "Object-based auditory and visual attention," *Trends in cognitive sciences*, vol. 12, no. 5, pp. 182–186, 2008. DOI: 10.1016/j.tics.2008.02.003

[49] G. Lehnert and H. D. Zimmer, "Auditory and visual spatial working memory," *Memory & cognition*, vol. 34, pp. 1080–1090, 2006. DOI: 10.3758/BF03193254

[50] W. Setti, L. F. Cuturi, E. Cocchi, and M. Gori, "A novel paradigm to study spatial memory skills in blind individuals through the auditory modality," *Scientific reports*, vol. 8, no. 1, pp. 1–10, 2018. DOI: 10.1038/s41598-018-31588-y

[51] T. L. Bonebright and M. A. Nees, "Memory for auditory icons and earcons with localization cues," in *13th International Conference on Auditory Display (ICAD13)*, Montréal, Canada, June 26-29, 2007. Georgia Institute of Technology, 2007, pp. 419–422.

[52] C. M. Vanden Bosch der Nederlanden, C. Zaragoza, A. Rubio-Garcia, E. Clarkson, and J. S. Snyder, "Change detection in complex auditory scenes is predicted by auditory memory, pitch perception, and years of musical training," *Psychological Research*, vol. 84, pp. 585–601, 2020. DOI: 10.1007/s00426-018-1072-x

[53] F. Haiduk, C. Quigley, and W. T. Fitch, "Song is more memorable than speech prosody: Discrete pitches aid auditory working memory," *Frontiers in psychology*, vol. 11, p. 586723, 2020. DOI: 10.3389/fp-syg.2020.586723

[54] R. Bianco, P. M. Harrison, M. Hu, C. Bolger, S. Picken, M. T. Pearce, and M. Chait, "Long-term implicit memory for sequential auditory patterns in humans," *ELife*, vol. 9, p. e56073, 2020. DOI: 10.7554/eLife.56073

[55] W. W. Gaver, "Auditory icons: Using sound in computer interfaces," *ACM SIGCHI Bulletin*, vol. 19, no. 1, p. 74, 1987. DOI: 10.1145/28189.1044809

[56] M. M. Blattner, D. A. Sumikawa, and R. M. Greenberg, "Earcons and icons: Their structure and common design principles," *Human–Computer Interaction*, vol. 4, no. 1, pp. 11–44, 1989. DOI: 10.1207/s15327051hci0401_1

[57] Á. Csapó, G. Wersényi, H. Nagy, and T. Stockman, "A survey of assistive technologies and applications for blind users on mobile platforms: a review and foundation for research," *Journal on Multimodal User Interfaces*, vol. 9, pp. 275–286, 2015. DOI: 10.1007/s12193-015-0182-7

[58] J. Sterkenburg, M. Jeon, and C. Plummer," Auditory emoticons: Iterative design and acoustic characteristics of emotional auditory icons and earcons," in *Human-Computer Interaction. Advanced Interaction Modalities and Techniques: 16th International Conference*, *HCI International 2014, Heraklion, Crete, Greece, June 22-27, 2014, Proceedings, Part II 16*. Springer, 2014. pp. 633–640. DOI: 10.1007/978-3-319-07230-2_60

[59] T. Hermann, A. Hunt, and J. G. Neuhoff, *The sonification handbook*. Logos Verlag Berlin, 2011, vol. 1.

[60] J. P. Cabral and G. B. Remijn, "Auditory icons: Design and physical characteristics," *Applied ergonomics*, vol. 78, pp. 224–239, 2019. DOI: 10.1016/j.apergo.2019.02.008

[61] Á. Csapó and G. Wersényi, "Overview of auditory representations in human-machine interfaces," *ACM Computing Surveys (CSUR)*, vol. 46, no. 2, pp. 1–23, 2013. DOI: 10.1145/2543581.2543586

[62] A. Shamei and B. Gick, "The effect of virtual reality environments on auditory memory," *The Journal of the Acoustical Society of America*, vol. 148, no. 4, pp. 2498–2498, 2020. DOI: 10.1121/1.5146928

[63] "Sound of vision. https://soundofvision.net/ (accessed on 20 july 2023)," 2018. [Online]. Available: https://soundofvision.net/

[64] H. Nagy and G. Wersényi, "Evaluation of training to improve auditory memory capabilities on a mobile device based on a serious game application," in *Audio Engineering Society Convention 142*. Audio Engineering Society, 2017.

[65] A. R. Jensen, "Individual differences in visual and auditory memory," *Journal of Educational Psychology*, vol. 62, no. 2, p. 123, 1971. DOI: 10.1037/h0030655

[66] M. E. Gloede and M. K. Gregg, "The fidelity of visual and auditory memory," *Psychonomic Bulletin & Review*, vol. 26, pp. 1325–1332, 2019. DOI: 10.3758/s13423-019-01597-7

[67] J. Rodríguez, T. Gutiérrez, O. Portillo, and E. J. Sánchez, "Learning force patterns with a multimodal system using contextual cues," *International Journal of Human-Computer Studies*, vol. 110, pp. 86–94, 2018. DOI: 10.1016/j.ijhcs.2017.10.007

[68] K. Lindner, G. Blosser, and K. Cunigan, "Visual versus auditory learning and memory recall performance on short-term versus long-term tests," *Modern Psychological Studies*, vol. 15, no. 1, p. 6, 2009.

[69] K. M. Visscher, E. Kaplan, M. J. Kahana, and R. Sekuler, "Auditory short-term memory behaves like visual short-term memory," *PLoS biology*, vol. 5, no. 3, p. e56, 2007. DOI: 10.1371/journal.pbio.0050056

[70] G. Ward, S. Avons, and L. Melling, "Serial position curves in short-term memory: Functional equivalence across modalities," *Memory*, vol. 13, no. 3-4, pp. 308–317, 2005. DOI: 10.1080/09658210344000279

[71] M. J. Watkins and Z. F. Peynircioğlu, "Interaction between presentation modality and recall order in memory span," *The American journal of psychology*, pp. 315–322, 1983. DOI: 10.2307/1422314

[72] R. Pillai and A. Yathiraj, "Auditory, visual and auditory-visual memory and sequencing performance in typically developing children," *International Journal of Pediatric Otorhinolaryngology*, vol. 100, pp. 23–34, 2017. DOI: 10.1016/j.ijporl.2017.06.010

György Wersényi was born in 1975 in Győr, Hungary. He received his MSc degree in electrical engineering from the Technical University of Budapest in 1998 and PhD degree from the Brandenburg Technical University in Cottbus, Germany. Since 2002 he has been member of the Department of Telecommunications at the Széchenyi István University in Győr. From 2020 to 2022 he was the dean of Faculty of Mechanical Engineering, Informatics and Electrical Engineering, as well as the scientific president of the Digital Development Center at the university. Currently, he is a full professor, member of the European Acoustics Association (EAA) and the Audio Engineering Society (AES). His research focus is on acoustic measurements, virtual and augmented reality solutions, sonification, cognitive infocommunications, and assistive technologies.

Ádám Csapó obtained his PhD degree at the Budapest University of Technology and Economics in 2014. Between 2016 and 2022, he has worked as an Associate Professor at the Széchenyi István University, and between 2022 and 2023, at Óbuda University in Budapest, Hungary. Currently he is an Associate Professor at Corvinus University of Budapest. Dr. Csapo's research focuses on soft computing tools for developing cognitive infocommunication channels in assistive technologies and virtual collaboration environments, with the goal of enabling users to communicate with each other and with their spatial surroundings in novel and effective ways. He has also participated in the development of a commercial desktop VR platform serving both educational and industrial use cases. Dr. Csapo has over 100 publications, including 1 co-authored book and over 20 journal papers.