From Automatic Sound Analysis of Gameplay Footage [Echos] to the Understanding of Player Experience [Ethos]: an Interdisciplinary Approach to Feedback-Based Gameplay Metrics

Raphael Marczak, Pierre Hanna, Jean-Luc Rouas, Jasper van Vught, Gareth Schott

From Automatic Sound Analysis of Gameplay Footage [Echos] to the Understanding of Player Experience [Ethos]: an Interdisciplinary Approach to Feedback-Based Gameplay Metrics

Raphael Marczak
School of Arts
University of Waikato
New Zealand
raphaelm@waikato.ac.nz

Pierre Hanna
LaBRI
University of Bordeaux
France
pierre.hanna@labri.fr

Jean-Luc Rouas
LaBRI
University of Bordeaux
France
jean-luc.rouas@labri.fr

Jasper van Vught
School of Arts
University of Waikato
New Zealand
jasperv@waikato.ac.nz

Gareth Schott
School of Arts
University of Waikato
New Zealand
g.schott@waikato.ac.nz

ABSTRACT

In line with the ICMC|SMC|2014 conference theme "from digital echos to virtual ethos", and the conference interdisciplinary main objective; the present paper is seeking to demonstrate that the sound feedback stream produced by videogames when activated by players (echos) can be automatically analyzed in order to study how sound can, not only, describe a gameplay performance, but also help to understand player experience and emotions (ethos). To do so, the present paper illustrates how sound processing algorithms can be applied in the game studies discipline in order to assess and understand better how players engage with videogames. The present paper proposes to adapt the Feedback-based Gameplay Metrics method, successfully applied to the analysis of gameplay footage video stream [17], to the sound stream, via the automatic detection of musical sequences and speech segment.

1. INTRODUCTION

Over a period of several decades videogames have been established as an important new part of our everyday lives. Although still showing clear similarities to other more traditional manifestations of culture such as traditional games and fictions, videogames are digital media with their own distinctive (combination of) qualities [1]. Games require the player to perform recursive actions [2, 3] that lead to polysemic performances and readings [4]. It is this hybrid nature of games that has encouraged their study from a range of different fields and with a range of different approaches.

In general, however, the study of games has put specific emphasis on the active role of the player in its production. This means that game research has focused much attention on seeking a better understanding of the different experiential components of gameplay (e.g., enjoyment [5] flow, [6] or immersion [7]) and the player's behavior as pivotal aspect to these experiences [8]. In an attempt to achieve precision accounts of player behavior and discern the game's experiential triggers, different quantitative techniques such as the analysis of gameplay metrics [9] and biometrics [10] have been employed. Modeling the player's behavior with the use of gameplay metrics allows for an examination of how players actually activate the games under investigation thereby creating a better understanding of the type of content encountered by players. For a more exhaustive analysis, these quantitative approaches are even sometime combined to build an empirically based understanding of the cognitive and affective impact that games exert during play (like Biometric Storyboard [11]).

These quantitative methodologies do however have their limitations. Gameplay metrics do for instance not account for reasons or motivations behind the player's behavior [8] nor are there any psychophysiological measures that differentiate the intentionality of the measured outputs. But these are limitations imposed by an exclusive use of quantitative methodologies that can simply be overcome by triangulating different qualitative and quantitative techniques [11, 12]. More concerning is however the fact that the use of gameplay metrics is reserved only for those with access to the source code of the studied game, or at least a modding system. It might be for this reason that gameplay metrics have so far mostly been used for playability research [9, 13, 14, 15] rather than research towards player experience [16].

To counteract this research tendency and expand the use of gameplay metrics to other research domains Marczak et al. [17] have previously proposed an alternative methodology to acquire gameplay metrics through an analysis of the game's video feedback. This opens up the possibility to use gameplay metrics in the study of any commer-

cially available game without the need for the source code or a strong collaboration with the relevant game company. Although these video feedback gameplay metrics are useful and accurate in segmenting the gameplay experience and quantifying elements of the player's actions, they do limit the gameplay experience to the graphical representations on screen. This means that these gameplay metrics quantify gameplay with a focus on the game's 'visualism'. It is however a mistake to assume that games 'present only one type of experience and foster one type of engagement' [18]. So to move beyond this 'visualism' as a focal point for exposing the mechanics by which games operate we propose a new way of acquiring gameplay metrics through an analysis of the audio feedback of the game.

Although game audio is still a hugely under-researched subject, it is clear that sound plays an important role in our gameplay experience. Sound can help us orient our actions or identify certain game functions [19], it plays an important role in triggering certain emotions [20], it has an essential 'preparatory function' [21], and it can significantly increase the 'immersiveness of the gameplay experience' [22]. Gathering gameplay metrics from the game's audio feedback therefore allows us to segment the player actions according to an essential component of the gameplay experience.

Signal processing of audio streams constitutes an established and active research field within computer science. Algorithms have been created to detect speech and music in radiophonic streams [23], to retrieve artist and song title information from music recognition [24, 25] or to indicate moments of 'story intensity' in movies through audio tempo analysis [26]. But these algorithms have yet to be applied to videogames.

In this paper we outline the construction of two different algorithms used to acquire gameplay metrics from the game's audio feedback stream. By implementing the algorithms for the analysis of the game *Bioshock 2* (2K Games, 2010), we were able to show the usability of the methodology for the obtainment of gameplay metrics. This means that this methodology can accurately calculate the temporal position of encountered game content, which can significantly help us in our understanding of the gameplay experience.

2. AUTOMATIC AUDIO FEEDBACK STREAM ANALYSIS

2.1 Sequence detection system

One possible way to segment gameplay sessions and detect key-moments of interest and their temporal locations, is to identify parts of the gameplay session based on their musical and atmospheric audio rendition. For this detection to work automatically, the system is required to have prior knowledge of the different musical parts of the game under investigation. The main difficulty for the system to learn and subsequently detect gameplay moments on their audio rendition is the irregular presence of noise in the form of gunshots, screams, speech, etc. The perceived intensity level of this noise is usually higher than that of the musical environment which means the audio recognition system applied has to be very robust to rule out any of this overpowering noise.

The principle of the detection relies on the estimation of the similarity between representations of music. The originally handpicked musical pieces are successively compared to short cut out (sometimes overlapping) excerpts of the recorded game footage. For each comparison, an identification process has to be computed that indicates which part of the musical piece has been recognized.

Estimating the similarity between two musical pieces is a very complex task [27]. This task becomes even more challenging when the musical pieces differ in their presence of different noises. Existing methods generally rely on fingerprinting techniques [28, 29]. These techniques consist of first encoding the original musical piece as multiple fingerprints, which are generally related to spectral properties. Other fingerprints are then extracted from the query which is expected to be identified, and are compared to all the fingerprints encoded and stored in a database. The piece of the database with the highest number of similar fingerprints is then identified as similar thus computing the temporal location of the query.

Instead of this technique we propose here to identify representations based on tonal properties because it is dedicated to musical pieces and seems robust to the presence of noise [25, 30]. In this approach musical pieces are encoded as sequences of symbols corresponding to the distribution of the energy in the amplitude spectra. The comparison of these symbols is computed by aligning the sequences with the use of local alignment algorithms [31]. Local alignment algorithms compute a score similarity. The higher the score, the more similar the segments compared are. Therefore, all the recorded gameplay footage is compared to each musical excerpt. The higher value of the similarity score indicates the presence of the corresponding musical excerpt: the related timestamp indicates the beginning of the searched musical part.

During our experiments, the gameplay footage is decomposed into frames (that are overlapping with a hop size of 1.85 seconds) corresponding to the size of the researched musical excerpts.

2.2 Speech detection system

Gameplay segmentation is also achieved by automatically identifying speech and music segments within the game's audio track.

The audio segmentation is based on the Hidden Markov Models (HMM). Four states HMM are used for this task, each state being modeled by a Gaussian mixture with 256 components. The features used are classical speech processing features, Perceptual Linear Prediction coefficients (PLP). We used 12 coefficients together with their first and second-order derivatives.

Models training

Since we aimed for a wide implementation of the models over a range of videogames, we decided to train the models using radio broadcast data. This allowed for a more generic speech recognition instead of the highly specific use of speech in a game such as *Bioshock 2*. The radio

broadcast data consisted of a French broadcast corpus comprising a total of 50 hours designed for the ESTER evaluation campaign [32]. The available labels on this corpus are: Acappella, Advertising, Applause, Jingle, Laugh, Multiple_speech, Music, Speech, Other.
Acappella is a label used for singing unaccompanied by instrumental music, multiple_speech is used when several people speak at the same time, and other is used for sounds that do not pertain to any other category. The other labels can be considered as self-explanatory.
The system is similar to a phonetic decoder, which means it assumes that only one class may be encountered at a time. For this reason, we use mixed-event classes (i.e., music+speech, jingle+music) to take into account events happening simultaneously. Because the number of classes grows exponentially when mixed events are considered, we decided to preselect those mixed events that were significantly represented in the radio broadcast data (with at least an accumulated duration of more than 100 seconds).
Recognition
During the recognition process, the test file is segmented and labeled using the Viterbi [33] algorithm. Then the labels corresponding to each of the target classes (e.g; speech and music) are collected from the output stream. After that, the target class speech is calculated by merging all the mixed-event classes containing a speech event, (e.g., speech+music, speech+laugh, speech+advertising).
Afterwards, a post-processing scheme is applied to remove segments that are too short (i.e. under half a second of duration).

3. RESULT

To assess the usefulness of these methods for the identification of speech and key audio moments during gameplay sessions, approximately 30 hours of footage from the game *Bioshock 2* was collected for analysis. During a five week long study, ten participants were asked to play *Bioshock 2* during four consecutive sessions of 45-50 minutes, and were asked to comment on selected moments of their footage in week 5. The data set is composed of both the screen and sound capture of the participants' gameplay, participants' psychophysiological responses (heart rate, galvanic skin response), and the participants' facial expressions and keystrokes. The current paper is focusing on presenting the sound analysis of gameplay footage, but the others modalities have been gathered for further correlation works (see for instance the video synchronized presentation system [17]).
The results presented in Section 3 and Section 4 correspond to the results of the sequence identification algorithm and the speech recognition algorithm with the first 45-minute session of nine of the participants (one of the participant first session was not successfully recorded, as the first ten minutes of sound are missing, so the results are discarded in this section).

3.1 The game

Bioshock 2 is a first person survival horror/shooter game that takes place in the underwater city of Rapture. The player assumes the role of Subject Delta, a Big Daddy that is symbiotically connected to Little Sisters and acts as their protector against Rapture's evil citizens. Purpose of the player's movement across Rapture is explained in the form of Subject Delta's separation from his particular Little Sister that he sets out to save from the main antagonist.
The player is regularly given the choice to follow and understand the story in more detail by collecting audio tapes that he can decide to listen to (the tapes are not automatically played upon collection). The player can also choose to ignore these narrative driving opportunities and focus his attention more exclusively on the action affordances of shooting and melee fighting that the game offers.

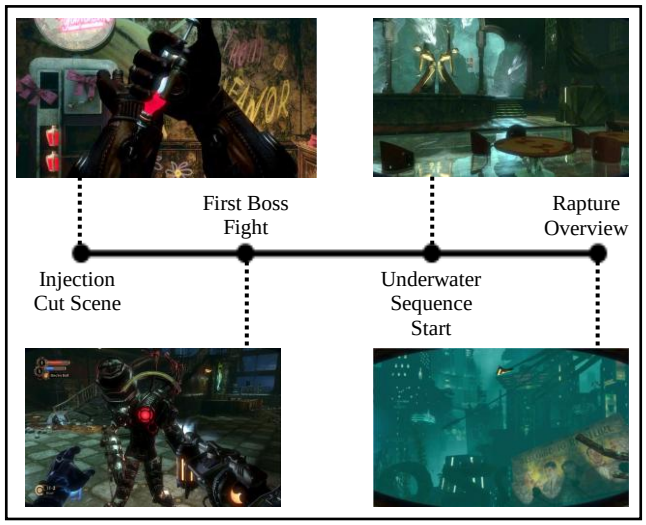

Figure 1. Memorable moments in the first 45 minutes of the game

3.2 Game sequence detection

During the first 45 minutes of the game, the player plays through several memorable moments that deemed important for progressing the story forwards (Figure 1):

- The player acquires his first super human power during a non-interactive cut-scene where he painfully injects a plasmid (a specific power) in his arm.
- The player encounters the first boss fight (Big Sister), introduced by a dark non-interactive cut-scene. The player is unable to finish the battle, as the Big Sister eventually flees from the scene.
- The player is sent underwater after a second confrontation with the Big Sister ends with a window breaking and consequently a flooding of the room.
- The player is presented an overview of the abandoned underwater city of Rapture.

All these sequences are recognizable by their specific sound scheme: the player character screams in agony during the injection; the fight with the Big Sister is preceded by suspense building music; the breaking window makes a loud cracking noise; and the overview of the city is accompanied by majestic music increasing in volume.
As these key moments possess a specific sound scheme, they can be detected by the game sequence algorithm

presented in the previous section. The results are presented in Figure 3, where the absolute detection times can be observed, and in Figure 2, where the detection is displayed relative to the other players.

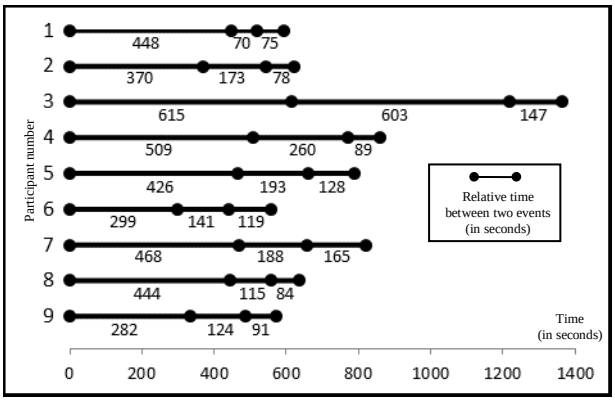

Figure 2. Relative detection times, by participant (1-9)

Detecting these four key moments and their temporal location for all the players, allows a sense of pace in which players move through this game. For instance, by looking at Figure y, several empirically validated comments can be made about the difference in player behavior in *Bioshock 2*:

- It took most of the players around 600s to go through these key moments, but four of them (participants 3, 4, 5 and 7) took more time. It is possible to assume the existence of different ways of engaging with the game system, as participants 3, 4, 5 and 7 seem to perform a greater desire to explore the nooks and crannies of the game world.

- A completely deniable yet visually stunning scene can be attuned to just after being underwater, and before the overview of the city. Looking back inside through a window, the player can see another Big Daddy engaged in combat in the protection of a Little Sister. Because the underwater sequence is extremely linear and no other action but movement is required from the player, we can speculate with reasonable certainty that some players choose to ignore this sequence and go straight to the overview of the city (participants 1, 2, 4, 8 and 9 that completed this sequence between 75s and 90s), while some players will have stopped to watch the fight unfold (participants 3, 5, 6 and 7 that took between 119s and 147s).

- When the Big Sister flees the scene during the first boss fight, most of the participants were observed leisurely exploring the environment for items in the assumption that the fight was over. However, Participant 1 was observed running after the Big Sister which also clearly shows in Figure 2 since this participant arrives at the second confrontation with only 70s after (the others participants needed at least two minutes). It seems that participant 3 did not even try to follow the prescribed path, although very linear in this sequence, since it took him ten minutes to come to the second confrontation.

Figure 3 shows the absolute time detection of these events. The absolute time is also meaningful in understanding the players' behavior. Participant 3, for instance, seems to either have a strong desire to explore the game world in detail or has a tendency to get lost or stuck in several places since it took this participant almost 1400s to encounter the four events. Knowing this can also be useful in automatically segmenting the game sessions, especially if we seek correlation between these moments or sequences and other feedback-based gameplay metrics (e.g. speech detection or video-based one [17]). This information can also be useful in highlighting specific moments of interest for further qualitative analysis (e.g., post-sessions commentaries based on the player's own gameplay behavior).

3.3 Speech detection

The speech detection results are presented on Figure 3 (grey areas). In *Bioshock 2*, moments of speech have an important role in communicating the story of *Bioshock 2* in a more detailed manner. In other words they help the player create a better understanding of what happened to the once so glorious art deco city of Rapture. Speech in *Bioshock 2* consists of propaganda messages broadcasted over the speakers by *Sophia Lamb* (the main antagonist in the game), radio messages sent by (seemingly) friendly characters, and diary tapes containing short spoken messages by previous citizens of Rapture telling their mini-histories during the demise of the city.

Similar to the game sequence identification system, the results can be used to segment a player's gameplay session. Because the player can chose to listen to any tape whenever he wants to, manual identification of moments of speech is a highly challenging and time consuming task. Automatic segmentation of a game session based on speech detection then gives researchers an easy and useful tool to exactly identify when the player is encountering moments of speech. One application of such an automatic segmentation is that it can aid the selection of key-moments for post-game commentary session to assess a player's recall and understanding of the game.

Moreover, automated speech detection provides metrics capable of empirically validating the different ways in which players engage with the game.

- The presented sequences in Figure 3 correspond to the first 45 minutes of the game, which starts with a cut-scene presenting the player character and his relationship with the game's main antagonist. This scene, comprising of an intense monologue by the main antagonist, is highlighted for each participant around 200s. However, participants 4, 8 and 9 miss this detection. These participants clearly skipped the cut-scene, more interested in rushing into the game action than understanding the underlying story.

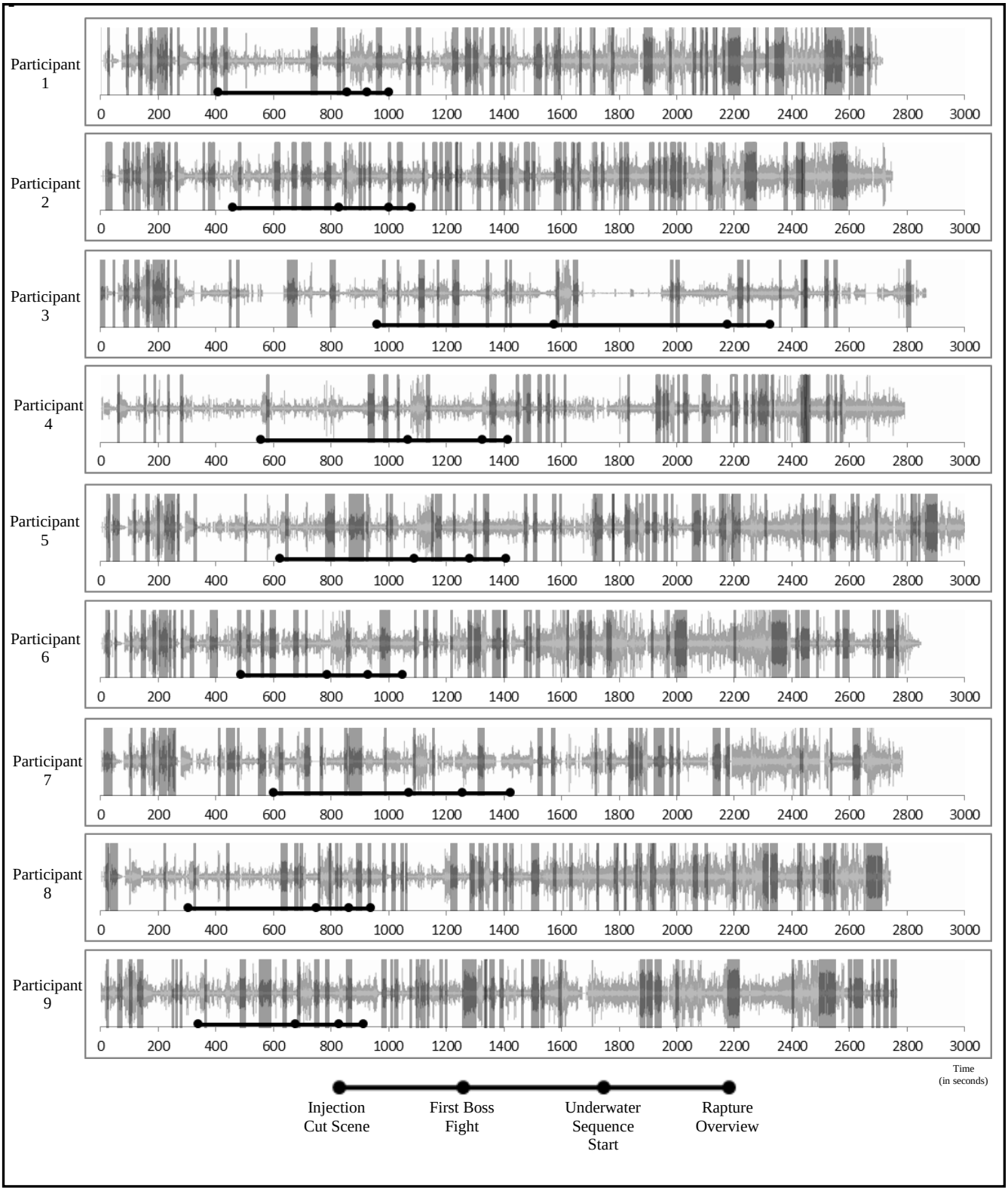

Figure 3. Sound processing results (the speech detection is represented by a grey highlight on waveform and the key sequence detections are represented by the four linked dots)

- As highlighted above, participant 3 took ten minutes to go from the first Big Sister fight to the second encounter. The graph in Figure 3 shows that there is also no speech detection between 1650s and 1950s. Furthermore, the waveform in Figure 3 shows a silent moment. This sparked interest for further analysis which then showed that this particular participant decided to enter the game's story menu to read up on information provided on entities and objects existent within Rapture. This menu does indeed not contain any sound or music. Participant 3 is obviously more interested into the complex story content of the game rather than the action content. This is also highlighted by the time it took him to encounter the four key moments (1400s, see Figure 2). Participant 3 is clearly showing an interest for the game world he is evolving in.
- When counting the speech moments between the injection scene and the overview of the city, we can see that participant 1 only encountered 4 moments while participant 3 encountered 13. Participant 1 also got to the city overview around the same time that participant 3 finished watching the injection cutscene. It thus seems that counting the moments of speech can validate inferences about certain play styles. Participant 1 had a more action driven play style, trying to move quickly through the game not taking much note of diaries scattered around the environment, whereas the behavior of Participant 3 shows a much greater interest in the story aspect of *Bioshock 2*.

4. VALIDITY

4.1 Sequence detection system

The game sequence detection system has been executed on the complete 30 hours of available game footage. The results are presented in Table 1. For each participant, the four sessions have been processed, and the best detections (the detected moments with the highest similarity score) have been selected. It is important to acknowledge that, while executed on the four sessions for detecting the different memorable moments, the algorithm has always returned the best detection score inside the first sessions. This is in line with our expectations (the memorable moments discussed in this paper all occurred at the beginning of the game), and also demonstrates that the algorithm is robust to false-alarm. In Table 1, Detect stands for the best detection (in seconds) and Ref. for the manually performed detection (in seconds). For each participant and each sequence, the system execution time was less than 1 second.

When comparing the best detections and the matching hand coded references, there is no time difference exceeding 3 seconds. Furthermore, for Participant 10 - whose data have been discarded in the previous section because the first ten minutes of sound were not successfully recorded - no detection of the Injection Cut Scene has been highlighted. This is an important result demonstrating the robustness of this method with regard to falsely identified moments.

For the purpose of detecting the temporal position of key moments in a videogame session, this detection can be considered highly accurate.

	Injection Cut Scene		First Boss Fight	
Participant	Detect.	Ref.	Detect.	Ref.
1	408	407	856	858
2	458	456	828	830
3	960	957	1575	1575
4	557	555	1066	1069
5	622	620	1088	1090
6	488	486	787	790
7	601	600	1069	1072
8	304	302	748	750
9	341	339	676	678
10	-	-	682	685

	Underwater Sequence		Rapture Overview	
Participant	Detect.	Ref.	Detect.	Ref.
1	926	927	1001	1003
2	1001	1001	1079	1081
3	2178	2178	2325	2328
4	1326	1325	1415	1416
5	1281	1278	1409	1411
6	928	928	1047	1048
7	1257	1256	1422	1425
8	863	862	939	941
9	828	828	912	914
10	1006	1007	1097	1098

Table 1. Result of the sequence detection system (time in seconds). For each participant, the detection time of the four memorable moments (Detect.) is compared with a reference, hand-coded one (Ref.)

4.2 Speech detection system

The audio segmentation system was tested on each first session of the nine participants presented in the results section. In assessing the validity of this system, hand-coded identification of speech moments in these same files have been compared to the processed results. For each session, the system execution time was around twenty minutes, while the hand-coding of each session took around two hours. Results are presented in Table 2. *Tar* stands for target class time (the total speech time calculated from the hand coding), *non* for non-target class (the total of hand coded non-speech segments), *miss* is the amount of time where speech is hand coded but not detected by the system, and *ins* is the amount of time where speech is detected (inserted) without a speech reference in the hand coded result. With these results, several statistical analyses can be performed.

	tar.	non	miss	ins	%err	%miss	%fa	%rec	%prec	F
Summary	2577.25	21641.95	1387.73	292.11	6.935	53.845	1.350	46.155	80.284	0.586
Average	286.36	2404.66	154.19	32.46	6.956	52.747	1.348	47.253	78.456	0.585

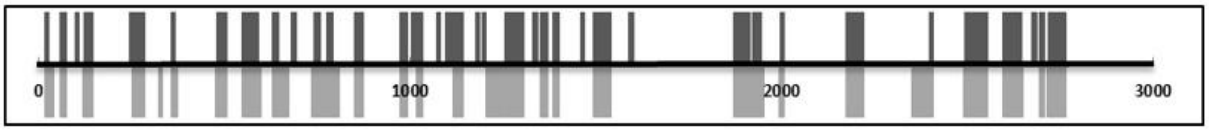

Table 2. Result of the audio segmentation system for the speech class, and example of comparison between hand-coded data (light grey) and speech detection results (dark grey) (participant 9)

Let *total* be the total time of the file

error rate: $\%err = 100 \times \frac{(miss+ins)}{total}$

miss rate: $\%miss = 100 \times \frac{miss}{tar}$

false alarm rate: $\%fa = 100 \times \frac{ins}{non}$

recall: $\%rec = 100 \times \frac{(tar-miss)}{tar}$

precision: $\%prec = 100 \times \frac{(tar-miss)}{(tar-miss+ins)}$

F-measure: $F = 2 \times \frac{(prec \times rec)}{(prec+rec)}$

As displayed in Table 2, the results obtained from the labeled data used to evaluate the performance of the speech detection system on *Bioshock 2* show some flaws of the performing system. An explanation of the significant miss rate (an average of 50%) may be sought in the nature of the speech acts in *Bioshock 2*. Many speech acts are subject to audio effects that deform them making the detection more difficult. Also the existence of extensive background noise (fighting and war-like noises) overpowering the speech acts and the existence of synthesized sounds that may share properties with vocal sounds can be factors attributing to the miss rate.

Nevertheless, the false alarm rate is very low (1.3%) and a good precision rate (80%) is achieved which means that most detected speech parts occur within the hand-coded sequences (this is highlighted by the chart accompanying Table 3). Therefore, these results are still considered useful for the purpose of automatically identifying moments of interest for further scrutiny and provide a new analysis layer capable of segmenting the gameplay session for the analysis of player behavior in games.

5. CONCLUSION

This paper has aimed to show the usefulness of acquiring gameplay metrics through an automated analysis of sound in games. Similar to other gameplay metrics these audio based gameplay metrics allow for an empirical assessment of a player's engagement with the game. This increases our understanding of what it actually means to play games. Empirically analyzing existing sounds in games is especially interesting since sound is deemed such an essential experiential game component. By quantifying elements of the encountered soundscape per player, audio based gameplay metrics can show both the types of sound (speech and music) encountered as well as their temporal location in a gameplay session. An analysis of these metrics can then help our understanding of different play-personas [8] and the different (pace in) behaviour they exhibit.

Audio based gameplay metrics can thus empirically support more theoretical understandings of the gameplay experience by providing robust data on the way that players actually negotiate their way through games. And because audio based gameplay metrics do not have the source code requirements of other type of gameplay metrics these metrics can be implemented in the study of any available game. By extracting metrics from the game's audio feedback [echos], the use of quantitative gameplay data can therefore truly start to move beyond a more exclusive use in playability research into the realm of player experience studies [ethos].

Acknowledgments

The Royal Society of New Zealand, Marsden Grant for funding 'Videogame Classification: Assessing the experience of play' (10-UOW-024).

The University of Waikato, International Scholarship Grant, for funding 'Quantitative Data and Gameplay Segmentation Applied to the assessment of Players' Experience with Digital Games'.

Thank you to Lennart E. Nacke for his valuable help and support.

6. REFERENCES

[1] Tavinor, G. The art of videogames. John Wiley and Sons (2009).

[2] Aarseth, E. Cybertext. Perspectives on Ergodic Literature. London: The John Hopkins University Press, (1997).

[3] Lauteren, G. Pleasure of the Playable Text: Towards an aesthetic theory of computer games, in F. Mäyrä (Ed.) In *Proc. Computer Games and Digital Cultures Conference 2002*. Tampere University Press. (2002)

[4] Consalvo, M. Hot Dates and Fairy-Tale Romances: Studying sexuality in videogames, in J. P. Wolf and B. Perron (Eds.) *The Video Game Theory Reader*, New York: Routledge. (2003).

[5] Vorderer, P., Hartmann, T., and Klimmt, C. Explaining the enjoyment of playing video games: the role of competition. In *Proc. ICEC '03*. (2003) 1-9.

[6] Kivikangas, J. M. Psychophysiology of flow experience: An explorative study. Faculty of Behavioural Sciences, Department of Psychology. Helsinki, Finland, University of Helsinki. (2006).

[7] Nacke, L., and Lindley, C. A. Flow and Immersion in First-Person Shooters: Measuring the player's gameplay experience. In Proceedings of the Conference on Future Play, (2008), 81-88.

[8] Canossa, A. Play-Persona: Modeling Player Behaviour in Computer Games. PhD thesis. Danmarks Designskole, (2009).

[9] Drachen, A., and Canossa, A. Towards Gameplay Analysis via Gameplay Metrics. In *Proc. of the 13th MindTrek*, ACM-SIGCHI Publishers, (2009) 202-209.

[10] Nacke, Lennart E. Affective Ludology: Scientific Measurement of User Experience in Interactive Entertainment. Ph.D. Thesis. Blekinge Institute of Technology, Karlskrona, Sweden.(2009)

[11] Mirza-Babaei, P., and McAllister, G. Biometric Storyboards: visualizing meaningful gameplay events. BBI *Workshop CHI 2011*. ACM. (2011)

[12] Canossa, A., Drachen, A., Sørensen., J.R.M. Arrrgghh!!! – Blending Quantitative and Qualitative Methods to Detect Player Frustration. In *Proc. of FDG 2011,* ACM, (2011)

[13] Drachen, A., Canossa, A.(2009) Analyzing Spatial user Behavior in Computer Games using Geographic Information Systems. In *Proc. of the 13th MindTrek*, ACM, (2009). 182-189.

[14] Kim, J.H., Gunn, D.V., Schuh, E.,Phillips, B.C., Pagulayan, R.J, Wixon, D. (2008) Tracking Real-Time User Experience (TRUE): A comprehensive instrumentation solution for complex systems. In *Proc. CHI 2008*, ACM. (2008)

[15] Nacke, L. E., Drachen, A. and Goebel, S. Methods for Evaluating Gameplay Experience in a Serious Gaming Context. International Journal of Computer Science in Sport, vol. 9 (2) (2010).

[16] Nacke, L., Drachen, A., Korhonen, H., Kuikkaniemi, K., Niesenhaus, J., van den Hoogen, W. Poels, K., IJsselsteijn, W., de Kort, Y. (2009) DiGRA Panel: Playability and Player Experience Research. In *Proc. DiGRA 2009* (2009)

[17] Marczak, R., van Vught, J., Schott, G., Nacke, L. Feedback-based gameplay metrics: measuring player experience via automatic visual analysis. *In Proc. IE 2012*, ACM Press (2012).

[18] Newman, J.The Myth of the Ergodic Videogame, Game Studies, 2(1), www.gamestudies.org, (2002).

[19] Jørgensen, K. On the Functional Aspects of Computer Game Audio. In *Proc. of the Audio Mostly Conference*. Interactive Institute, Piteå, Sweden. (2006)

[20] Moffat, D., and Kiegler, K. Investigating the effects of music on emotions in games, in *Proc. of the Audio Mostly Conference*.

[21] Collins, K. Game Sound. An Introduction to History, Theory, and Practice of Video Game Music and Sound Design. Cambridge, Massachusetts: The MIT Press, (2008).

[22] Grimshaw, M., Schott, G. Situating Gaming as a Sonic Experience: The Acoustic Ecology of First-Person Shooters. In *Proc. DiGRA 2007*. (2007).

[23] Richard, G., Ramona, M., and Essid, S. Combined Supervised and Unsupervised Approaches for Automatic Segmentation of Radiophonic Audio Streams. In *Proc. ICASSP 2007*. IEEE International Conference (2007).

[24] Orio. N. Music retrieval: A tutorial and review, volume 1 of Foundations and Trends in Information Retrieval. Now Pub, (2006).

[25] Bandera, C.D.L., Barbancho, A.M., Tardón, L.J., Sammartino, S., and Barbancho, I. Humming Method for Content-Based Music Information Retrieval. *In Proc. ISMIR* (2011), 49-54.

[26] Yeh, C.H., Kuo, C.H., and Liou, R.W. Movie story intensity representation through audiovisual tempo analysis. Multimedia Tools Appl. 44, 2 (2009).

[27] Ferraro P., Robine M., Allali J., Hanna P., and Rocher T. Detection of Near-Duplicate Musical Documents from a Multi-Level Comparison of Tonal Information in Information Extraction from the Internet (2011), 129-143.

[28] Cano, P., Batle, E., Kalker, T., and Haitsma, J. A review of algorithms for audio fingerprinting. In *Proc. of the Int. Work- shop on Multimedia Signal Processing* (2002), 169-173.

[29] Wang, A. An Industrial-Strength Audio Search Algorithm. *Proc. Int. Symp. On Music Info. Retrieval* (2003), 7-13.

[30] Serrà, J., Gómez, E., Herrera, P., and Serra, X. Chroma Binary Similarity and Local Alignment Applied to Cover Song Identification. *IEEE Trans. on Audio, Speech and Language Processing* (2008), 16(6): 1138-1152.

[31] Smith, T.F., and Waterman, M.S. Identification of common molecular subsequences. J. Mol. Biol., (1981) 147:195-197.

[32] Galliano, S., Geoffrois, E., Gravier, G., Bonastre, J.-F., Mostefa, D., and Choukri, K. Corpus description of the ESTER evaluation campaign for the rich transcription of French broadcast news Language Resources and Evaluation Conference (2006).

[33] Rabiner, L.R. A tutorial on hidden Markov models and selected applications in speech recognition. In *Proc. of the IEEE*, (1989) vol.77, no.2, 257-286