Audio Engineering Society

Convention Paper 5856

Presented at the 115th Convention
2003 October 10–13 New York, New York

COMPUTER GAMES AND MULTICHANNEL AUDIO QUALITY PART 2 – EVALUATION OF TIME-VARIANT AUDIO DEGRADATIONS UNDER DIVIDED AND UNDIVIDED ATTENTION

RAFAEL KASSIER, **SŁAWOMIR K. ZIELIŃSKI** , and **FRANCIS RUMSEY**

Institute of Sound Recording, University of Surrey, Guildford, GU2 7XH, UK

ABSTRACT

The effect of division of attention between the evaluation of multichannel audio quality degradations and involvement in a visual task (playing a computer game) was investigated. Time-variant impairments (drop-outs) were used to provide degradations in audio quality. It was observed that involvement in a visual task may significantly change the results obtained during the evaluation of audio impairments for some experimental conditions.

1. INTRODUCTION

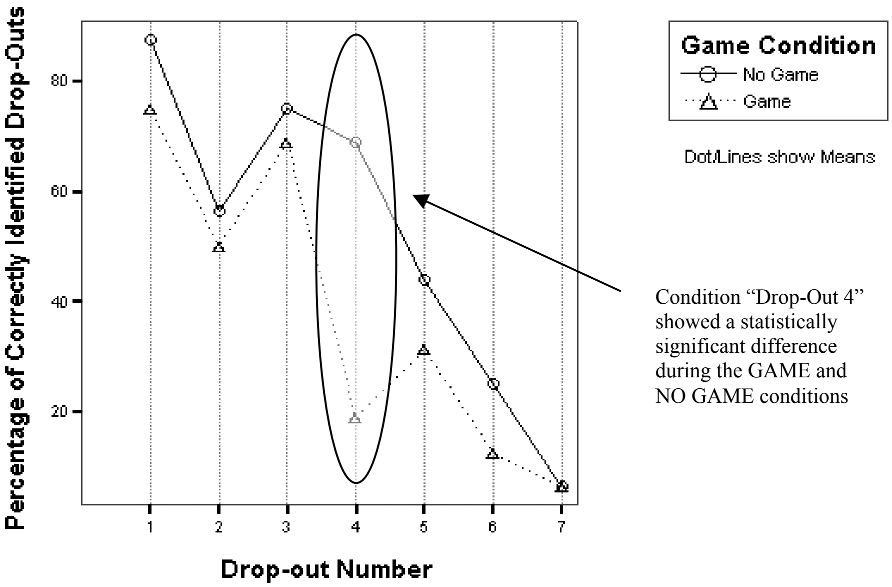
The rapid development of audio-visual systems in telecommunications and the entertainment industry gives rise to the question “in what way should the quality of these systems be evaluated?” According to some studies undertaken in this area, quality of audio and quality of video should not be evaluated in isolation due to the possibility of a cross-modal interaction [1][2][3] which requires complex, time consuming and thus expensive subjective tests. On the contrary, some other studies show that in some cases the effect of audio-visual interaction is very small and therefore can be neglected in the design of subjective tests [4].

The drawback of all previously quoted studies is that the division of attention between visual and auditory tasks was not controlled and therefore experimental conditions can be characterised as passive in relation to watching visual content. Therefore, these conditions were different from a domestic scenario in which listeners are involved in the story line of a movie, atmosphere of a concert, etc. Massaro and Warner undertook an experiment in which they successfully managed to control the division of attention between auditory and visual tasks, however their studies were limited only to the aspect of stimulus recognition – they have not investigated the issue of audio quality perception under selective or divided attention [5].

In a previous paper [6] we showed that involvement in a visual task (playing a computer game) may significantly change the grades obtained during evaluation of audio quality (up to 15 %) for some subjects and for some levels of audio quality. This result is in line with the results of the study undertaken by Massaro et al [5]. It was also found that this effect is subject-specific and the global effect observed after averaging the results across all listeners is very small. The results obtained confirm the existence of significant interactions between auditory and visual modalities. However, the observed interactions are subject-dependent and their magnitude is small after averaging the results across all the subjects. Therefore it could be concluded that the effect of audio-visual interaction is very small and can be neglected in subjective evaluation of the audio quality of audio-visual systems, at least in the case of trained listeners, such as those used in the experiment.

However, the nature of the audio impairments employed in the previous experiment can be characterised as static (stationary) and therefore easily noticed during prolonged exposure. It was expected that time-variant degradations (such as drop-outs) would be much more difficult to notice under the condition of divided attention (when the subject is actively involved in a visual task). This supposition was to some extent confirmed by the results of a pilot study undertaken at the Institute of Sound Recording [7]. According to the obtained results (shown in Figure 1) it is clear that for one experimental condition (Drop-out No. 4), involvement in a visual task made subjects less reliable at detecting drop-outs. For the remaining experimental conditions it is possible to note the trend that subjects are less reliable in detecting drop-outs whilst involved in the game, however the differences were statistically insignificant.

In this paper, we discuss the results of a new experiment using dynamically changing multichannel audio impairments, to verify the hypothesis that it is more difficult to notice dynamically changing audio degradations than stationary degradations whilst involved in a visual task.



Condition “Drop-Out 4” showed a statistically significant difference during the GAME and NO GAME conditions

Figure 1: Pilot Experiment Results (after [7])

2. SELECTION OF A COMPUTER GAME

The main criteria for selection of the computer game in [6] were:

- Constant involvement of a subject in the game (it was a requirement that the degree of the involvement should be the same throughout the whole period for which a subject plays the game in order to avoid any short-term situations where the subject pays no attention to the game and is entirely concentrated on audio evaluation)
- Short period of required training
- Gender independency
- Consistent audio characteristics

It was found that a majority of the popular computer games did not meet these criteria (e.g. action and sport games). For most of the examined games the level of the involvement in a game was highly variable and depended on the current game conditions. Moreover, a game's audio content and its characteristics were also variable and depended on the game events and conditions. Consequently, it was difficult to use a typical action or a sport game in the experiment in which repeatability and consistency of conditions were of high importance. Moreover, the state-of-the-art games require a long period of training, which may prolong the experiment and thus make it more expensive. Additionally, some games are particularly violent which may appeal only to a limited group of subjects.

Taking into account all these considerations, it was decided to use a mind/skill type game providing a relatively constant involvement in the task and requiring relatively little training. The game chosen was a variant of the popular "Tetris" style game for Windows PC [8] (See Figure 2).

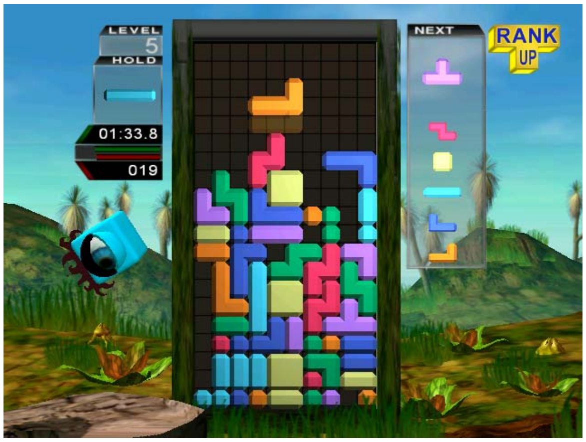


Figure 2: Tetris in action [8]

The goal of Tetris is to manipulate a series of falling blocks by rotating and moving them, such that they form a tight wall of blocks with no gaps. Once a complete horizontal line of blocks or parts of blocks has been assembled it disappears, allowing additional lines to build up in their place. The game itself is simple in operation, but demands concentration and can be increasingly stressful at times, as the pieces build up and the rate at which pieces fall increases.

Useful features of this game are:

- It is widely known – therefore little training is required (no subjects needed the basic rules explained to them)
- The subjects were able to play at their own pace, so it was possible for the players to involve themselves to a similar level, regardless of relative skill levels
- The particular version used had the ability to display a summary screen that included useful information about the subject's activity during the task

The drawback of this particular game was that the accompanying audio material was recorded in a two-channel stereo format and therefore was not suitable for the purposes of the experiment, since it was intended to use a game with surround audio. Therefore it was decided to mute the native background music in the game's software mixer and to use a high quality surround 5.1 recording instead, played back by a separate computer (SGI) equipped with software for running subjective tests (see Appendix B). After the informal pilot tests, it was also decided to mute the game's sound effects since, in the authors' opinion, they were annoying in the long term and also caused occasional beating effects when mixed with the external recording. The game's voice messages, which were kept intact in the previous experiment [6], were also muted during this experiment, as they might coincide with one of the intended time-variant degradations.

Subjects played the game within approximately 2-minute trials. In the previous experiment [6] the subjects were instructed to attain the highest game score possible. This was measured in terms of difficulty levels attained during the task as a whole - a rather vague measure at best. For the new experiment it was decided that more data should be collected of the subject's involvement in the game.

Table 1: Scoring for Tetris

Tetris "task"	Description	Score
Single	Clearing a single line of blocks	1
Double	Clearing two lines of blocks simultaneously	3
Triple	Clearing three lines of blocks simultaneously	6
Tetris	Clearing four lines of blocks simultaneously	10
B2B Tetris	Achieving "Back to Back" Tetris clearances	5
T Spin	Rotating a "T" block into a tight position.	10

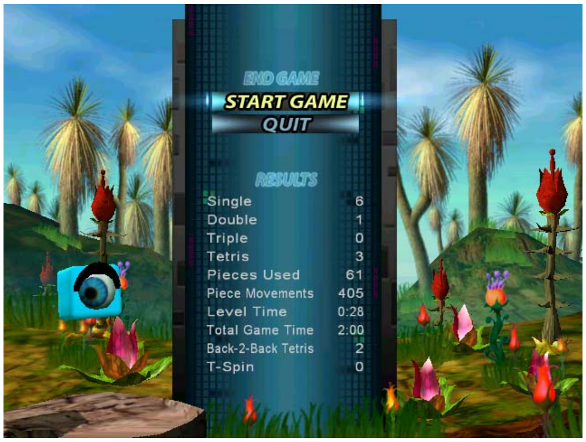
Because the game itself had no internal scoring system other than the concept of difficulty levels attained by the subject, it was decided to introduce a different scoring system for the subjects to work against. This would allow for a more intuitive scoring system to rate one subject's performance against another within the trials' short duration, and to help to involve the subjects more in the task of playing the game.

The scoring system is shown in Table 1.

The rationale for the scoring system was that a "Double" should be worth more than two "Singles", a "Triple" worth more than three "Singles", a "Tetris" worth more than two "Doubles" or a "Triple" plus a "Single" etc. The points given for "B2B Tetris" and "T-Spin" were bonus points awarded for rare or more complicated tasks.

At the end of each 2-minute game session, it was possible to view a summary page which contained information about how many of the above events happened during the session. In addition to the scoring events, "pieces used" and "piece movements" are also shown. These are totals of how many falling pieces were slotted into place during the two minutes and how many times the falling blocks were rotated or moved left and right and down. The importance of these last two totals is that they give a measure of the subject's activity during the task that should be more accurate than the game score (it is possible to move the same number of pieces the same amount of times and generate completely different game scores!). See Figure 3, for an example of the game summary screen.

Summary screens for each 2-minute game item in the tests were captured whilst the subject was evaluating the audio. This data was later entered manually into the statistical analysis package.



Single	6
Double	1
Triple	0
Tetris	3
Pieces Used	61
Piece Movements	405
Level Time	0:28
Total Game Time	2:00
Back-2-Back Tetris	2
T-Spin	0

Figure 3: Game Activity Summary Screen [8]

3. SELECTION OF AUDIO MATERIAL

The same source material was used in this experiment as in the previous experiment [6]: an instrumental jazz music recording (without vocals). The instruments (acoustic guitar, piano, bass guitar, synthesizers, drums and percussion) were mixed across all 3/2 channels. The duration of the excerpt was 2 minutes and 10 seconds. The music itself was of a similar nature to other computer game background music tracks.

4. PROCESSING OF AUDIO MATERIAL

In the previous experiment [6] the degradations in audio were obtained using a static low pass filter at specific cut-off frequencies. For this experiment, time-varying degradations were used in order to see if dynamically changing degradations were perceived differently under divided and undivided attention.

The pilot study [7] had shown that the introduction of drop-outs is a satisfactory way of adding time-variant degradations to multichannel audio. However the exact nature of these drop-outs (length, depth, frequency and channel) needed to be decided upon.

Table 2: Main degradations chosen

Degradation	Drop Out Nature		
	Frequency	Length	Channels
1	Once	1 Second	L+C+R
2	Three Times	1 Second	L or R (randomised)
3	Three Times	1 Second	L+LS or R+RS (randomised)
4	Three Times	1 Second	C+LS+RS

During pilot tests it was clear that an impairment scale as described in [9] would be more appropriate than the "Basic Audio Quality" scale used in the previous experiment [6]. This was because it was found to be difficult to average the effect of drop-outs across the entire excerpt o give an overall "Basic Audio Quality" grade to the items, but it was easier to find a descriptive term from the impairment scale to describe adequately the effect of the degradations in terms of their perceptibility and level of annoyance caused.

To finalise the degradations to be used in the experiment, a single subject (who was not used subsequently in testing), was exposed to a range of drop-out impairments with varying lengths, frequencies, as well as number and placement of channels, and asked to evaluate them using a multiple stimulus test. Because of time constraints and because it was considered unlikely that all subjects would rate the stimuli in a similar way, final degradation patterns were chosen that had been graded at around the mid-point of the impairment scale. This was in order to allow for subjects to be more critical, or less critical than the pilot subject.

Because of experimental time constraints, there was space for 6 items (with 6 iterations each), which would include a reference and a nominal "anchor", which would be the most severely degraded item.

The anchor was created during one of the pilot tests, and featured eight 1-second drop-outs every 10 seconds (apart from the first and last 30 seconds of the item), one in each of the following channels: L, R, LS, LS+RS, L+C+R, L+LS, R+RS (where: L = left front channel, R = right front channel, C = centre channel, LS = left surround channel and RS = right surround channel). The position of the drop-outs was randomised in each 10 second slot. There was one version of the anchor, iterated six times in each of the main experimental conditions.

There was therefore space for 4 additional items. The chosen degradations are shown in Table 2.

In order to minimise the chance that specific degradation types would be detected by the subjects, and to average the influence of particularly noticeable or indeed undetectable drop-outs, six different patterns of drop outs were created for each of the four main degradation patterns.

It was decided that audio degradations would be introduced into the period of time during which the subjects' involvement in the game was the highest. It was hypothesised that subjects would need some time (about 30 sec.) in order to get fully involved in the game. It is also possible that a subject's attention may "drift" to the audio evaluation task towards the end of the game. Therefore the original recording was processed in such a way that drop-outs occurred only during the period in which the highest involvement in the game was predicted to occur. The original recording was 2 min. 10 sec. long, which left a 1 min. 10 sec. period for the impairments to take place.

For the main degradations (1-4), either one or three drop-outs were randomly placed into the appropriate channels. In the case of the degradations containing three drop-outs, these were placed one each into three approximately equal sections of the 1min 10 second period, with randomised placement within each of the three sections.

Drop-outs were created using the same custom envelope within an audio editing software program, containing a short fade out (approx 0.1 sec in length) to digital silence and fade back (approx 0.1 sec in length) to unity. The total envelope length was one second.

This created 6 multichannel audio files for each of the four degradations, plus anchor and reference files.

Details of the different impairment patterns for the main degradation items can be found in Appendix A.

5. EQUIPMENT

Five loudspeakers were arranged according to the ITU-R BS. 775 Recommendation [10] (see Figure 4).

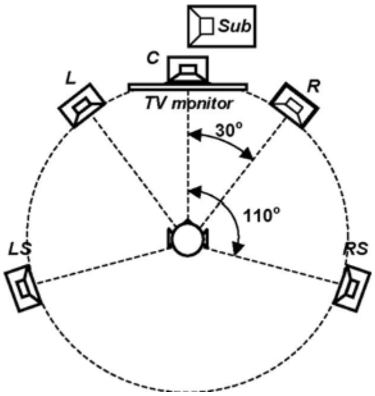


Figure 4: Arrangement of the audio-visual equipment.

The distance between the loudspeakers and the optimum listening position was equal to 2.1 m. The subwoofer was located behind the centre loudspeaker about 20 cm from the wall and 35 cm from the centre loudspeaker. A TV monitor (42" plasma display, 16:9 aspect ratio) was used for visual presentation of the game. The distance between the TV monitor and the listener was set to 4 H, where H is the height of the viewing area (this distance conformed to [11]). The technical specifications of the loudspeakers used in the experiment and other details related to the equipment are presented in [4]. The subject was seated at the optimum listening position.

The audio stimuli were played-back with the use of the "Alex" software running on an SGI computer. The audio items were stored using 6-channel uncompressed 16-bit audio files. The audio signals were transmitted digitally from the SGI computer to a digital mixing desk (Yamaha O2R) and converted using 20-bit D/A converters operating at the 48 kHz sampling rate.

There were two additional computers installed in the control room: a laptop running presentation software (to display messages at the beginning and end of each trial, and to provide the static picture during non-game playing sessions), and a standard PC which was used to run the game. A diagram of this equipment set-up can be found in Appendix B.

Running the experiments required the experimenter to perform a series of timed manual operations. Each experimental session was timed and controlled via the laptop's presentation software, which normally displayed its video output on the plasma screen in listening room. The presentation software would display greetings before and after each session, countdown timers before each of the items, prompts for the subjects to evaluate each item, and during play-back in the non-game sessions, would provide a static picture output. At the appropriate times, audio playback was cued manually on the SGI. During game sessions, the plasma screen was switched to display the output of the game PC. 10 seconds into playback, the experimenter would start a new game on the local keyboard of the game PC, then immediately use the keyboard switch to activate the game controller keyboard in the listening room, simultaneously disabling his own keyboard. At the end of playback, the video and keyboard switches were used to present a "Time is over" message from the laptop and to disable the subject's keyboard. During non-game sessions the laptop would run a continuous presentation, supplying a static picture during playback, although audio playback was started manually on the SGI. An additional "control" video monitor was installed in the control room to mirror the plasma screen.

6. ACOUSTICAL CONDITIONS

The listening tests were conducted in the Listening Room of the Institute of Sound Recording, University of Surrey. The acoustical parameters of this room conform to the requirements of the ITU-R Recommendation BS. 1116 [9]. All channels (L, R, C, LS, RS) were aligned relative to each other with a tolerance within ± 0.3 dB SPL (measured at the reference listening position). Absolute level alignment conformed to the ITU-R BS.1116 Recommendation [9]. All measurements were performed using a 1/2" pressure microphone (Bruel & Kjaer, Type 4134) at the centre listening position (measurements were carried out only at one listening position). The microphone was positioned at a height of 1.2 m pointing upwards. The average level of the audio stimuli was 73 dBA.

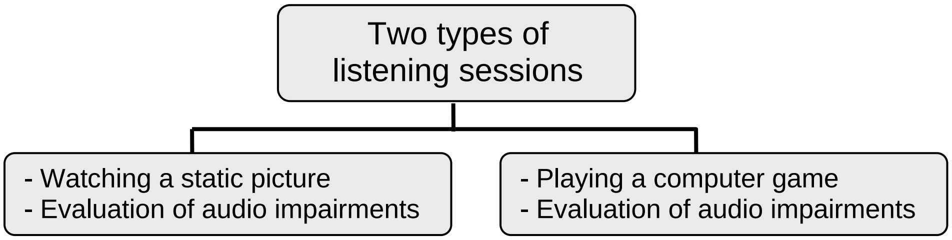

Two types of listening sessions
- Watching a static picture
- Evaluation of audio impairments
- Playing a computer game
- Evaluation of audio impairments

Figure 5: Two types of listening sessions.

7. EXPERIMENTAL DESIGN

There were two types of listening session in the experiment (Figure 5), corresponding to the two main experimental conditions.

The first type of session involved simultaneously playing the computer game and evaluating the audio quality — during these sessions subjects' attention was divided between the evaluation of the audio impairments and playing the game. In the second type of session listeners were asked to evaluate the audio impairments and to watch a static picture of a typical screenshot from the game — during these types of session the listeners' attention could be focussed mainly on the evaluation of the audio impairments (no involvement in the game). Originally in the previous experiment [6] it was planned to use a moving picture containing a demonstration of the game, however, during informal tests at that time, it was found that this drew too much attention towards the visual task, making this condition similar to that of an active involvement in the game.

Seven experienced listeners took part in the experiment. Each listener was given a one-hour familiarisation period. During this, they were exposed to the Reference audio followed by the most degraded item (the so-called "anchor"). Both items were then repeated. During both iterations of the "anchor", the subjects were asked to assign a grade to the item from the grading scale and to record this on a grading form (see Appendix C). The main reason for this was to allow the subjects to hear the nature and degree of degradations that would be presented in the main test, and to begin thinking about how to assign grades to the items on the impairment scale. The subjects were then given the remainder of the hour (about 45 minutes each) to practice playing the game. No game scores were recorded during the familiarisation phase.

Because involvement in playing the game might decrease each subject's consistency in the grading of audio quality it was decided to repeat each experimental condition six times.

The main tests consisted of 6 half-hour sessions, each containing 12 items (2 iterations each of the Reference and Anchor, and 2 different patterns of the 4 other degradations). The main reason for including the Reference and Anchor was to make the listeners more consistent in using the full range of the scale by exposing them both to the original and severely impaired recordings in each session (a form of listener calibration). There were 3 sessions with game, and 3 sessions without game.

Both types of session used a single stimulus paradigm (one stimulus was evaluated at a time). There were 12 items consecutively evaluated within each session (Figure 6). Each item was approximately 2 min. and 10 sec in length. A short pause after each item was scheduled for evaluation purposes. Both the order of sessions and the order of presentation of stimuli were randomised to minimise the carry-over effect. During "game" sessions, subjects were instructed that the accurate evaluation of audio quality and achieving the highest possible game score were of the same importance.

In addition to the trial items, each subject performed two "benchmark" game items, one at the beginning of the first game session, and one after the last game session. During these "benchmark" items, the subjects were exposed to the reference audio track, informed of this, and instructed *not* to grade the audio, but to concentrate on achieving the highest possible game score which would count with the other game scores to determine the ultimate "winner" of the game. For purposes of the analysis, this provided a useful performance benchmark, against which the activity of the subjects during the game condition trials could be measured. This data would allow verification that subjects' attention to the visual task was not "dipping" to allow them to concentrate on audio evaluation.

Subjects were asked to assign grades to each item using the form shown in Appendix C.

The form is divided into 20 "minor ticks", with a "major tick" every 5 "minor ticks". When entering the data from the grading sheets, this 20 point scale was converted to a 100 point quality scale, shown in Table 3.

8. DATA ANALYSIS

The obtained results were analysed using the following ANOVA model:
Rating = GM + GAME + DEGRAD + SUB + All interactions + residuals

where:

GM	-	General mean
GAME	-	Main experimental variable having two levels (game / no game),
DEGRAD	-	Degradation type (either one of the 4 main degradations, the anchor or reference)
SUB	-	Subject number (listener number), 1-7

All factors used in the ANOVA model were fixed.

Residuals were attributed to inconsistencies in grades between the different "time frames" of degradations 1-4, and between the grading of repetitions of the hidden reference and the hidden anchor.

Figure 6: Structure of test sessions

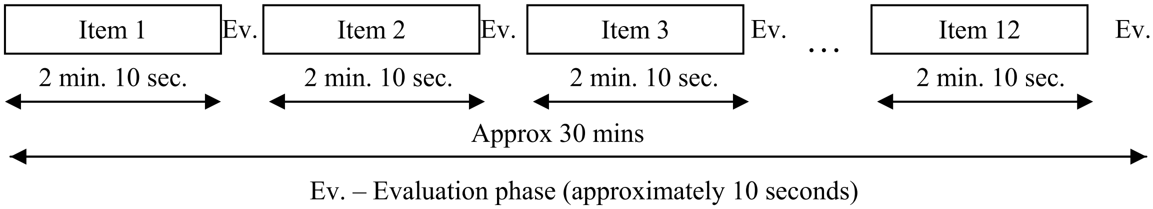


Table 3: Grading scale used in the experiment

Grade	Description	Grades after conversion
5	Imperceptible	100
4	Perceptible, but not annoying	75
3	Slightly annoying	50
2	Annoying	25
1	Very annoying	0

Table 4: Tests of Between-Subjects Effects

Dependent Variable: Quality

Source	Type III Sum of Squares	df	Mean Square	F	*p* (Significance)	Partial Eta Squared (Magnitude of effect)
Corrected Model	347159.468 (a)	83	4182.644	17.184	.000	.773
Intercept	1474636.198	1	1474636.198	6058.469	.000	.935
GAME	1213.341	1	1213.341	4.985	.026	.012
DEGRAD	252628.492	5	50525.698	207.582	.000	.712
SUB	50320.663	6	8386.777	34.457	.000	.330
GAME * DEGRAD	7488.492	5	1497.698	6.153	.000	.068
GAME * SUB	5929.909	6	988.318	4.060	.001	.055
DEGRAD * SUB	23512.480	30	783.749	3.220	.000	.187
GAME * DEGRAD * SUB	6066.091	30	202.203	.831	.725	.056
Error	102228.333	420	243.401			
Total	1924024.000	504				
Corrected Total	449387.802	503				

(a) R Squared = .773 (Adjusted R Squared = .728)

According to the ANOVA test (table 4), all investigated factors and second-order interactions were significant at $p<0.05$ level.

This means that, unlike in the previous experiment [6], there was a global effect of the 'GAME' factor on the results of the evaluation of audio. As in the previous experiment [6] the 'GAME' factor was significant in interactions with other experimental factors. This means that playing a game affected the results of audio quality evaluation of the time-variant audio degradations used in this experiment for some experimental conditions (for some degradations, and for some subjects), and that playing the game had an overall effect when averaged across all subjects and degradations.

The magnitude of each effect is shown by the partial eta squared value in table 4. This shows that DEGRAD had the largest effect ($\eta^2 = 0.712$), whereas GAME had the least effect ($\eta^2 = 0.012$), although it was statistically significant.

The third order interaction between degradation nature, the main experimental condition (game / no game) and subjects 'DEGRAD * GAME * SUB' was not significant at $p<0.05$ level. This means that each subject's grading of the different degradations with and without game need not be studied individually.

8.1 Testing of ANOVA Assumptions

ANOVA makes three assumptions about the data, which need to be checked.

8.1.1 ANOVA Assumption 1: Independence of grading

Dependence was minimised through randomisation of experimental factors.

8.1.2 ANOVA Assumption 2: Normal distribution of scores for each group

Checking normal distribution of scores for each group is equivalent to checking for normal distribution of residuals. This was initially checked graphically (see Figure 7).

Graphical examination shows that the distribution is close to normal distribution, but to analyse how this distribution deviates from normal, a formal Kolmogorov-Smirnov test was conducted (see table 5).

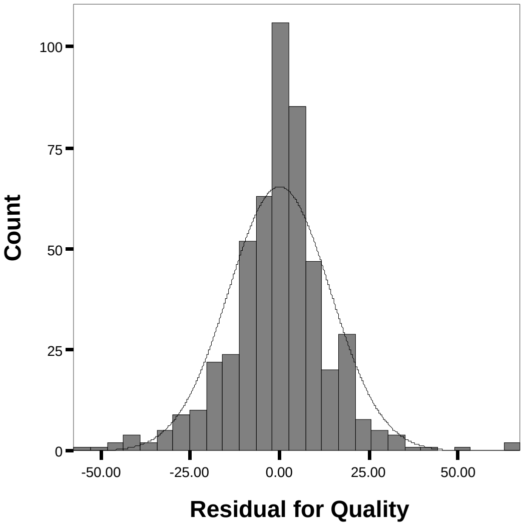


Figure 7: Histogram of Residuals for Quality (with normal distribution curve superimposed)

Table 5: One-Sample Kolmogorov-Smirnov Test

		Residual for Quality
N		504
Normal Parameters(a,b)	Mean	.0000
	Std. Deviation	14.25613
Most Extreme Differences	Absolute	.099
	Positive	.093
	Negative	-.099
Kolmogorov-Smirnov Z		2.230
Asymp. Sig. (2-tailed)		.000

(a) Test distribution is Normal.
(b) Calculated from data.

The Kolmogorov-Smirnov test shows a significant departure from normality, but it is known that the ANOVA test is robust to the violation of the normality assumption provided that the sample size is large (minimum 15 cases per group [12]). The minimum number of analysed cases per group was 6 (for GAME * DEGRAD * SUBJECT), but was 12 or higher for the other groups, so the ANOVA assumption was roughly fulfilled. Checking for this assumption does however expose the weakest point in the analysis, and larger numbers of cases should be included in future experiments.

8.1.3 ANOVA Assumption 3: Homogeneity of variance between cases.

The third assumption for the ANOVA test is that there is homogeneity of variance between groups. According to a formal Levene's test, the homogeneity of variance between groups was not equal. However the ANOVA test is known to give reliable results even when the variances are not equal across different groups, provided that the number of cases in each group is the same [13], which was the case in the current experiment.

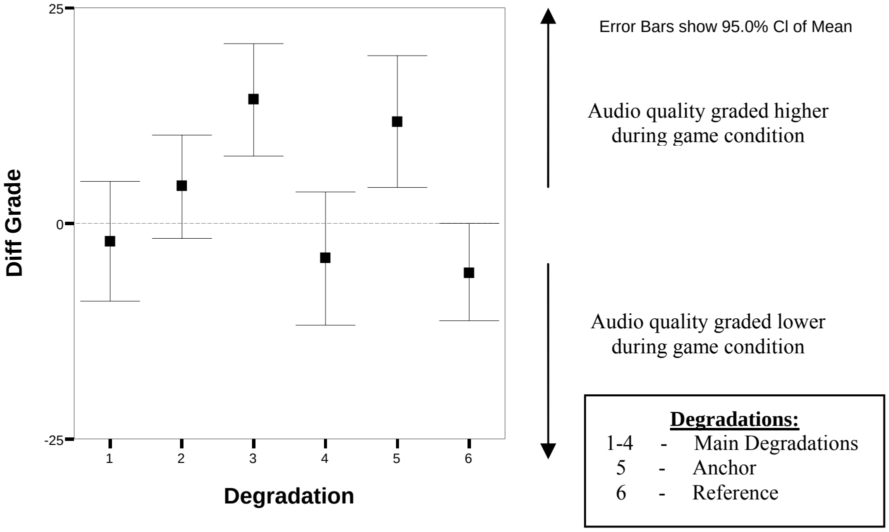


Figure 8: Audio quality diff grades for different degradations

9. RESULTS

9.1 Global GAME effect

In contrast to the previous experiment [6], the GAME factor in this experiment had a significant effect on the grading of audio quality. The magnitude of this effect was small (about +3%), but was significant at p<0.05 level and thus showed that involvement in the game produced an upward shift in audio quality grading.

This means that even though the degradations were relatively easy to detect, playing the game significantly increased audio quality grades. However, due to significant interactions with other factors, the effect of the GAME condition is seen to vary between degradations and between subjects.

9.2 GAME Interactions

9.2.1 Interaction between Game and Degradation (GAME * DEGRAD)

In order to study interactions between the game condition and different audio degradations, “diff grades” were calculated by subtracting the NO GAME condition quality grades from those from the GAME condition, for each item. These diff-grades were then plotted by degradation, shown in Figure 8.

Figure 8 shows that audio quality for degradations 3 and 5 were being rated significantly higher during the GAME condition due to involvement of listeners in the game. Other degradations show insignificant “diff grades”, except degradation 6 (the “Reference”), which was being graded significantly lower during the GAME condition. Errors in the detection of the reference result in changes to quality grading in a negative direction only (due to the ceiling effect), and one of the possible interpretations as to why there is a statistically significant decrease in audio quality for the reference during the GAME condition is that more errors were made detecting the hidden reference while playing the game than while not playing the game.

The magnitudes of the upward shifts in diff-grades of degradations 3 and 5 are approximately +14% and +12% respectively (the overall effect of the game on diff grades was just +3%), and the effect of the GAME * DEGRAD interaction itself is relatively small ($\eta^2 = 0.068$).

It is difficult to explain entirely the tendencies observed in Fig. 8. For example, it is not clear why degradations 3 and 5 being graded especially high, and how do they differ from the other degradations?

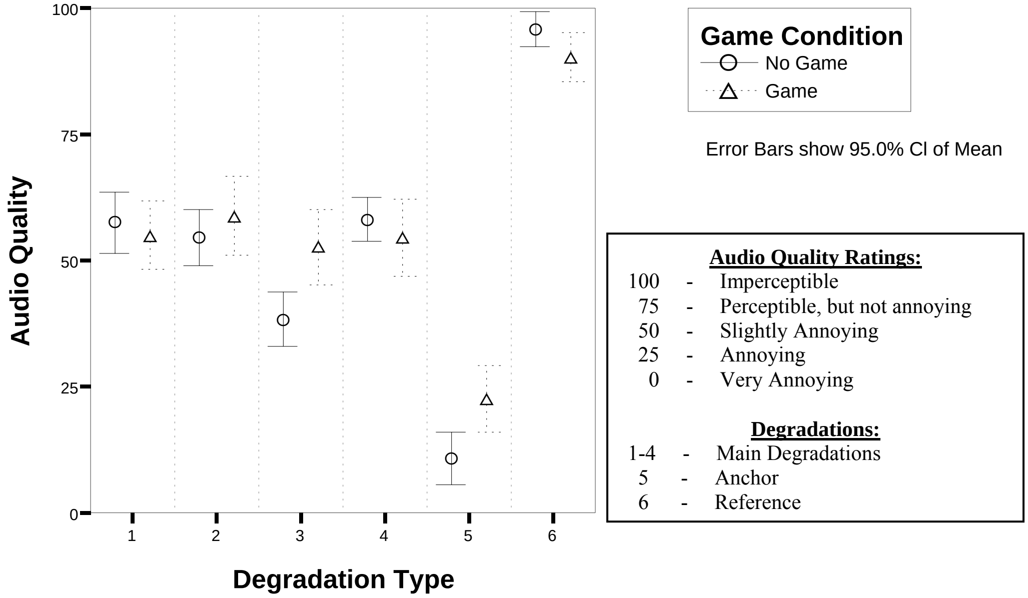


Figure 9: Audio quality for GAME and NO GAME conditions for different Degradations

To investigate this in more detail, a graph of the absolute values of the GAME and NO GAME audio quality grades was plotted by degradation (see Figure 9). This representation of the data shows that the difference between degradations 3 and 5 was that they were, on average, graded with lower audio quality than the others during the NO GAME condition.

According to Figure 9, involvement in the game is seen to improve the quality of the items that were graded as being severely impaired by the subjects during the non-game condition.

It is difficult to interpret why this happened, especially since the results of the previous experiment [6] showed that involvement in the game affected the audio quality of only slightly impaired items. Further research would be needed to investigate a wider range of degraded items in order to work out where the onsets of these effects lie in static and in time-varying degradations.

9.2.2 Interaction between Game and Subject (GAME * SUB)

To investigate the interaction between GAME and SUBJECT, diff-grades were again used to create a graph showing the difference in audio quality ratings between the NO GAME and GAME conditions for all degradations averaged, displayed by subject (Figure 10).

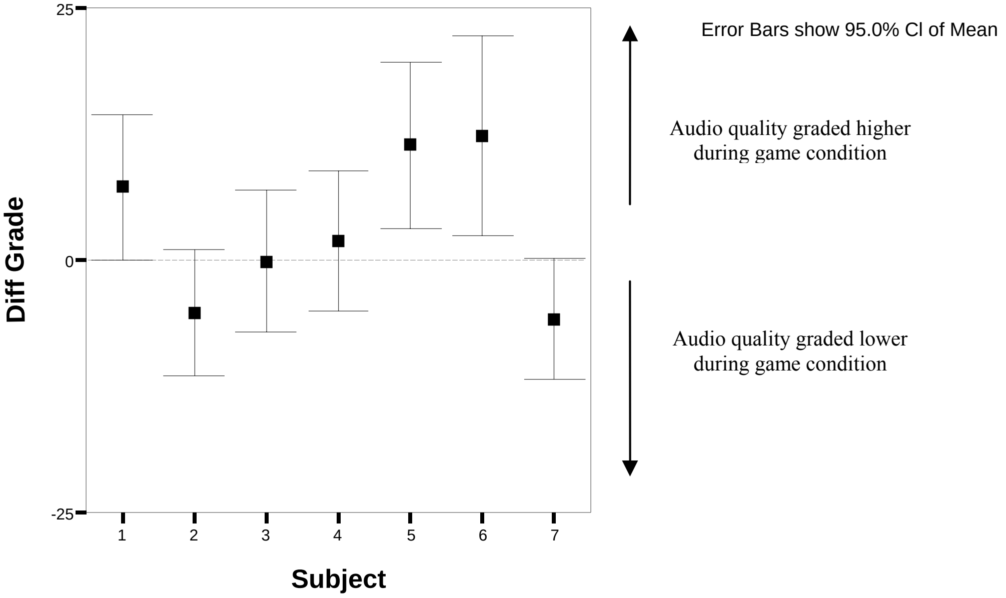


Figure 10: Audio quality diff grades for different subjects

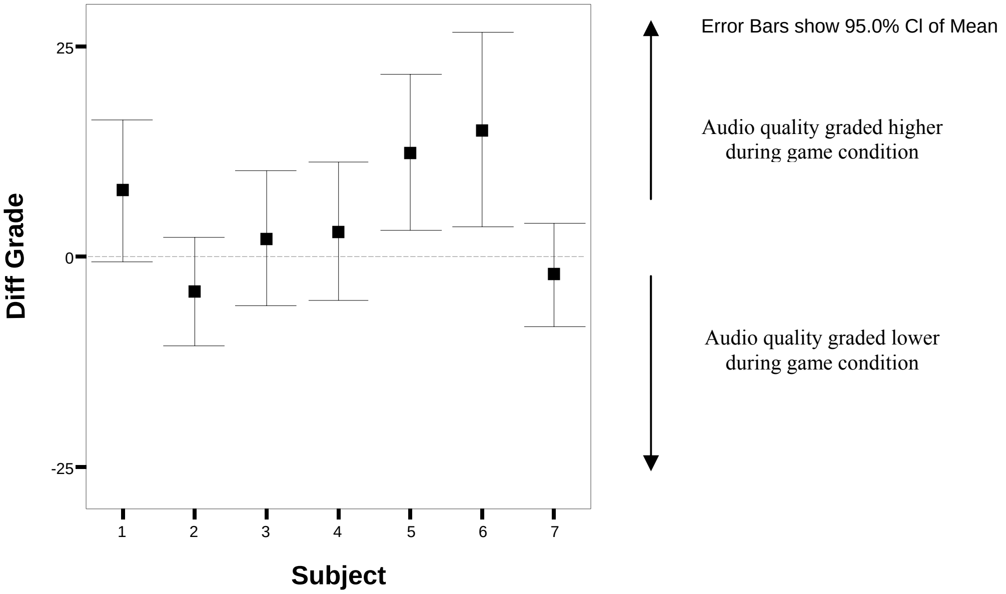


Figure 11: Audio quality diff grades for different subjects (excluding grades for the Reference item)

Figure 10 shows that subjects 5 and 6 graded audio quality higher during the game condition than when they graded the same item during the non-game condition. Subject 7, on the other hand seemed to grade audio quality lower while involved in the game. For the remaining subjects no statistically significant interaction was found. To test whether this downward effect in subject 7 was due to the "ceiling effect" of the reference, the graph was plotted again, this time excluding grades for the reference (Figure 11).

Figure 11 provides graphical evidence that the significant downward shift in diff grades for subject 7 was due to grading errors and the ceiling effect on the reference item dominating the results for that subject.

Investigating the GAME * SUBJECT interaction has shown that the effect of playing the game on the evaluation of audio quality is subject-specific, and may cause significant increases in audio quality grades for some subjects.

9.3 Grading Error Considerations

Looking at the confidence interval sizes in Figure 9 (which shows absolute values of audio quality for the GAME and NO GAME condition, plotted by degradation) one may note that the grading error appears slightly larger when subjects were involved in the game. Study of the absolute values of the residuals from the ANOVA model can provide statistical verification for this claim, as absolute values of residuals reflect grading error.

Since only the "anchor" and "reference" items (DEGRAD 5 and 6) were repeated, it was decided to examine the grading error in those two conditions. The square root of the absolute values of their residuals from the ANOVA model were calculated and plotted in Figure 12. The reason for calculating the square root was to allow the grading error to be expressed in the same units as the grading scale.

As the confidence intervals do not overlap, Figure 12 shows statistical evidence that there is a significant increase in grading error during the GAME condition. This is not surprising because subjects would be expected to be less consistent when sharing their attention between two tasks.

9.4 Monitoring Game Performance

One of the concerns felt when designing the experiment was that it was important that the subjects' attention on the game did not dip due to their concentration on the evaluation of audio quality.

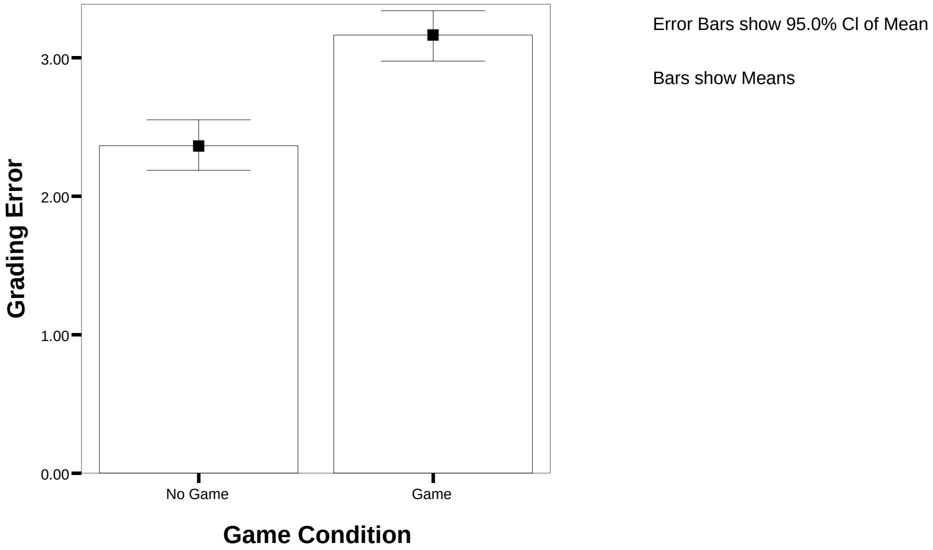


Figure 12: Grading Error for DEGRAD 5 and 6, for the GAME and NO GAME conditions

Figures 13, 14 and 15 show exemplary plots of the game performance scores against "game trial" which indicates the order in which the game items occurred. In other words, they show the selected subject's game performance over time. The "benchmark" items are coded as game trials 1 and 38.

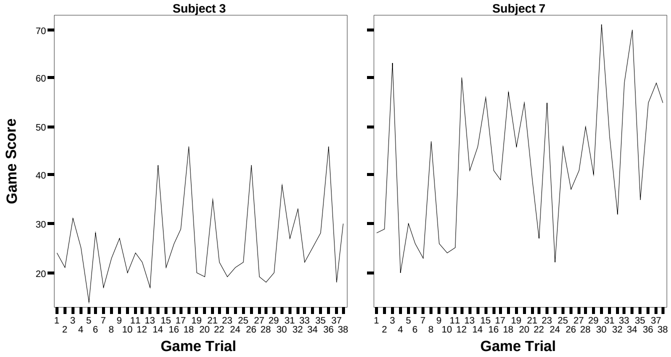


Figure 13: Exemplary plot of Game Scores over time

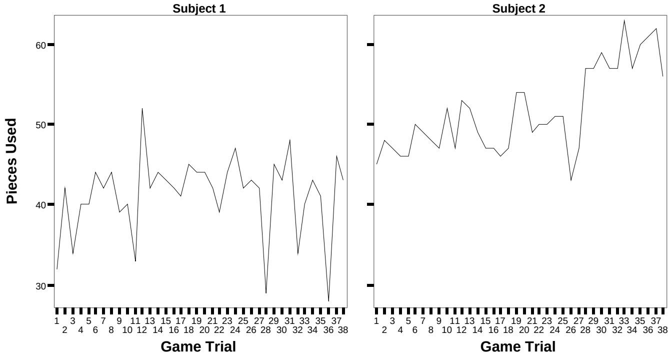


Figure 14: Exemplary plot of Pieces Used over time

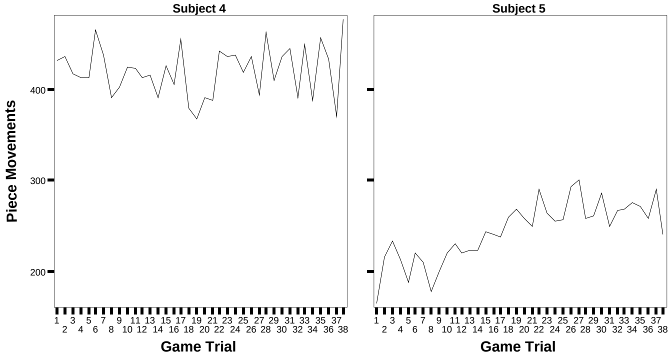


Figure 15: Exemplary plot of Piece Movements over time

The main purpose of plotting Figures 13, 14 and 15 is to show that they are not “U-shaped”, ie: where both “benchmarks” show significantly higher activity levels to the main game experiments. In the exemplary plots shown, one can see that activity during the main experiment for subject 3 in Figure 13, subject 1 in Figure 14 and subject 4 in Figure 15 show peaks of higher activity than the benchmarks, and troughs of lower activity than the benchmarks, with the activity staying more or less constant throughout. This is in contrast to the plots for subject 7 in Figure 13, subject 2 in Figure 14 and subject 5 in Figure 15 that show an overall increase between the first and last benchmarks, evidence of an improvement in game score and activity as subjects acquired a greater degree of skill with the game.

It is also worth noting that game score varies widely, even between consecutive items (see Figure 13). One can see, for example, that subject 7 achieves two very high game scores towards the end of the experimental run, with relatively low game scores between these two “peaks”. In contrast to this, graphs of pieces used or piece movements show generally smoother lines, less prone to the large changes found in the game score graphs. This is explained by the simple fact that two very different game scores can be achieved by moving the same number of pieces the same number of times, with the subject involved in the task to approximately same degree.

It can be concluded that the game successfully controlled the subject’s attention, because subjects showed a similar or increased level of activity between the controlled “benchmark” items where no audio evaluation was undertaken, and the experimental game trials.

9.5 Learning Effects

9.5.1 Audio Quality Grading “Learning” Issues

It was decided to check for bias due to carry-over, learning effect, or the effect of boredom or fatigue (due to the requirement to listen to 78 iterations of the same source material in various degraded forms).

In order to check this, a graph of averaged audio quality grades was plotted by session (each session contained 2 iterations of each degradation type, game and no game conditions were randomised for subjects, so can be assumed to be evenly distributed across sessions).

As Figure 16 shows, the 95% confidence intervals for the grades of each session overlap, therefore there were no significant changes in audio quality grades over time due to carry-over, fatigue, boredom or learning effects.

Because there were no significant time-dependant changes in audio quality grades, the previously presented ANOVA table (table 4) does not require the inclusion of any time-dependant variables (such as session or individual trial number) as co-variants.

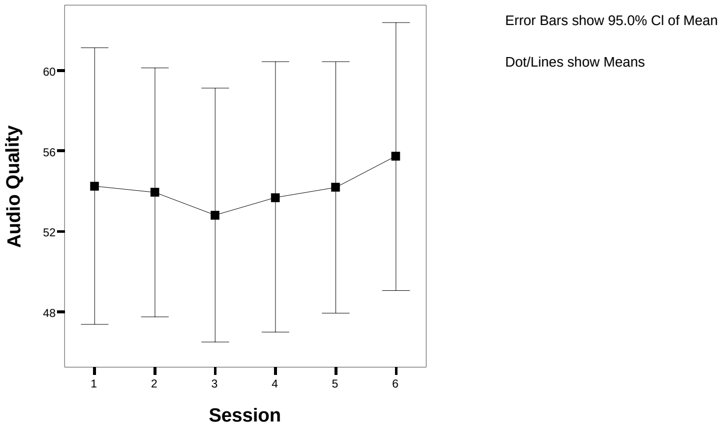


Figure 16: Audio Quality Grades by Session

9.5.2 Game Performance “Learning” Issues

It clear that there are no significant changes in audio quality due to ongoing experimental time, however, there could be time-dependant changes in game performance which could be examined. To this end, the “game trial” variable, which sequentially numbered individual game trials was used to plot how game performance variables changed over time.

Figure 17 shows an overall increase in game scores across subjects over time. According to regression analysis, the increase in game scores over time is statistically significant, and can be described using the following equation: Game score = 21.22 + (0.41 * item order). This means that subjects increased their game score by 0.41 points per trial on average.

As previously mentioned, game score is not a very accurate measure of subjects' activity, therefore we should analyse the other performance data (pieces used and piece movements).

Figure 18 shows a graph of mean pieces used against time, and Figure 19 shows mean piece movements against time. Both appear to show evidence of learning, which is confirmed by regression analysis:

Pieces used = 43.20 + (0.23 * game trial)
ie: subjects use 0.23 more pieces with each trial.

Piece movements = 267.96 + (0.93 * game trial)
ie: subjects perform 0.93 more piece movements per trial.

In summary, game performance was seen to increase on average over time (through the duration of the experiment). This is further evidence to support the claim that the subjects had been successfully involved in the visual task, as they tended to improve upon previous activity rates and game scores.

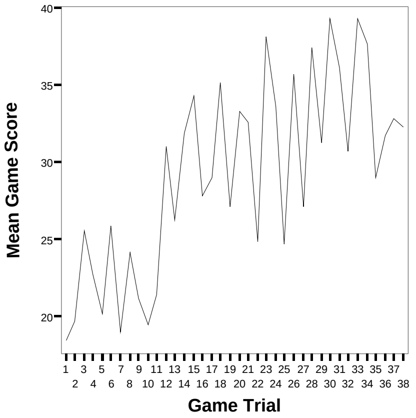


Figure 17: Mean game score against time

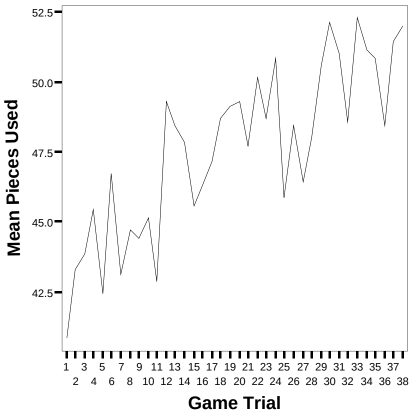


Figure 18: Mean pieces used against time

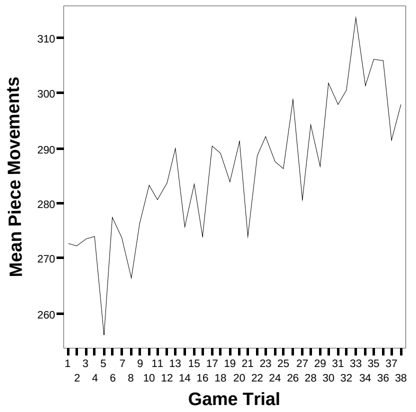


Figure 19: Mean piece movements against time

9.6 Correlation between audio quality and game performance

A bivariate correlation analysis was performed to check if there was any correlation between the game performance variables and audio quality diff grades for each item, the results of which are shown in table 6.

According to table 6, there is a small negative correlation between diff grades and number of pieces used. This correlation is clearly illustrated in Figure 20.

Table 6: Correlations (for all subjects)

		Game Score	Piece Movements	Pieces Used
Diff Grade	Pearson Correlation	-0.104	-0.041	-.152(*)
	Sig. (2-tailed)	0.099	0.518	0.016
	N	252	252	252
Pieces Used	Pearson Correlation	.649(**)	.621(**)	
	Sig. (2-tailed)	0	0	
	N	252	252	
Piece Movements	Pearson Correlation	.332(**)		
	Sig. (2-tailed)	0		
	N	252		

* Correlation is significant at the 0.05 level (2-tailed).
** Correlation is significant at the 0.01 level (2-tailed).

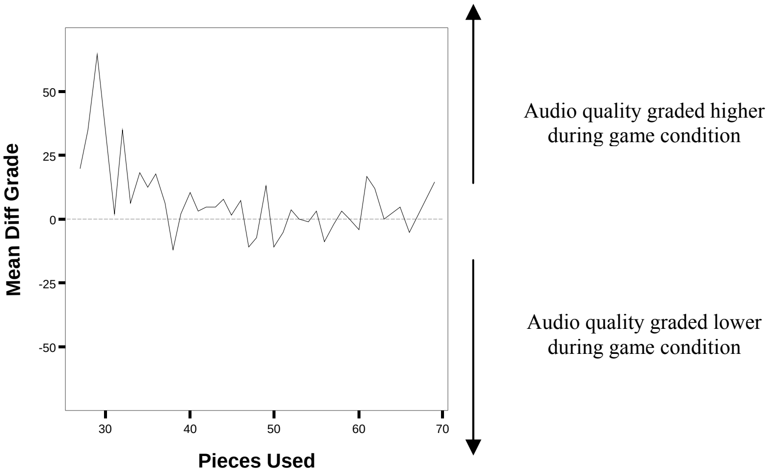


Figure 20: Mean Diff Grades against Pieces Used

Table 7: Tests of Between-Subjects Effects

Dependent Variable: Diff Grade

Source	Type III Sum of Squares	df	Mean Square	F	*p* (Significance)	Partial Eta Squared (Magnitude of effect)
Corrected Model	18625.128 (a)	6	3104.188	6.759	.000	.142
Intercept	4528.114	1	4528.114	9.859	.002	.039
PIECES USED	3593.962	1	3593.962	7.825	.006	.031
DEGRAD	15590.327	5	3118.065	6.789	.000	.122
Error	112522.189	245	459.274			
Total	133574.000	252				
Corrected Total	131147.317	251				

(a) R Squared = .142 (Adjusted R Squared = .121)

Figure 20 shows that during game trials where small numbers of pieces were used, there was an increased tolerance to audio impairments compared to game trials where greater numbers of pieces were used. This is a surprising result because one would expect that an increase in game activity (pieces used) would result in increased uncertainty of audio quality rating, or increased tolerance to audio impairments (positive increase in diff grade).

It is difficult to interpret what could be causing this negative correlation. One possible explanation is that for particularly bad trials (where fewer pieces were used), the subjects were more distracted, possibly turning additional attention towards correcting game-playing errors, and were therefore more tolerant towards any audio impairments.

Table 6 also shows relatively high correlation between game score and the two game activity variables "pieces used" and "piece movements". This appears to be logical, as the more pieces that are used in the game, the greater the total number of movements of the pieces would be for the trial, and the greater the game score is likely to be. The correlation was not higher, however, because attaining a certain game score can be achieved using varying numbers of pieces and moving those pieces a varying number of times, depending on playing skill and random factors such as the computer's selection of the order of the playing pieces.

Due to the significant correlation between diff grades and pieces used, it was decided to calculate an ANOVA of diff grade using degradation number as a factor and "pieces used" as a covariate.

Table 7 shows that both factors are significant, but the magnitude of the PIECES USED factor is small (partial eta squared = 0.031)

In this case all ANOVA assumptions were fulfilled, and the ANOVA confirms that number of pieces used affected diff grades in a very small, but statistically significant way.

10. DISCUSSION

The current experiment has shown that certain time-variant degradations are tolerated to a greater extent by subjects under conditions of divided attention. This also indicates that the results of the previous experiment [6] cannot be generalised.

Results of the current experiment show significant changes to the audio quality of more severely degraded items due to active involvement in the visual task. It is difficult to interpret why this happened, especially since the results of the previous experiment [6] showed that involvement in the game affected the perceived audio quality of only slightly impaired items. Further research would be needed to investigate a wider range of degraded items in order to work out where the onset of these effects lies for both static and in time-varying degradations.

Partial violation of the second ANOVA assumption indicates the need to test larger groups of subjects in future experiments, in order to take full advantage of parametric statistical analysis.

11. CONCLUSIONS

The effect of changes in evaluation of multichannel audio quality under conditions of divided and undivided attention was investigated. A computer game was successfully used as a means of dividing subjects' attention. Time-variant degradations (drop-outs) were used to provide audio quality impairments.

Involvement in the computer game had a significant but very small overall effect (+3%) in the grading of audio quality. This effect was also seen to be subject-specific and most obvious for the more severely degraded items. Also, as hypothesised, active involvement in a visual task decreased the consistency of audio quality grading.

In comparison with the results obtained in the previous experiments, the maximum magnitude of the effect associated with the involvement in a visual task is similar (about +15%) for both static and the time-variant degradations.

However, these conclusions cannot be generalised at this stage of research, since the audio degradations did not span the quality scale for each subject, and the results were obtained using a small group of highly trained listeners. Further research in this area is needed.

12. ACKNOWLEDGEMENTS

This project was carried out with the financial support of the Engineering and Physical Sciences Research Council, UK.

The authors are grateful to Russell Mason (Institute of Sound Recording) for his help in the selection of the computer game, and Ben Supper for proofreading the paper.

REFERENCES

[1] J.G. Beerends, F.E. De Caluwe, "The Influence of Video Quality on Perceived Audio Quality and Vice Versa," J. Audio Eng. Soc., vol. 47, pp. 355-362 (1999 May).

[2] A.N. Rimell, M.P. Volcker: "The Influence of Cross-Modal Interaction on Audio-Visual Speech Quality Perception" presented at the AES 105th convention, San Francisco, California. Preprint No. 4791, 1998 Sept. 26-29.

[3] R.L. Storms, "Audio-Visual Cross-Modal Perceptual Phenomena", Ph.D. Dissertation, Naval Postgraduate School, Monterey, California (1998).

[4] S.K. Zielinski, F. Rumsey, S. Bech, "Effects of bandwidth limitation on audio quality in consumer multichannel audio-visual delivery systems," J. Audio Eng. Soc., vol. 51 (6), pp. 475–501, (2003).

[5] D.W. Massaro, D.S. Warner "Dividing Attention Between Auditory and Visual Perception", Perception & Psychophysics, Vol. 21 (6), pp. 569-574, (1977).

[6] S. K. Zielinski, F. Rumsey, S. Bech, B. de Bruyn, R. Kassier, "Computer Games And Multichannel Audio Quality – The Effect Of Division Of Attention Between Auditory And Visual Modalities", presented at the AES 24th International Conference on Multichannel Audio, Banff 2003 26-28 June.

[7] A. Lister, "An Investigation into the Effect of a Visual Task on Auditory Task Performance", Final Year Technical Project. Institute of Sound Recording, University of Surrey (2003 - Not published).

[8] Tetris Worlds PC-CD ROM, (THQ Inc., 2001).

[9] ITU-R Recommendation BS. 1116, "Methods for Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems," International Telecommunications Union (1994).

[10] ITU-R Recommendation BS. 775-1, "Multi-channel Stereophonic Sound System With or Without Accompanying Picture" International Telecommunications Union (1992-1994).

[11] EBU Recommendation Tech 2376-E, "Listening Conditions for the Assessment of Sound Programme Material. Supplement 1: Multichannel Sound," European Broadcasting Union, Geneva (1999).

[12] S.B. Green, N.J. Salkind, and T.M Akey, Using SPSS for Windows (Prentice-Hall, Englewood Cliffs, NJ, 2000)

[13] D.C. Howell, Statistical Methods for Psychology (Duxbury, New York, 1997)

APPENDIX A – IMPAIRMENT PATTERNS FOR MAIN DEGRADATIONS

Degradation 1 (One, 1 sec. drop out in L+C+R)

Pattern	Start Point (in seconds from beginning of item)
A	75
B	40
C	32
D	43
E	50
F	89

Degradation 2 (Three, 1 sec. drop outs in L or R)

Pattern	Start Point (in seconds from beginning of item) and L/R
A	30 R, 68 R, 77 L
B	34 R, 72 L, 90 L
C	42 R, 71 L, 98 R
D	52 L, 75 L, 81 R
E	31 R, 63 R, 97 L
F	42 L, 58 R, 79 R

Degradation 3 (Three, 1 sec. drop outs in L+LS or R+RS)

Pattern	Start Point (in seconds from beginning of item) and L/R
A	47 R, 54 R, 91 L
B	46 L, 56 R, 97 R
C	32 R, 57 L, 93 R
D	30 R, 74 R, 83 L
E	40 L, 67 R, 82 L
F	45 L, 74 L, 99 R

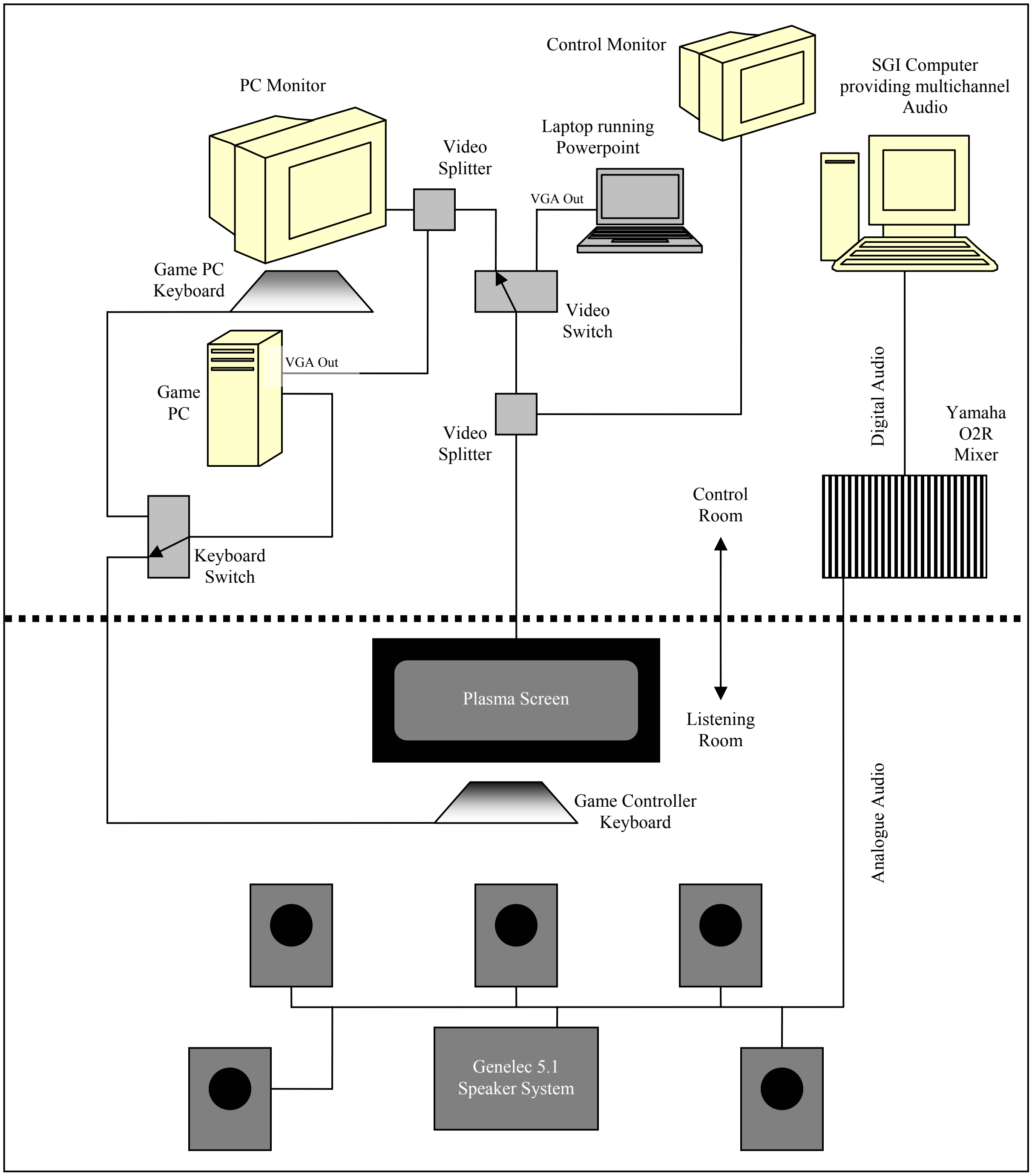
Degradation 4 (Three, 1 sec. drop outs in C+LS+RS)

Pattern	Start Point (in seconds from beginning of item)
A	48, 56, 78
B	42, 76, 82
C	44, 67, 85
D	40, 57, 91
E	31, 61, 95
F	51, 60, 93

Where

- **L = Left Channel**
- **R = Right Channel**
- **C = Centre Channel**
- **LS = Left Surround Channel**
- **RS = Right Surround Channel**

APPENDIX B: Experimental set-up



APPENDIX C: GRADING FORM

Rate the annoyance of the impairments
Please put an “X” in the appropriate box

(Grade non-impaired items as “5”)

□-	**5**	**Imperceptible**
□		
□		
□		
□		
□-	**4**	**Perceptible, but not annoying**
□		
□		
□		
□		
□-	**3**	**Slightly annoying**
□		
□		
□		
□		
□-	**2**	**Annoying**
□		
□		
□		
□		
□-	**1**	**Very annoying**