Binaural sound rendering improves immersion in a daily usage of a smartphone video game

Julian Moreira, Laetitia Gros, Rozenn Nicol, Isabelle Viaud-Delmon, Cécile Le Prado, Stéphane Natkin

BINAURAL SOUND RENDERING IMPROVES IMMERSION IN A DAILY USAGE OF A SMARTPHONE VIDEO GAME

Julian Moreira[1] **Laetitia Gros**[2] **Rozenn Nicol**[2]
Isabelle Viaud-Delmon[3] **Cécile Le Prado**[1] **Stéphane Natkin**[1]
[1] Cnam (CEDRIC), Paris, France
[2] Orange Labs, Lannion, France
[3] CNRS, Ircam, Sorbonne Université, Ministère de la Culture, Paris, France
Correspondance: julian.moreira.fr@gmail.com

ABSTRACT

Binaural rendering is a technology that could be advantageously coupled with a smartphone application, but is still not commonly used. The aim of this study is to investigate how this technology could enrich the experience of a video game application delivered on a smartphone, in terms of immersion, memorization and performance. We have used a longitudinal research procedure, the Experience Sampling Method, asking individuals to accomplish short game sessions in their daily life. With this procedure, we want to determine if a significant effect of binaural rendering can be detected while data are noised by realistic contextual variations. The results indicate a better feeling of immersion during the binaural sessions, but no improvement was shown for memorization and performance. However, as the experiment is deployed in realistic contexts of use, this result justifies the implementation of a binaural rendering in smartphone applications.

1. INTRODUCTION

Binaural synthesis is a technology that spatializes sound outside of the head of a listener wearing headphones. It allows to reproduce the acoustic properties of a sound coming from a specific direction by using the Head Related Transfer Functions (HRTF) filters owned by this listener [1]. Being the only solution of spatialized sound to be easily transportable everywhere, it particularly fits with the use of a smartphone, where contexts are numerous, dynamic and unpredictible.

Most of the works evaluating binaural technologies have been dedicated to measure the accuracy of source localization (see for instance [2–5]). However, in a daily life, audiovisual experiences with binaural sounds (video games, movies, videoconferencing, etc.) are not necessarily to be limited to a localization task. As such, a few other studies have focused their attention to the contribution of a binaural rendering (compared to mono or stereo renderings, or a purely visual scene, etc.) by looking at various other attributes: immersion (e.g., [6–8]), navigation performance (e.g., [6, 7]), memorization (e.g., [6]), feeling of spatialisation (e.g., [9]), consistence or correlation with visual (e.g., [9, 10]), global appreciation (e.g., [10]), etc.

To our knowledge, none of these works have been dedicated so far to a binaural rendering coupled with a smartphone application. Yet, smartphone usages raise additional questions related to contextual factors, that might influence the perception of a spatialized sound scene (see for instance [12–17] about the role of some of these factors on binaural-enhanced experiences). Other questions rely on methodology: contexts bring noise, whether it be auditory, visual or even cognitive (e.g., walking and using a smartphone at the same time), that can be hardly reproduced in a laboratory environment. In this paper, we propose an experiment to study the contribution of a binaural rendering to a smartphone video game application, in terms of immersion, memorization and performance. We address the question of influence factors by deploying our experiment in real life situations. By doing this, we want to determine if a significant effect of binaural rendering can be detected while data are noised by realistic contextual variations.

In Section 2, we introduce the notion of data validity and the deployment method we choose. In Section 3 we present our experiment, including our video game, the data we collect and the setup. Section 4 is dedicated to the results, discussed in Section 5.

2. DATA VALIDITY AND DEPLOYMENT METHOD

Data validity and deployment method are closely linked. In [18], Scriven explains that collected data are valid if they are likely to answer the question initially asked. He distinguishes two types of validity: internal and external. Internal validity is linked to the control experimenters have on influence factors, like environmental conditions (location, temperature, luminosity, time of the day, ambiant noise, etc.), expertise of the assessors, or similarity of the conditions when presenting stimuli (order, number of occurences, etc.) The more under control these factors are,

doi:10.25836/sasp.2019.23

the more internally valid the data become. External validity represents how close to a real use case the experiment is. As such, if a laboratory represents an experimental place with a high internal validity, on the contrary external validity is potentially low, especially when it comes to reproduce smartphone usage situations. For this reason, we propose a protocole oriented to external validity, where our experimental sessions are deployed outside the laboratory.

We use the Experience Sampling Method (ESM). This method consists in dividing a long experiment into small sessions over a long period of time, in order to collect data that represent the subjects' daily life [19]. Several times per day, they are notified to accomplish a session that should not exceed a few minutes. The short duration is intended to prevent the experiment to be intrusive, and to keep the motivation of the subjects. With this process, the experience is designed to last one or several weeks. The ESM has already been used a few times with mobile phones, for instance in [20,21]. Advantageously, the whole experiment can be processed on the device: notifications, application usages, additional questionaries, etc. Furthermore, contextual data can be collected, by extracting automatically information from sensors (which requires a solid theoretical model to interpret them) or by directly asking subjects. This is of particular interest to observe trends about daily life usages, and to see if experimental data are consistent with their context.

3. EXPERIMENT

3.1 Progress of the experiment

In our experiment, we use the ESM as follows: during five weeks, subjects are notified twice a day by SMS to play a five minutes session (for a total of 70 sessions per subject). In order to diversify contexts, one session is randomly scheduled in the morning between 8am and 1pm, the other is randomly scheduled in the afternoon between 1pm and 6pm (see Table 1). To maximize the external validity of the data, the experiment takes place on the personal smartphones and headphones of the subjects.

Each session is divided into four phases, all embedded in the same application: the context questionary, the calibration of sound volume, the game itself and the end questionary. The data gathered during all these steps are sent at the end of the session to a database located on a distant server. If a subject is not connected to the internet at this time, the data are stored locally and added for sending to those of the next session.

3.2 Contextual data collection and sound calibration

Contextual data are collected via a preliminary questionary at each session. In order to keep the session short, we limit the number of questions to four, retrieving the current location of the subjects, their social surrounding, their level of mobility and their level of occupation. The purpose of this questionary is to observe trends about smartphone contexts, and because they are not controlled factors (not equally distributed between subjects), they are not meant to be used to intepret the contribution of binaural sounds. As such, they are not in the scope of this paper, and not treated in the next sections.

The sound calibration step consists in a panel that reminds the subjects to plug their headphones, put the sound volume of the phone to its maximum value, and to calibrate the volume of the application via a slider in-app. This helps us to control sound volumes and prevent subjects from muting sound.

3.3 The video game

Our smartphone application is an Infinite Runner developped for the occasion with the game engine Unity (see Figure 1). The player controls an avatar that walks automatically on a procedurally generated road, and with an increasing speed. The goal is to bring the avatar as far as possible, avoiding obstacles and gathering bonus items. To do that, the player can swipe a finger on screen to move the avatar on the right, on the left, make him jump or slide down.

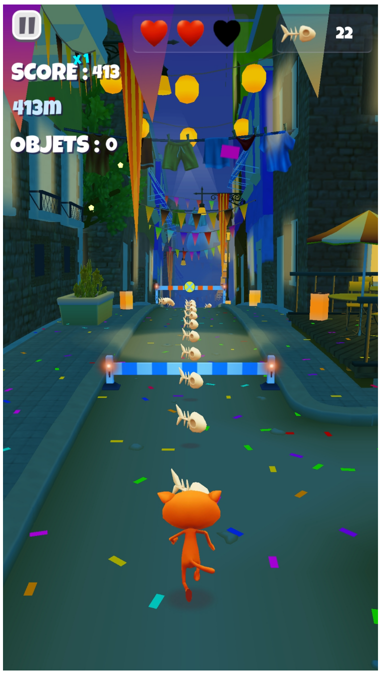

Figure 1. A screenshot of the Infinite Runner used for the experiment.

The game exists in two versions: one with binaural sounds and one in mono. Half of the sessions is presented to the subjects with a binaural rendering, the other half with a mono rendering (see Table 1). Relying on the results from a previous study [22], we set the point of listening of the binaural version to the position of the virtual camera. The sound source positions are set by using a unique set of non-individual HRTF measured from a dummy head Neumann KU 100.

Several areas have been developped: forest (spring and autumn), city (standard and snowy) and seaside town (standard and carnival), all available in a day and a night ver-

 doi:10.25836/sasp.2019.23

sion. As their purpose is only to bring some graphical and sound variety, they are not subject to a precise distribution across sessions. At each session, one of them is picked up randomly and presented to the subject.

3.4 Measuring immersion, memorization and performance

Immersion, memorization and performance are measured via different methods. Immersion is measured via a questionary at the end of the session. Still keeping the idea of a short session, we elaborate four questions, inspired from previous studies on the feeling of immersion in virtual reality [23] or with spatialized sound systems [24]: 1) In the game generated world, you had a sense of ”being there”, 2) Have you experienced a ”3D sound” effect?, 3) Sound did contribute to your feeling of immersion and 4) The external context (ambiant noise, parallel task, etc.) troubled your immersion. Except for the question 2 with a yes/no answer, all responses are given on a discret scale between 0 and 3, 0 being tagged as ”not at all”, and 3 being tagged as ”fully agree”.

Memorization is measured via a task: during the game, various audiovisual objects are randomly spread along the roadside. The objects are randomly picked up among a list of twenty, and are designed to be graphically and aurally emphasised in the scenery, to ease their memorization. At the end of the session, the last seven objects that have been met during the game are presented in a random order to the subjects, and they are asked to sort them.

Immersion and memorization both require an answer from the subjects at the end of the session. In order to prevent the subject to anticipate them (bringing a potential bias in their responses), over the 70 sessions, 30 sessions have the immersion questionary only, 30 others have the memorization task only, and the 10 remaining ones don’t have any questionary (see Table 1).

Finally, performance is computed as a mix between the bonus collected and the distance traveled in the game.

In this experiment we favor external validity rather than internal validity, and we expect this choice to bring noise in our experimental data. By measuring the contribution of binaural with three different methods (a subjective questionary for immersion, a task for memorization and interactivity data for performance), we aim to obtain different kind of information, one of the method being hopefully less impacted by the influence factors.

3.5 Summary: distribution of the sessions

Table 1 summarizes the distribution of the 70 sessions considering the type of audio rendering, the period of the day and the end session questionary. In the experiment, all the sessions were presented in a random order to each subject.

Number of sessions	Sound type	Half day	End questionary
7	binaural	morning	immersion
8	binaural	afternoon	immersion
8	binaural	morning	memorization
7	binaural	afternoon	memorization
2	binaural	morning	nothing
3	binaural	afternoon	nothing
8	mono	morning	immersion
7	mono	afternoon	immersion
7	mono	morning	memorization
8	mono	afternoon	memorization
3	mono	morning	nothing
2	mono	afternoon	nothing

Table 1. Distribution of the 70 ESM sessions.

3.6 Subjects

Thirty subjects (8 women) take part in the experiment, aged between 16 and 67 (average 28.1). Among them, five are unpaid volonteers from the laboratory, the other twenty-five are externals and receive a 50 EUR voucher at the end of the experiment, to compensate for their time. The experiment are conducted in compliance with the Declaration of Helsinki as well as national and institutional guidelines for experiments with human subjects.

Subjects are recruited based on the following requirements: owning headphones and a smartphone that runs on Android OS, and having a minimal knowledge about mobile video games. Before starting the core experiment, they all are invited to come at the laboratory to get the game installed on their phone (except for five subjects who install the game by themselves, following instructions via videoconferencing). Subjects are also informed about the details of the experiment, but not about its purpose. They are instructed to accomplish their sessions as much as possible immediately after receiving the SMS notification. However, to make the experiment more fluid, any delay or anticipation are permitted within the half day of the session. Exceeding the half day, the session is postponed to the next day, extending the whole experiment at the same time.

3.7 Hypotheses

Comparing mono to binaural sessions, we provide the following hypotheses:

- for immersion, we expect better responses in the binaural sessions, resulting in a better sense of presence, a sound effect more often experienced, a better contribution of sound to the feeling of immersion and an external context less disturbing;
- for memorization, we expect objects to be better memorized in the binaural sessions;
- for performance, we expect a better global score in the binaural sessions.

4. RESULTS

Results comprise 2100 sessions (70 sessions for 30 subjects). For technical reasons, twelve sessions have been lost, owned by a unique subject, resulting in 2088 sessions. Among them, 896 are provided with an immersion questionary, 894 with a memorization questionary and 298 sessions with no questionary.

4.1 Immersion

Figure 2 shows the answer of the subjects to the four questions related to their feeling of immersion. We observe that on average, subjects experienced a better sense of presence (question 1), a better contribution of sound to immersion (question 3) and a lower trouble caused by the external context (question 4) when audio was rendered in binaural. The answers to the question 2 reveal that most of the time, the subjects detected a 3D sound effect, whatever the audio type is, binaural or mono (366 detections of a 3D sound effect in binaural sessions, i.e., 83% of all the binaural sessions, and 349 detections in mono sessions, i.e., 76% of all the mono sessions).

Three ANOVA were performed, with the answers of questions 1, 3 and 4 as successive dependent variables, the audio rendering type as a within-subject factor, and the subject as a random factor. Results indicate a significant effect of the audio type on the sense of presence (F(1, 29.03)=6.31, $p<0.05$), on the contribution of sound to immersion (F(1, 29.04)=5.05, $p<0.05$), but not on the trouble caused by the external context (F(1, 29.06)=1.09, p=0.30). Finally, a χ^2 test is performed for question 2, with the answer type (yes or no) as the dependent variable and audio rendering type as the independent variable. Results reveal a highly probable reject of the null hypothesis (χ^2=6.81, $p<0.01$), meaning that the audio rendering type has a statistically significant effect on the answers. In other words, subjects detect significantly more often a 3D sound effect when sessions are in binaural.

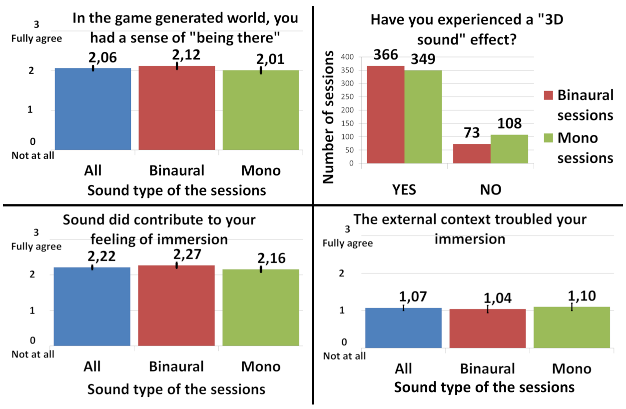

Figure 2. Subjects' answers to the four immersion related questions. The answers are systematically given in function of the type of audio rendering (mono or binaural). The answers were given on a discrete scale between 0 and 3, except for the question 2 (on the top right) where the answer was yes or no. Vertical bars are the 95% confidence intervals.

4.2 Memorization

For memorization, we have to compare the sequence of objects memorized by the subjects with the correctly ordered sequence. We compare them following several metrics: the Hamming distance [26], the Levenshtein distance [27], the Damerau-Levenshtein distance [28], the longest common subsequence [29], and the Jaro similarity [30]. The basic principle of these metrics is to compare what objects are in common in the two sequences (their number, their distance), and how many operations are required to convert one sequence to another. More details can be found in the papers associated to each method. Figure 3 shows the average normalized distances, given the audio type of the sessions. All data represent here a distance, i.e., a value of 0 if sequences are the same (good memorization), a value of 1 if sequences are very different (bad memorization). We observe for each distance that values are nearly the same for both audio renderings, but systematically lower for the mono sessions. Five ANOVA are performed, with the distances as successive dependent variables and the audio rendering type as a within-subject factor and the subject as a random factor. No significant effect of the audio rendering type is revealed, whatever the distance is: Hamming (F(1, 29.24)=0.10, p=0.76), Levenshtein (F(1, 29.22)=0.71, p=0.41), Damerau-Levenhstein (F(1, 29.21)=0.45, p=0.51), longest common subsequence (F(1, 29.28)=2.89, p=0.10) and Jaro distance (F(1, 29.21)=0.34, p=0.56).

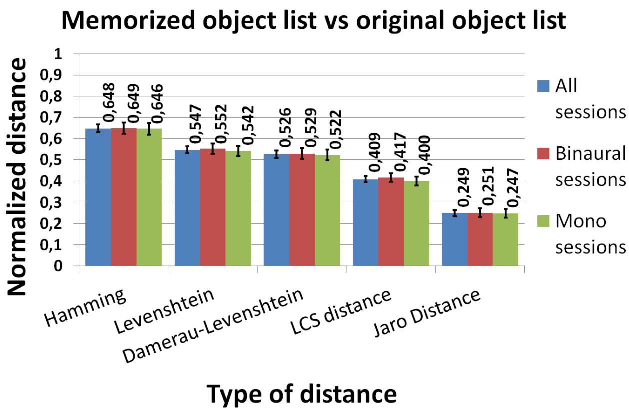

Figure 3. Average normalized distances between the memorized sequences of objects and the correctly sorted sequences. A distance of 0 means that the two sequences are identical (good memorization), and a distance of 1 means that the two sequences are highly different (bad memorization). Vertical bars are the 95% confidence intervals.

4.3 Performance

Performance of the subjects as a function of the audio rendering type can be observed in Figure 4. Values are better for binaural sessions, but here again, an ANOVA performed on data reveals no effect of the audio type (F(1, 28.05)=0.18, p=0.68).

doi:10.25836/sasp.2019.23

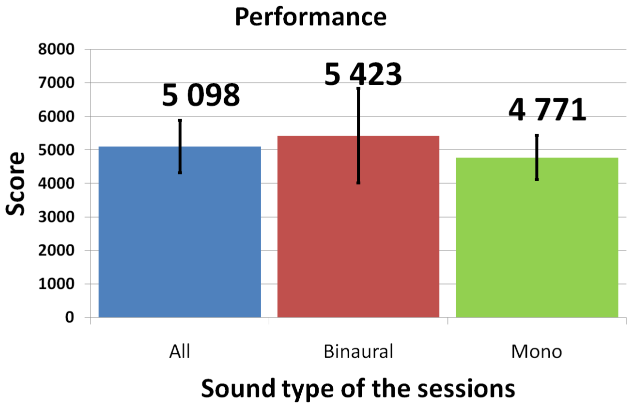

Figure 4. Vertical bars are the 95% confidence intervals.

5. DISCUSSION

Among all the questions and variables, it seems that only the sense of presence and the contribution of sound to immersion are significantly improved by the binaural rendering. Besides, a 3D effect is more often experienced in binaural sessions, meaning that subjects consciously perceive the sound spatialization.

On the contrary, memorization is not influenced by the audio rendering. One possible reason is the difficulty of the task: remembering the order of 7 objects is probably too hard, considering that 7 is usually seen as the upper limit of the memory span [31]. This was confirmed by several subjects after the experiment, during an informal debriefing. We also raise the idea that we might have mitigated the beneficial effect of the binaural rendering, by positioning objects only on the right or left of the avatar's track, instead of all around the user.

For the performance, no effect of the audio rendering type was neither found. However, many influence factors come under consideration in this experiment. First, we suppose that the difficulty of the game (and consequently performance) is probably related to the context of use, which was highly variable and not possible to study systematically in our experiment. Second, several subjects confessed having stopped their sessions intentionally, for three different reasons: some of them found that the game was too easy; some others had sometimes to stop for external reasons (end of the bus ride, end of the break at work, etc.); and some others, focused on the memorization task, died purposely to keep in mind the 7 memorization objects. In these conditions, the interpretation of the performance regarding the binaural rendering is quite difficult. In further studies, the conflict between the memorization task and the game itself may be softened by integrating the former in the core gameplay of the latter.

Finally, among the three assessed attributes though, having a significant improvement of immersion in the binaural condition is an important result. It means that in realistic use cases, despite the large variability of contexts (different places, time, situations, smartphones, headphones, inter-individual differences, etc.), binaural rendering still has a positive effect, justifying its implementation in a smartphone application.

6. CONCLUSION

We proposed an experiment to measure the contribution of a binaural rendering in a smartphone video game application. The video game was an Infinite Runner, largely spread among the general public. We measured three attributes, i.e., immersion, memorization and performance, using the Experience Sampling Method, that aims to replace the experiment in realistic contexts of use. Results revealed a better feeling of immersion with the binaural rendering, compared to the same game in mono. No significant effect was found for memorization and performance. As the study focused on external validity (realistic conditions) rather than internal validity (controlled conditions), we claim that this result justifies the use of binaural sounds in this kind of smartphone applications.

Many prospects are possible. Considering the contribution of the binaural sound, further studies should assess other attributes, other smartphone applications, and compare binaural rendering to other sound systems (e.g., stereo). Other studies may focus on the method itself, investigating how to gradually control more conditions, while keeping the external validity at the same thime.

7. REFERENCES

[1] J. Blauert, *The technology of binaural listening*. Springer, 2013.

[2] F. L. Wightman and D. J. Kistler, "Headphone simulation of free-field listening. ii: Psychophysical validation," *The Journal of the Acoustical Society of America*, vol. 85, no. 2, pp. 868–878, 1989.

[3] H. Møller, C. B. Jensen, D. Hammershøi, and M. F. Sørensen, "Using a typical human subject for binaural recording," in *Audio Engineering Society Convention 100*, Audio Engineering Society, 1996.

[4] E. H. Langendijk and A. W. Bronkhorst, "Fidelity of three-dimensional-sound reproduction using a virtual auditory display," *The Journal of the Acoustical Society of America*, vol. 107, no. 1, pp. 528–537, 2000.

[5] R. L. Martin, K. I. McAnally, and M. A. Senova, "Free-field equivalent localization of virtual audio," *Journal of the Audio Engineering Society*, vol. 49, no. 1/2, pp. 14–22, 2001.

[6] P. Larsson, D. Vastfjall, and M. Kleiner, "Better presence and performance in virtual environments by improved binaural sound rendering," in *Audio Engineering Society Conference: 22nd International Conference: Virtual, Synthetic, and Entertainment Audio*, Audio Engineering Society, 2002.

[7] A. Gonot, M. Emerit, and N. Château, “Usability of 3d-sound for navigation in a constrained virtual environment,” in *AES, 120th Audio Engineering Society Convention, Paris*, Audio Engineering Society, 2006.

[8] S. Moulin, R. Nicol, and L. Gros, “Spatial audio quality in regard to 3d video,” in *Acoustics 2012*, 2012.

[9] M. Cobos, J. J. Lopez, J. M. Navarro, and G. Ramos, “Subjective quality assessment of multichannel audio accompanied with video in representative broadcasting genres,” *Multimedia Systems*, vol. 21, no. 4, pp. 363–379, 2015.

[10] F. Grani, F. Argelaguet, V. Gouranton, M. Badawi, R. Gaugne, S. Serafin, and A. Lecuyer, “Design and evaluation of binaural auditory rendering for caves,” in *2014 IEEE Virtual Reality (VR)*, pp. 73–74, IEEE, 2014.

[11] U. Reiter, K. Brunnström, K. De Moor, M.-C. Larabi, M. Pereira, A. Pinheiro, J. You, and A. Zgank, “Factors influencing quality of experience,” in *Quality of Experience*, pp. 55–72, Springer, 2014.

[12] D. R. Begault, E. M. Wenzel, and M. R. Anderson, “Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source,” *Journal of the Audio Engineering Society*, vol. 49, no. 10, pp. 904–916, 2001.

[13] C. Guastavino, *Étude sémantique et acoustique de la perception des basses fréquences dans l’environnemet sonore urbain*. PhD thesis, Université Paris 6, 2003.

[14] T. Walton, M. Evans, D. Kirk, and F. Melchior, “Exploring object-based content adaptation for mobile audio,” *Personal and Ubiquitous Computing*, vol. 22, no. 4, pp. 707–720, 2018.

[15] S. Werner and F. Klein, “Influence of context dependent quality parameters on the perception of externalization and direction of an auditory event,” in *Audio Engineering Society Conference: 55th International Conference: Spatial Audio*, Audio Engineering Society, 2014.

[16] A. Fiebig, “Influence of context effects on sound quality assessments,” in *Proceedings of EuroNoise*, pp. 2555–2560, 2015.

[17] T. Walton and M. Evans, “The role of human influence factors on overall listening experience,” *Quality and User Experience*, vol. 3, no. 1, p. 1, 2018.

[18] F. Scriven, “Two types of sensory panels or are there more?,” *Journal of sensory studies*, vol. 20, no. 6, pp. 526–538, 2005.

[19] R. Larson and M. Csikszentmihalyi, “The experience sampling method,” in *Flow and the foundations of positive psychology*, pp. 21–34, Springer, 2014.

[20] S. Ickin, K. Wac, M. Fiedler, L. Janowski, J.-H. Hong, and A. K. Dey, “Factors influencing quality of experience of commonly used mobile applications,” *IEEE Communications Magazine*, vol. 50, no. 4, pp. 48–56, 2012.

[21] J. Traer and J. H. McDermott, “Statistics of natural reverberation enable perceptual separation of sound and space,” *Proceedings of the National Academy of Sciences*, vol. 113, no. 48, pp. E7856–E7865, 2016.

[22] J. Moreira, L. Gros, R. Nicol, and I. Viaud-Delmon, “Spatial auditory-visual integration: The case of binaural sound on a smartphone,” in *Audio Engineering Society Convention 145*, Audio Engineering Society, 2018.

[23] B. G. Witmer and M. J. Singer, “Measuring presence in virtual environments: A presence questionnaire,” *Presence*, vol. 7, no. 3, pp. 225–240, 1998.

[24] R. Nicol, L. Gros, C. Colomes, E. Roncière, and J. Messonier, “Etude comparative du rendu de différentes techniques de prise de son spatialisée après binauralisation,” *CFA/VISHNO*, 2016.

[25] N. Liu, Y. Liu, and X. Wang, “Data logging plus e-diary: towards an online evaluation approach of mobile service field trial,” in *Proceedings of the 12th international conference on Human computer interaction with mobile devices and services*, pp. 287–290, ACM, 2010.

[26] R. W. Hamming, “Error detecting and error correcting codes,” *The Bell system technical journal*, vol. 29, no. 2, pp. 147–160, 1950.

[27] V. I. Levenshtein, “Binary codes capable of correcting deletions, insertions, and reversals,” in *Soviet physics doklady*, vol. 10, pp. 707–710, 1966.

[28] F. J. Damerau, “A technique for computer detection and correction of spelling errors,” *Communications of the ACM*, vol. 7, no. 3, pp. 171–176, 1964.

[29] L. Bergroth, H. Hakonen, and T. Raita, “A survey of longest common subsequence algorithms,” in *Proceedings Seventh International Symposium on String Processing and Information Retrieval. SPIRE 2000*, pp. 39–48, IEEE, 2000.

[30] M. A. Jaro, “Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida,” *Journal of the American Statistical Association*, vol. 84, no. 406, pp. 414–420, 1989.

[31] G. A. Miller, “The magical number seven, plus or minus two: Some limits on our capacity for processing information.,” *Psychological review*, vol. 63, no. 2, p. 81, 1956.

doi:10.25836/sasp.2019.23