Audio Engineering Society

Convention Paper

Presented at the 132nd Convention
2012 April 26–29 Budapest, Hungary

Determining the Threshold of Acceptability for an Interfering Audio Programme

Jon Francombe[1], Russell Mason[1], Martin Dewhirst[1], and Søren Bech[2]

[1]*Institute of Sound Recording, University of Surrey, Guildford, Surrey, GU2 7XH, UK*

[2]*Bang & Olufsen, Peter Bangs Vej 15, 7600, Struer, Denmark*

Correspondence should be addressed to Jon Francombe (j.francombe@surrey.ac.uk)

ABSTRACT

An experiment was performed in order to establish the threshold of acceptability for an interfering audio programme on a target audio programme, varying the following physical parameters: target programme, interferer programme, interferer location, interferer spectrum, and road noise level. Factors were varied in three levels in a Box-Behnken fractional factorial design. The experiment was performed in three scenarios: information gathering, entertainment, and reading/working. Nine listeners performed a method of adjustment task to determine the threshold values. Produced thresholds were similar in the information and entertainment scenarios, however there were significant differences between subjects, and factor levels also had a significant effect: interferer programme was the most important factor across the three scenarios, whilst interferer location was the least important.

1. INTRODUCTION

In today's media-driven society, there are ever more ways for users to access audio, with a plethora of products producing sound in the home, car or almost any other environment. Potential audio programmes include a large variety of music, speech, sound effects and combinations of the three. It is also increasingly common for products producing audio to be portable. This wide range of increasingly portable products which produce audio coupled with the ubiquity of audio in almost all facets of society naturally leads to an increase in situations in which there is some degree of audio-on-audio interference.

Examples of such situations might include audio produced by a laptop computer in a room with a television; a mobile phone conversation whilst a car radio is on, or in the presence of piped music in a shopping centre; or competing workstations in an office environment. It is therefore of interest in a number of areas, within the audio industry and beyond, to evaluate the perceived effect of audio interference upon a target audio programme.

Similar research areas include those in which the experience of listening to target audio (or performing a task) is altered by the presence of some interfering audio programme or noise. Such research includes the effects of environmental noise (e.g. road traffic or aeroplane [1, 2, 3]) and noise in the workplace (office [4], school [5], hospital [6]). These areas tend to focus solely on wide-band noise interference, and there has been little research into the effects of audio-on-audio interference; one area where similar work has been performed is in the perceptual evaluation of blind source separation algorithms [7].

In this paper, the design and results of a preliminary investigation into the effects of audio-on-audio interference are presented; the aim of the research described is to establish the threshold of acceptability for an interfering audio programme and to examine how this is affected by various physical parameters.

2. EXPERIMENT SCOPE

In order to determine the best way to investigate audio-on-audio interference, an informal listening test was performed. The design of this test and the conclusions drawn are presented in this section, followed by the aims of the subsequent formal tests.

2.1. Informal Listening Test

In order to qualitatively investigate the effects of audio-on-audio interference situations, a demonstration was set up in the ITU-R BS.1116 [8] standard listening room at the University of Surrey, Guildford. The target programme was reproduced over a single on-axis loudspeaker, and an array of interferer loudspeakers positioned equal distances apart at points on a circle with the listening position at the centre could be controlled to play the unwanted programme from any combination of the loudspeakers. The output of the interfering loudspeakers was controlled so that as more loudspeakers were added, the physical level of the interferer programme at the listening position remained constant. A road noise signal could also be added to the interferer loudspeakers in order to simulate the effect of audio-on-audio interference in an automotive environment.

A bespoke user interface facilitated control of the programme, level and low- and high-frequency cut-off for both the target and interferer signals. A wide range of audio, speech and multimedia programme items were available to use as the target or interferer signals; the programme items were perceptually loudness matched by the experimenter over headphones.

The following observations were made by a small group of experienced listeners. It was interesting to note that it seemed intuitive to adjust the level of the interferer programme to some point where the listener was happy with the situation, or the interferer was ‘no longer annoying’; this point can be considered the ‘threshold of acceptability’. This task seemed much more natural than trying to quantify the extent of the annoyance experienced. It also appeared that different listeners had vastly different thresholds for this ‘no-longer-annoying’ point, although discussion with participants suggested that listeners were performing different tasks; for example, some were trying to relax and enjoy the target audio, whilst others were simply setting the level to the point where they could still understand the target content. This suggested that the task being performed has a pronounced effect on the acceptability threshold, and that the task should be clearly defined to subjects in any future experiment.

Alongside these observations, it was found that all of the variable parameters (target programme, interferer programme, spectrum and location) had an effect on the experience of listening to the target audio in the presence of the interfering audio.

2.2. Experiment Scope and Aims

In order to begin to quantify the effects found in the informal listening tests, a formal experiment was designed. The methodology and results of this experiment are discussed in sections 3 and 4 respectively.

As audio-on-audio interference is a relatively novel research area, there is little in the way of research looking into the acceptable level of interfering audio. Therefore, the first stage of this research is to conduct a broad study, looking at a range of scenarios, programme material and other parameters that may affect the situation. The aims of the experiment described in this paper are as follows:

- to determine the target-to-interferer ratio required for an audio-on-audio interference situation to be acceptable;
- to determine the effect of the task being performed by a listener on this acceptable level;
- to quantify the magnitude of the effects of various physical parameters; and
- to investigate individual differences between participants.

The tasks and physical parameters selected for investigation were motivated in part by observations from the informal listening tests.

3. METHODOLOGY

In this section, the design of an experiment to realise the aims outlined in section 2.2 is described.

3.1. Method of Adjustment

A modified method of adjustment task was used to produce threshold values. In the method of adjustment, subjects are asked to alter the level of a stimulus until it matches a provided reference [9] or reaches some absolute threshold (commonly the threshold of audibility [10]). However, in this adapted procedure, subjects were asked to adjust the level of the interfering programme until it reached the threshold of acceptability. There are examples of the method of adjustment being used to determine an abstract threshold (such as acceptability) in the literature [11, 12]. The response attribute ‘acceptability’ was selected based on comments from the informal listening tests reported in section 2.1 and its common use in the literature [13, 14, 15].

The method of adjustment has a number of advantages [16, 17]: subjects are able to focus on a specific attribute even with stimuli in which multiple attributes are variable; the subject is in control of the stimulus presentation and therefore stimuli are presented in the most suitable manner for each subjects; the method is efficient; and, intra-subject reproducibility is higher than in forced choice procedures. Whilst the method of adjustment may be slightly less accurate than alternative methods, the gain in simplicity is a worthwhile trade-off in this preliminary experiment [18, 10].

In order to avoid bias due to subjects learning to associate certain movements or positions of the fader with the appropriate threshold value [19], an unlabelled rotary fader with no end-points (the *Griffin PowerMate* [20]) was used to control the level of the interferer. The fader adjusted the volume in steps of 0.3dB.

A pilot experiment, using the method described above, was performed to determine an appropriate starting level for the interfering programme. Subjects reported that the task was confusing when the interferer started at a level below the threshold of audibility and that they would initially increase the level to a point well above threshold before continuing with the task. This observation was supported by plots of the individual fader use (figure 1) alongside an analysis of variance (ANOVA), which suggested that there was no significant difference between the threshold produced when the interfering programme started at a level above or below the threshold of acceptability (F(1,31)=1.197, p=0.282).

Therefore in the full experiment, the interferer started at a level well above the threshold of

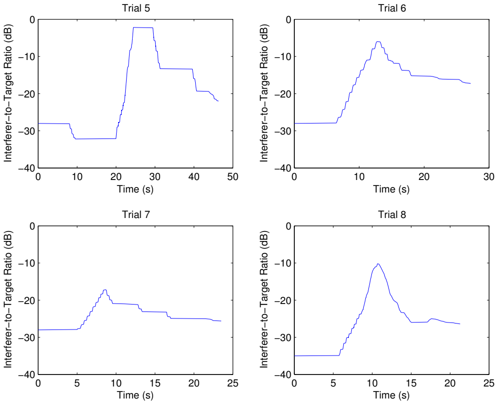


Fig. 1: Selection of Example Fader Use Plots from Pilot Experiment

acceptability (randomised between −3dB and +6dB with reference to the target programme).

3.2. Experiment Scenarios

It was apparent from the informal listening tests that the task being performed influenced the effect of interference upon the listening experience. In order to further investigate this, three ‘scenarios’ were included in the experiment. These scenarios were designed to reflect realistic tasks that people may carry out in the presence of an interfering audio programme, and were defined as follows:

- Information Gathering: *“Imagine that you are at home or in the car, listening as if you were required to understand, act on and/or pass on the information presented”.*
- Entertainment: *“Imagine that you are relaxing (at home or in the car) by listening to music or a football match”.*
- Reading/Working: *“Please read the provided newspaper article. Imagine you are reading or working at home, the office or in the car”.*

The newspaper articles used in the reading/working scenario were print-outs of short current affairs articles from the online archives of the London Evening Standard [21] (mean length: 168 words) accompanied by a related picture. The texts were selected to have minimum variation in Flesch Reading Ease score [22] (mean score: 50.43). Articles were unique for each repeat of each trial and presented to each subject in a different random order.

3.3. Subjects

Nine subjects (six experienced listeners and three inexperienced listeners) participated in the listening tests. The experienced listeners were students and staff from the Institute of Sound Recording, University of Surrey. The inexperienced listeners were undergraduate and postgraduate students in various disciplines.

3.4. Factors

Five factors were selected for investigation: target programme, interferer programme, interferer location, interferer spectrum and road noise. These factors were selected as they all had an apparent effect on the listening experience in the informal listening tests described in section 2.1. For the purposes of experiment design it is beneficial to have the same number of levels of all factors, therefore the factors were each varied in three levels.

As potential target programme material is determined by the task, the target programme material was different in each scenario. In the reading/working task, the target is silence and therefore this variable was omitted. The levels for the information gathering and entertainment scenarios are detailed below (see appendix A for details of the programme material items used).

Target Programme (Information): the following programme items were selected to provide a selection of audio items containing information that may be appropriate in the situation where a listener was listening in order to understand the information presented.

- Male News Speech
- Sports Commentary
- Female News Speech

Target Programme (Entertainment): the programme items used in the entertainment scenario were selected to give a wide coverage of the type of audio that listeners may listen to in an entertainment or relaxation scenario.

- Vocal Pop Music
- Sports Commentary
- Instrumental Classical Music

The remaining factors relate to the interfering programme and were therefore the same in all scenarios, as interference could potentially come from any source regardless of the target task. The levels used are detailed below.

Interferer Programme:

- Male Speech
- Instrumental Classical Music
- Vocal Pop Music

Interferer location:

- 0-degree interferer
- 90-degree interferer
- Diffuse interferer (5 loudspeakers, stimulus processed using Pulkki's method [23] of convolution with white noise bursts in three frequency bands)

Interferer Spectrum: two spectral tilts were included to simulate attenuation of low- or high-frequencies in the interfering programme; it was felt that this would be likely in many audio-on-audio interference situations.

- Low Pass Filtered (200Hz, 9dB/oct)
- Flat
- High Pass Filtered (1kHz, 16dB/oct)

Road Noise: road noise was included to simulate listening in the automotive environment; replay levels are based on data from [24].

- No Noise
- 30mph Road Noise (60dBA)
- 70mph Road Noise (70dBA)

3.5. Experiment Design

With five factors at three levels in the information and entertainment scenarios, and four factors at three levels in the reading/working scenario, the number of combinations required in a full factorial experiment would be 567[1]. This number of combinations results in an unmanageably long test, especially when repeats, familiarisation and breaks are taken into account. To reduce the test time, a fractional factorial experiment design was used. Fractional factorial experiments feature a carefully designed subset of possible factor combinations, significantly reducing the length of the experiment, at the expense of confounding some (generally higher-order) interactions. Third- and higher-order interactions are often non-significant [9], therefore it can be considered a suitable trade-off to facilitate an otherwise unmanageable experiment. Such designs are often used in the preliminary stages of research to suggest directions for more complete observations in the most interesting areas.

In this experiment, a Box-Behnken [25] design was used, facilitating unconfounded analysis of all main effects, but leaving the second- and higher-order interactions confounded. As this design is normally used with continuous factors and intended for analysis with response surface methods, any conclusions made from the ANOVA must be treated with caution, however this was considered a reasonable trade-off for a preliminary experiment designed to investigate the order of magnitude of effects of various factors and ascertain reasonable threshold values, especially given the approximately 80% time saving that the Box-Behnken design afforded in the case of this experiment. The design gives a total of 46 runs for the information and entertainment scenarios and 27 runs for the reading/working scenario.

The experiment comprised of six sessions: three scenarios by two repeats. The presentation order of

[1] Number of combinations N is given by $N =$ (number of levels per variable)$^{\text{number of variables}}$ [9].

the stimuli was randomised within each session, and the order in which the sessions were administered was balanced (there are six possible orders for three scenarios: the six experienced listeners each participated in one of the orders, and the three inexperienced subjects each participated in a different order). For the information and entertainment tasks, the sessions were performed in separate sittings with the option to take a break given twice during the test. For the reading/working scenario, the sessions were performed back-to-back with the opportunity to take one break during each session and one break between sessions. Familiarisation trials were administered prior to each new scenario, using similar programme material to that used in the actual tests. There were four or five trials (depending on the scenario) with one factor varied independently of the others in each trial.

3.6. Stimuli

Stimuli were selected to be representative of real programme items from the categories presented in section 3.4. Minute-long excerpts were used in order that the excerpts were long enough that it was not necessary to loop the items and therefore subjects would not become familiar with the content which would alter the necessary level of concentration and potentially the acceptability of interference, but short enough to select items with minimal variation in loudness.

It was considered important to minimise the subjects' familiarity with the content of the speech items throughout the test in order that the acceptability threshold was not affected (for example, the threshold may become higher as subjects learn the information content of the target audio). Therefore, a pool of multiple excerpts taken from the same broadcast was created, and these items were selected randomly during the test. Each pool contained half of the number of samples required, therefore the number of repeats of each stimulus was reduced to two. It was not felt that it was possible to select a variety of music excerpts that were similar enough to be grouped as the same programme item, therefore the same music items were used throughout the tests.

All stimuli were loudness balanced to equal long-term loudness (LTL) level using the GENESIS Loudness Toolbox [26] implementation of Glasberg and Moore's [27] loudness model for time-varying sounds. The target programme was reproduced at $76\text{dBLA}_{eq(20s)}$ based upon the preferred listening level in a car with background noise at 60dB(A) reported in [24].

3.7. Physical Set-Up

All listening tests took place in the ITU-R BS.1116 [8] standard listening room at the University of Surrey, Guildford. Figure 2 shows the layout of the loudspeakers[2]. All loudspeakers were concealed using acoustically transparent sheets in order to eliminate any bias caused by visibility of the loudspeakers.

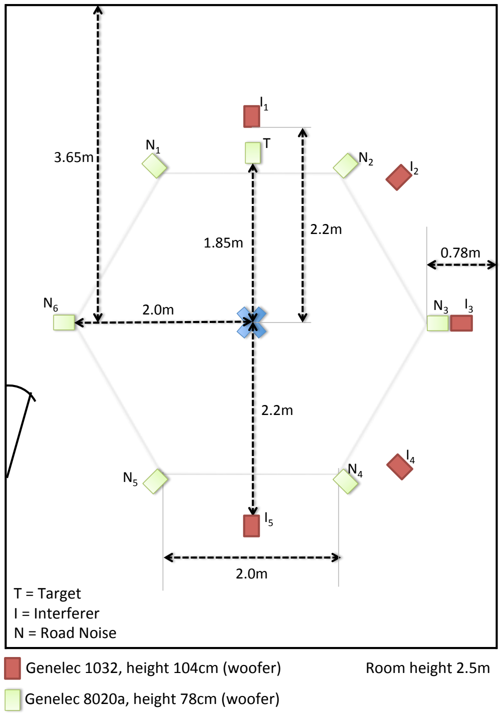


Fig. 2: Experiment Set-Up

[2] Due to unavailability of equipment, the road noise loudspeakers were positioned at a height of 65cm when the experiment was performed with inexperienced listeners.

3.8. Summary

Section 3 contained details of an experiment to determine the threshold of acceptability for an interfering audio programme when listening to a target audio programme, in three different scenarios and with five factors: target programme, interferer programme, interferer location, interferer spectrum, and road noise level. The experiment used a modified method of adjustment task to produce threshold values, and a Box-Behnken experiment design to facilitate analysis of the main effects of each factors as well as two-way interactions.

4. RESULTS

In this section, the results of the experiment described in section 3 are presented. Overall results are presented, followed by a break-down of the results for each scenario.

4.1. Overall Results

Figure 3 shows notched box-plots[3] of the threshold of acceptability for all conditions in the three scenarios, grouped by subject type (experienced/inexperienced). The threshold of acceptability (in dB) is the level at which the subject set the interfering programme with reference to the target programme.

In the information gathering scenario for the experienced listeners, the distribution of results for the experienced listeners was seen to be bimodal; on closer inspection it became apparent that there were large differences in performance between two groups of listeners (see figure 4), suggesting that these groups had possibly interpreted the task in different ways. The results from these groups of subjects were therefore considered separately in all further analysis.

The most striking observation from figure 3 is the difference in threshold between the experienced and inexperienced listeners, with the median threshold for inexperienced listeners approximately 13dB, 10dB and 19dB higher than that of the experienced listeners in the information (low-threshold), entertainment and reading/working scenarios respectively. These results support the conclusion that the high- and low-threshold groups had interpreted the information task differently: the difference between the low-threshold subjects and the inexperienced subjects in the information task is comparable to the differences between the experienced and inexperienced subjects in the other two scenarios.

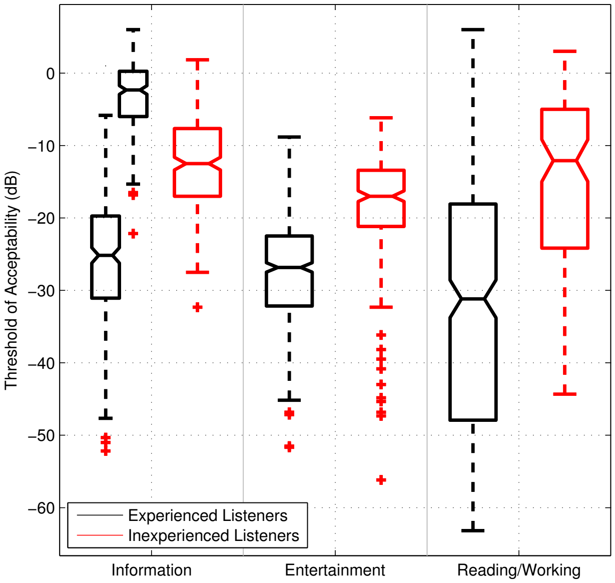


Fig. 3: Overall Results in All Scenarios

The results suggest that with the exception of the group of subjects with high thresholds in the information task, the information and entertainment tasks produce similar thresholds: the points at which 95% of cases are acceptable (for all subjects) fall at −37.3dB and −38.2dB respectively. It is clear that there is a much wider variation in the case of the reading/working scenario; this can be attributed to the greater influence of factor levels as will be seen in section 4.5.

[3] When the notches do not overlap, this gives some indication that the medians are significantly different at the 5% level [28]).

4.2. Analysis of Variance

The results for each scenario were analysed with separate ANOVAs. As discussed in section 3.5, the Box-Behnken design used in this experiment means that the second- and higher-order interactions are confounded; they have therefore been omitted from the models. The ANOVA results must be treated with caution as it is not possible to say what effect the second-order interactions may have, and Box-Behnken designs are generally used with continuous factors for analysis with response surface methods. However, the ANOVA tables do give some indication of the significance and size of the main effects.

Tables 1a, 1b and 1c show ANOVA results for the information, entertainment and reading/working[4] scenarios respectively, including estimates of effect size (η_p^2). The model used includes the main effects of subject, repeat and the physical parameters present in each scenario; all terms were modelled as fixed factors[5]. The adjusted r^2 values for each of the models are reasonably high (0.876, 0.600 and 0.600 for the information, entertainment and reading/working scenarios respectively), suggesting a good fit to the data. Kolmogorov-Smirnov (K-S) tests (with Lilliefors correction) performed on the residuals of each model suggest that the residuals are not normally distributed for the information and entertainment scenarios (p=0.009 and p=0.002 respectively), however histograms and Q-Q plots (included in appendix B) suggest that the residuals are approximately normally distributed with some deviations at the extremities of the data. The model residuals for the reading/working scenario were shown to be normally distributed (K-S test, p=0.188). The results for each scenario are discussed in more detail in the following sections.

4.3. Information Gathering Scenario

The results of the ANOVA (table 1a) suggest that there is a significant effect (p<0.01) of all of the terms with the exception of repeat.

Figure 4 shows box-plots of subject performance broken down by repeat. It can be seen that there are small differences between the two replicate ratings for subjects 1 and 9 only, hence the repeat term being non-significant in the ANOVA model. The figure also highlights the large differences between the low- and high-threshold groups of experienced listeners as discussed in section 4.1.

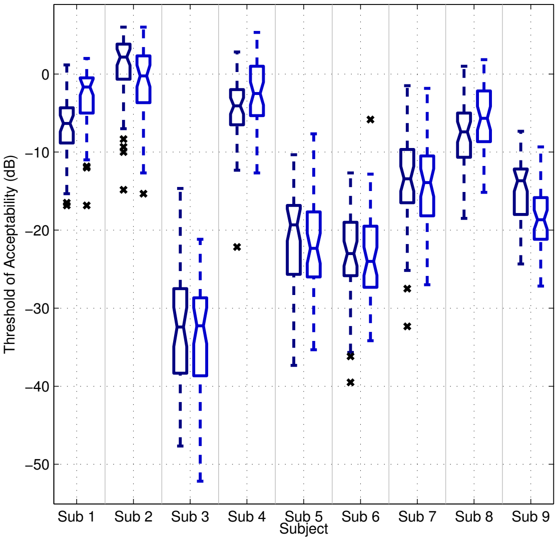
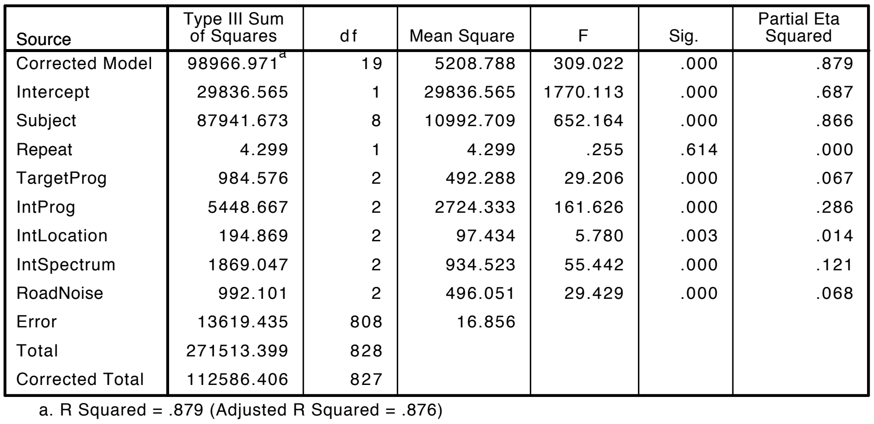


Fig. 4: Information Scenario: Effect of Subject

Table 2 shows the difference in acceptability threshold between the conditions producing the highest and lowest thresholds for each factor, detailing the factor levels producing the extreme threshold values. Figure 5 shows error bar plots for the 4 factors with the greatest effect on threshold (abbreviations are expanded in table 6, appendix A). The effect of all of the factors is fairly intuitive: speech-on-speech interference has a lower threshold of acceptability than music-on-speech; low-pass filtering increases the threshold (possibly because of a decrease in sibilance or transients); adding road noise increases acceptability (presumably as the interferer becomes more masked); and sports commentary targets produce a slightly higher threshold (possibly due to the consistent crowd noise). The effect of location was less pronounced.

[4] As discussed in section 4.5, the results of subject seven were omitted from the ANOVA model for the reading/working scenario.

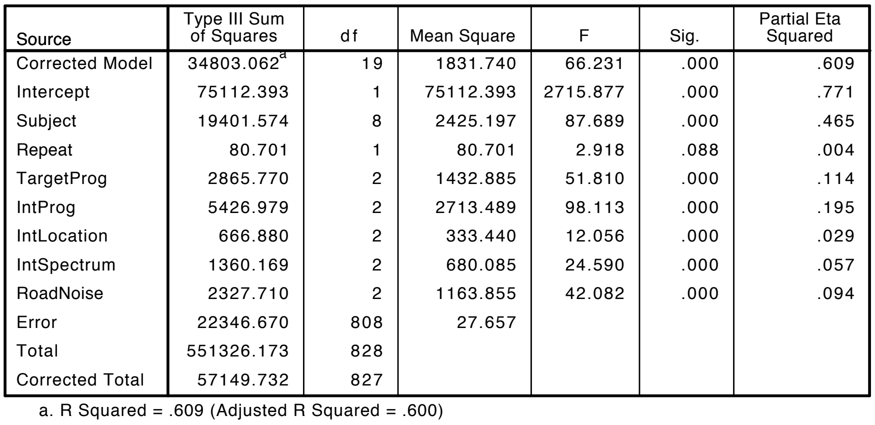
[5] Models were also created with subject and repeat as random factors; this produced no differences to the models presented.



Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	98966.971[a]	19	5208.788	309.022	.000	.879
Intercept	29836.565	1	29836.565	1770.113	.000	.687
Subject	87941.673	8	10992.709	652.164	.000	.866
Repeat	4.299	1	4.299	.255	.614	.000
TargetProg	984.576	2	492.288	29.206	.000	.067
IntProg	5448.667	2	2724.333	161.626	.000	.286
IntLocation	194.869	2	97.434	5.780	.003	.014
IntSpectrum	1869.047	2	934.523	55.442	.000	.121
RoadNoise	992.101	2	496.051	29.429	.000	.068
Error	13619.435	808	16.856			
Total	271513.399	828				
Corrected Total	112586.406	827				

a. R Squared = .879 (Adjusted R Squared = .876)

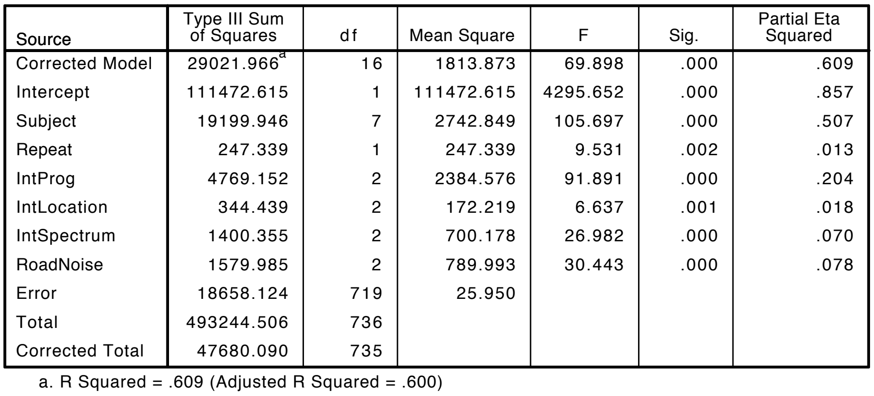
(a) Information Scenario ANOVA



Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	34803.062[a]	19	1831.740	66.231	.000	.609
Intercept	75112.393	1	75112.393	2715.877	.000	.771
Subject	19401.574	8	2425.197	87.689	.000	.465
Repeat	80.701	1	80.701	2.918	.088	.004
TargetProg	2865.770	2	1432.885	51.810	.000	.114
IntProg	5426.979	2	2713.489	98.113	.000	.195
IntLocation	666.880	2	333.440	12.056	.000	.029
IntSpectrum	1360.169	2	680.085	24.590	.000	.057
RoadNoise	2327.710	2	1163.855	42.082	.000	.094
Error	22346.670	808	27.657			
Total	551326.173	828				
Corrected Total	57149.732	827				

a. R Squared = .609 (Adjusted R Squared = .600)

(b) Entertainment Scenario ANOVA



Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	29021.966[a]	16	1813.873	69.898	.000	.609
Intercept	111472.615	1	111472.615	4295.652	.000	.857
Subject	19199.946	7	2742.849	105.697	.000	.507
Repeat	247.339	1	247.339	9.531	.002	.013
IntProg	4769.152	2	2384.576	91.891	.000	.204
IntLocation	344.439	2	172.219	6.637	.001	.018
IntSpectrum	1400.355	2	700.178	26.982	.000	.070
RoadNoise	1579.985	2	789.993	30.443	.000	.078
Error	18658.124	719	25.950			
Total	493244.506	736				
Corrected Total	47680.090	735				

a. R Squared = .609 (Adjusted R Squared = .600)

(c) Reading/Working Scenario ANOVA

Table 1: ANOVA Tables for All Scenarios

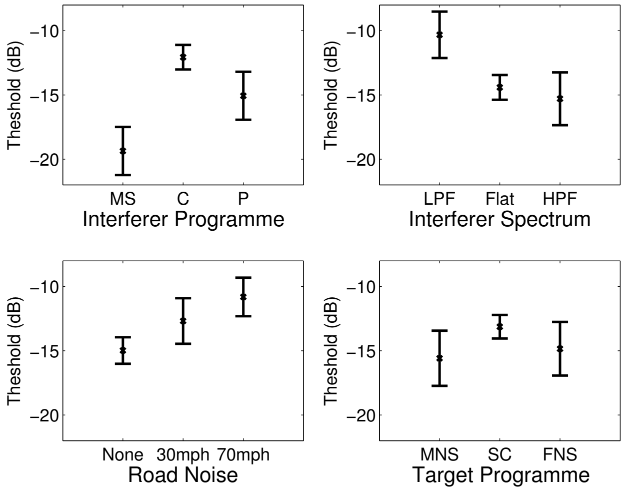


Fig. 5: Information Scenario Main Effects. Error bars show ±1.96*Standard Error

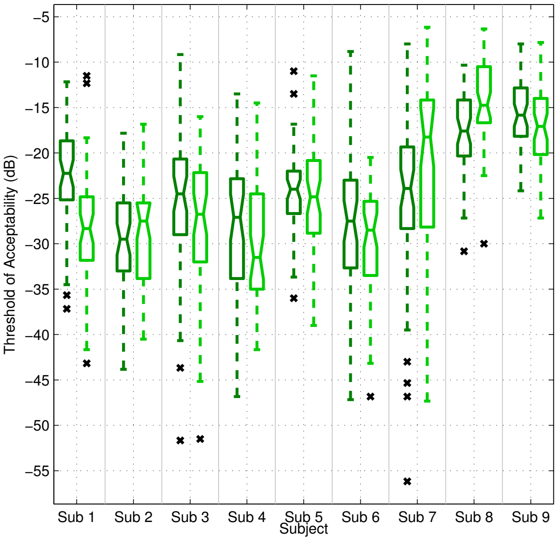


Fig. 6: Entertainment Scenario: Effect of Subject

4.4. Entertainment Scenario

Table 1b shows that as in the information scenario, all of the terms with the exception of the repeat ratings were shown to have significant effects (p<0.01).

It can be seen from figure 6 that, as in the information scenario, the only significant differences between the first and second replicate are marginal (subjects 1, 7 and 8), and the replicate effect is non-significant overall. The effect of subject on threshold is less pronounced than in the information scenario, as the experienced listeners are not divided into two groups. However, the inexperienced listeners still produce a higher threshold.

Table 3 shows the difference in acceptability threshold between the conditions producing the highest and lowest thresholds for each factor, detailing the factor levels producing the extreme threshold values. It can be seen that the interferer programme is again the most influential factor with a difference of 8dB between the highest and lowest thresholds; the factor levels producing the highest and lowest threshold are the same as in the information task. Target programme has a larger effect in the entertainment task; this could be attributed to the nature of the programme material used in this scenario, with vocal pop music more heavily compressed and therefore masking the interfering programme more consistently. The magnitude of the effect of road noise is similar to that in the information scenario, and that of spectrum slightly lower. Again, interferer location had the smallest effect on threshold. Figure 7 shows error bar plots for the most influential factors.

4.5. Reading/Working Scenario

Unlike the previous scenarios, the effect of repeat was found to be significant (p=0.001, η_p^2=0.022). Box-plots of subject performance (figure 9) showed that there was a large difference in threshold between the two replicates performed by subject seven. In order to assess the performance of this subject, fader use plots were created (figure 13, appendix C) for the trials where the absolute difference between replicates was greater than the mean absolute difference for all other subjects. These plots show large differences between the first and second replicate, often (though not always) displaying little use of the fader during the first replicate and a significantly lower threshold in the second replicate. Because of this inconsistency, the results of subject seven were removed from all

Factor	**Difference**	**High Threshold**	**Low Threshold**
Interferer Programme	7.30dB	Instrumental Classical Music	Male Speech
Interferer Spectrum	4.97dB	LPF	HPF
Road Noise	4.17dB	70mph	None
Target Programme	2.45dB	Sports Commentary	Male News Speech
Interferer Location	1.60dB	Diffuse	0 Degrees

Table 2: Influence of Factors in Information Scenario. 'Difference' indicates the difference in dB between the levels producing the highest and lowest thresholds.

Factor	**Difference**	**High Threshold**	**Low Threshold**
Interferer Programme	7.99dB	Instrumental Classical Music	Male Speech
Target Programme	6.25dB	Vocal Pop Music	Instrumental Classical Music
Road Noise	5.19dB	70mph	None
Interferer Spectrum	3.97dB	LPF	HPF
Interferer Location	2.50dB	Diffuse	90 Degrees

Table 3: Influence of Factors in Entertainment Scenario. 'Difference' indicates the difference in dB between the levels producing the highest and lowest thresholds.

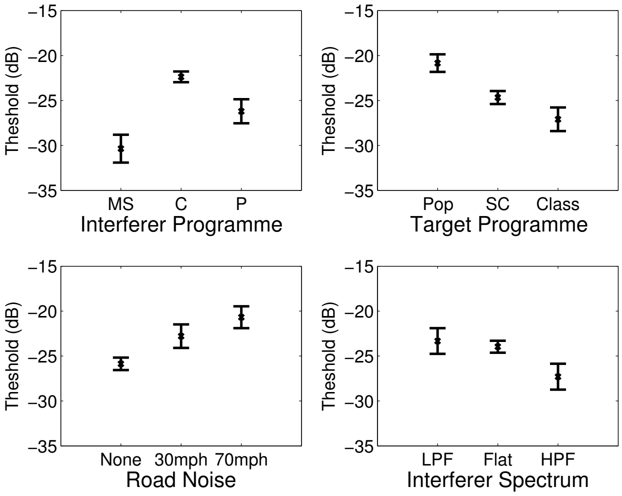


Fig. 7: Entertainment Scenario Main Effects. Error bars show ±1.96*Standard Error

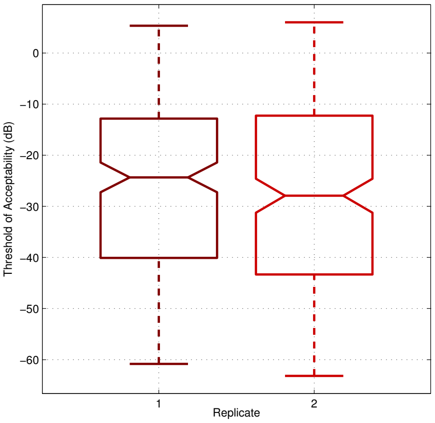


Fig. 8: Reading/Working Scenario: Effect of Replicate

further analysis. Table 1c shows that the effect of replicate (figure 8) is still significant, with the effect size approximately halved (η_p^2=0.013).

Figure 9 shows considerable differences between thresholds produced by various listeners, with no obvious groups or trends. It is difficult to explain these differences other than to say that reading in the presence of audio is seemingly a highly personal task; different people are able to perform the task with different levels of audio interference. This is supported by comments from some subjects, for example that they would or would not normally work with music on, and is also supported in the literature (e.g. [29]).

Table 4 shows the difference in acceptability

Factor	Difference	High Threshold	Low Threshold
Road Noise	19.38dB	70mph	None
Interferer Programme	15.27dB	Instrumental Classical Music	Male Speech
Interferer Spectrum	5.21dB	LPF	HPF
Interferer Location	2.16dB	90 Degrees	0 Degrees

Table 4: Influence of Factors in Reading/Working Scenario. 'Difference' indicates the difference in dB between the levels producing the highest and lowest thresholds.

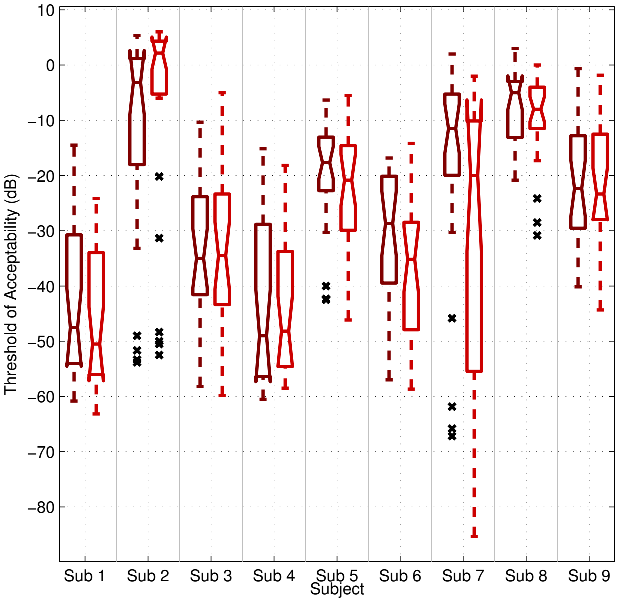


Fig. 9: Reading/Working Scenario: Effect of Subject. Note: Subject 7 was removed from further analysis (as discussed in section 4.5).

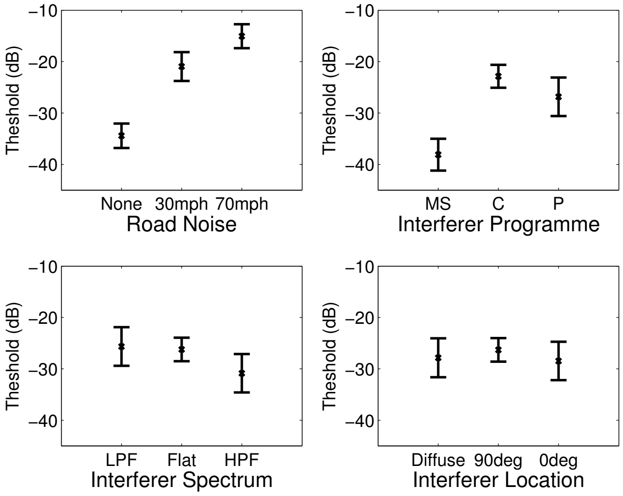


Fig. 10: Reading/Working Scenario Main Effects. Error bars show ±1.96*Standard Error

threshold between the conditions producing the highest and lowest thresholds for each factor, detailing the factor levels producing the extreme threshold values, and figure 10 shows error bars for the four factors. The order of importance of the factors is somewhat different to the previous scenarios, and the magnitude of the important effects is much larger. Introducing road noise at 70mph increases the threshold of acceptability by approximately 19dB; this can be attributed to the extra masking provided by the road noise when there is no target programme. The magnitude of the effect of interferer programme is similarly inflated to 15dB, with the same programme items as in the previous scenarios producing the lowest and highest thresholds. The interferer spectrum and location have similar effects to the information and entertainment scenarios.

5. CONCLUSIONS

In section 4, the results from an experiment to determine the threshold of acceptability for an interfering audio programme were presented. The 50% and 95% acceptable points for each scenario are detailed in table 5. These results provide useful information as to the level of audio interference which may be considered acceptable during the performance of certain tasks.

It is clear that there are pronounced differences

	Experienced (50% / 95%)	**Inexperienced (50% / 95%)**
Information (HT)	-2.33dB / -11.67dB	N/A
Information (LT)	-25.17dB / -42.23dB	-12.50dB / -23.62dB
Entertainment	-26.83dB / -39.17dB	-17.00dB / -31.35dB
Reading/Working	-31.17dB / -57.67dB	-12.08dB / -34.87dB

Table 5: 50% and 95% Acceptable Points for All Scenarios

between subjects: inexperienced listeners produced median threshold values between 10dB and 18dB above those of experienced listeners, and there were also large differences between individual subjects. Some of these differences were attributed to a different understanding of the task between subjects, suggesting that in future experiments it may be wise to introduce some manner of assuring that the subjects have understood the task in the same way. At the same time, some of these differences may be attributed to personal differences between listeners (e.g. temperament, mood, prior experience etc.).

It is apparent that the effect of physical parameters is somewhat determined by the task, and is seemingly heavily influenced by the target programme. In the reading/working scenario, there was up to 19dB difference between thresholds produced at different levels of road noise and for different interferer programmes. The effect of each factor was less pronounced in the information and entertainment scenarios, with the most influential parameters being interferer programme (approximately 8dB between the means for the highest and lowest threshold groups). In conclusion, it seems that interferer programme has the greatest effect on threshold, followed by road noise level, spectrum and target programme which are more or less important depending on scenario. Interferer location was found to be the least influential parameter in all cases.

The Box-Behnken design used in this experiment was sufficient for a broad look at the effect of various physical parameters, although it would be beneficial to perform a more detailed study in which the second- and higher-order interactions were estimable, as this would provide useful further insight into the perception of target audio in the presence of interfering audio.

5.1. Future Work

The experiment reported in this paper served to give an overview of the experience of listening to target audio in the presence of interfering audio, providing acceptability threshold values for various scenarios and detailing the effect of a number of physical parameters. This is a research area with a wide range of applications: the results may be applied to any system which aims to mitigate the effects of audio-on-audio interference, for example, noise-cancellation systems or source separation algorithms. However, the subjective effects of audio-on-audio interference is a research area which has not received a lot of attention to date.

This work leads towards further investigation into the relationship between physical parameters of target and interfering audio programmes and the experience of a listener. In order to gather incisive data about audio-on-audio interference situations, it is desirable to elicit suitable subjective attributes by which such situations can be rated. It is also necessary to perform a thorough investigation of the physical parameters which may effect listener experience in an audio-on-audio interference situation.

6. ACKNOWLEDGEMENTS

This research was carried out as part of the 'Perceptually Optimised Sound Zones' (POSZ) project funded by Bang & Olufsen. The authors would like to acknowledge the University of Surrey students and staff who participated in listening tests, the advice of Professor Per B. Brockhoff, and the input received from members of the POSZ project team.

7. REFERENCES

[1] R. Rylander, S. Sörensen, and A. Kajland. Annoyance reactions from aircraft noise exposure. *Journal of Sound and Vibration*, 24(4):419–444, October 1972.

[2] Sanford Fidell. Nationwide urban noise survey. *The Journal of the Acoustical Society of America*, 64(1):198, 1978.

[3] E. Öhrström, M. Björkman, and R. Rylander. Laboratory annoyance and different traffic noise sources. *Journal of Sound and Vibration*, 70(3):333–341, June 1980.

[4] J. Nemecek and E. Grandjean. Noise in landscaped offices. *Applied Ergonomics*, 4(1):19–22, March 1973.

[5] E. Boman and I. Enmarker. Factors affecting pupils' noise annoyance in schools. *Environment and Behavior*, 36(2):207 –228, March 2004.

[6] S.A. Falk and N.F. Woods. Hospital noise—levels and potential health hazards. *New England Journal of Medicine*, 289(15):774–781, 1973.

[7] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann. Subjective and objective quality assessment of audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2046–2057, September 2011.

[8] ITU-R. Recommendation BS.1116-1: methods for the subjective assessment of small impairments in audio systems including multichannel sound systems. 1997.

[9] S. Bech and N. Zacharov. *Perceptual audio evaluation: theory, method and application.* John Wiley and Sons, 2006.

[10] G.A. Gescheider. *Psychophysics: the fundamentals.* Routledge, 1997.

[11] W.F. Druyvesteyn, R.M. Aarts, A.J. Asbury, P. Gelat, and A. Ruxton. Personal sound. *Proceedings of the Institute of Acoustics*, 16(2):571–585, 1994.

[12] F.J. Hubach and B. Edwards. Empirical determination of sound isolation requirements for recording studio isolation booths. 1992.

[13] D.E. Bishop. Judgments of the relative and absolute acceptability of aircraft noise. *The Journal of the Acoustical Society of America*, 40(1):108–122, July 1966.

[14] E. C. Keighley. The determination of acceptability criteria for office noise. *Journal of Sound and Vibration*, 4(1):73–87, July 1966.

[15] S. Jumisko-Pyykkö, V.K. Malamal Vadakital, and J. Korhonen. Unacceptability of instantaneous errors in mobile television. In *Proceedings of the 8th conference on Human-computer interaction with mobile devices and services - MobileHCI '06*, page 1, Helsinki, Finland, 2006.

[16] S Bech. Spatial aspects of reproduced sound in small rooms. *The Journal of the Acoustical Society of America*, 103(1):434–445, January 1998.

[17] A. Hesse. Comparison of several psychophysical procedures with respect to threshold estimates, reproducibility and efficiency. *Acustica*, 59:263–273, 1986.

[18] I.J. Hirsh and C.S. Watson. Auditory psychophysics and perception. *Annual Review of Psychology*, 47(1):461–484, February 1996.

[19] E. C. Poulton. *Bias in quantifying judgements.* Erlbaum, 1989.

[20] 'PowerMate - USB Controller - Griffin Technology'. Available at: https://store.griffintechnology.com/powermate [accessed June 1, 2011], 2011.

[21] Evening Standard. 'Columnists'. Available at: http://www.thisislondon.co.uk/standard-home/columnist.do [accessed July 31, 2011], 2011.

[22] R. Flesch. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233, 1948.

[23] V. Pulkki. Spatial sound reproduction with directional audio coding. *Journal of the Audio Engineering Society*, 55(6):503–516, 2007.

[24] E. Benjamin and B. Crockett. Preferred listening levels in the automotive environment. New York, NY, USA, October 2005.

[25] G.E.P. Box and D.W. Behnken. Some new three level designs for the study of quantitative variables. *Technometrics*, 2(4):455–475, November 1960.

[26] GENESIS Acoustics. 'Loudness Online'. Available at: http://www.genesis-acoustics.com/en/index.php?page=32 [Accessed July 18, 2011], 2011.

[27] B.R. Glasberg and B.C.J. Moore. A model of loudness applicable to Time-Varying sounds. *Journal of the Audio Engineering Society*, 50(5):331–342, May 2002.

[28] R. McGill, J.W. Tukey, and W.A. Larsen. Variations of box plots. *The American Statistician*, 32(1):12–16, February 1978.

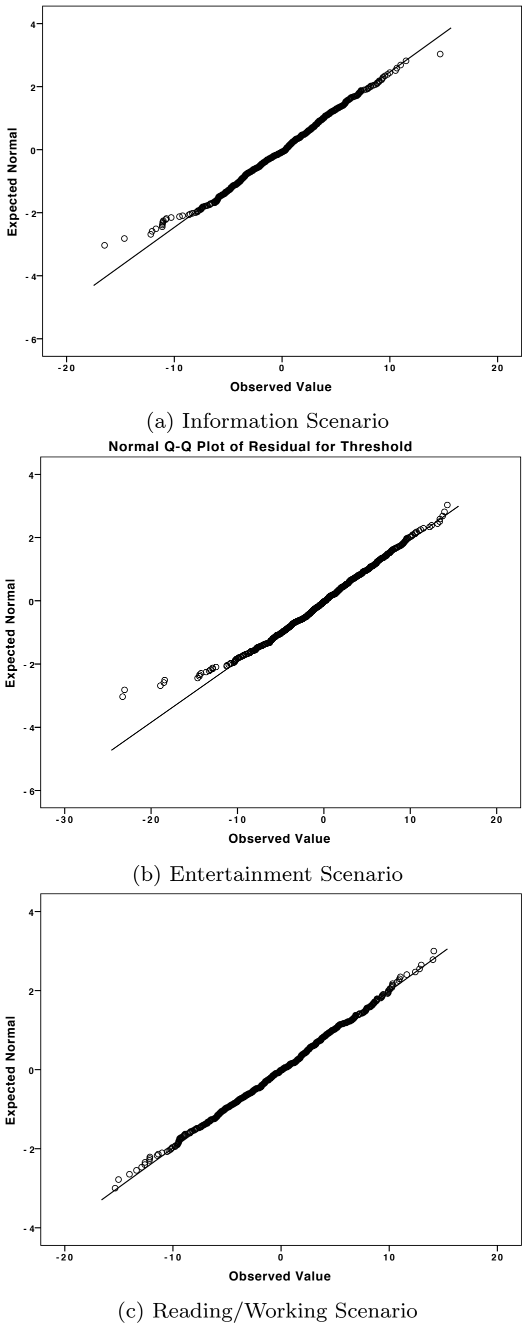
[29] L. Daoussis and S.J. Mckelvie. Musical preferences and effects of music on a reading comprehension test for extraverts and introverts. *Perceptual and Motor Skills*, 62(1):283–289, February 1986.

8. APPENDIX A: PROGRAMME MATERIAL

Programme Item	**T/I**	**Source**	**Abbrev.**
Male News Speech	T	BBC Radio News	MNS
Female News Speech	T	BBC Radio News	FNS
Sports Commentary	T	BBC Radio Football Commentary	SC
Vocal Pop Music	T	The Killers - *"On Top"*	Pop
Instrumental Classical Music	T	Brahms - *Hungarian Dance No. 18* (String Orchestra)	Class
Male Speech	I	BBC Radio 4 - *"Points of View"*	MS
Vocal Pop Music	I	The Bravery - *"Give In"*	P
Instrumental Classical Music	I	Mahler - *Symphony No. 5 Mov. 4* (String Section)	C

Table 6: Description of Programme Items (T = Target, I = Interferer)

9. APPENDIX B: Q-Q PLOTS AND HISTOGRAMS OF RESIDUALS

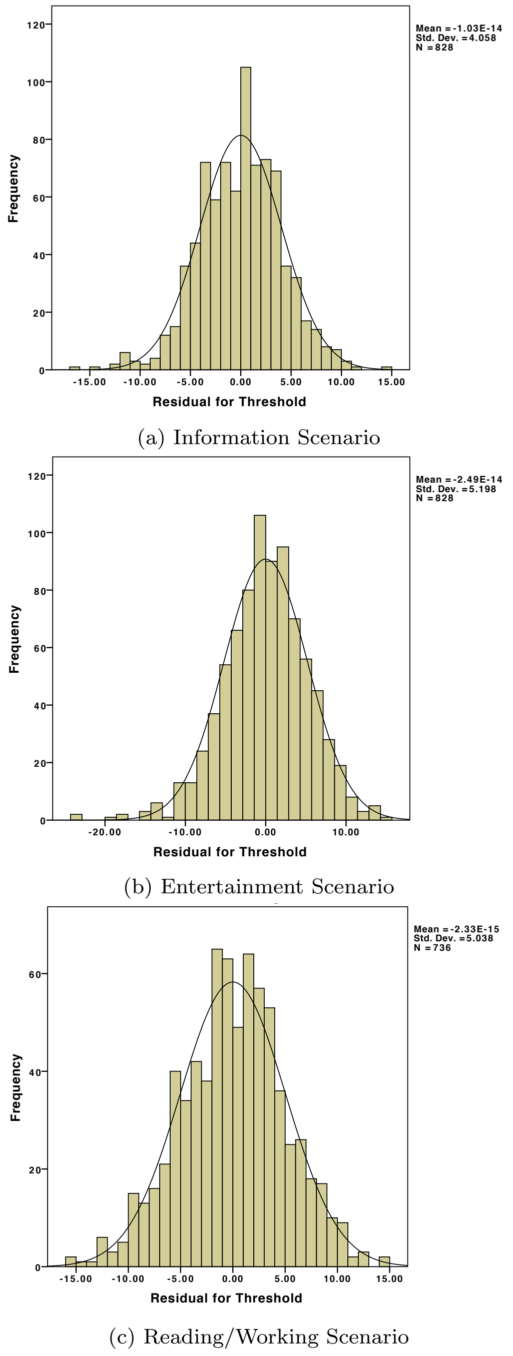


(a) Information Scenario

(b) Entertainment Scenario

(c) Reading/Working Scenario

Fig. 11: Q-Q Plots of Residuals for All Scenarios



(a) Information Scenario

(b) Entertainment Scenario

(c) Reading/Working Scenario

Fig. 12: Histograms of Residuals for All Scenarios

10. APPENDIX C: SUBJECT 7 FADER USE

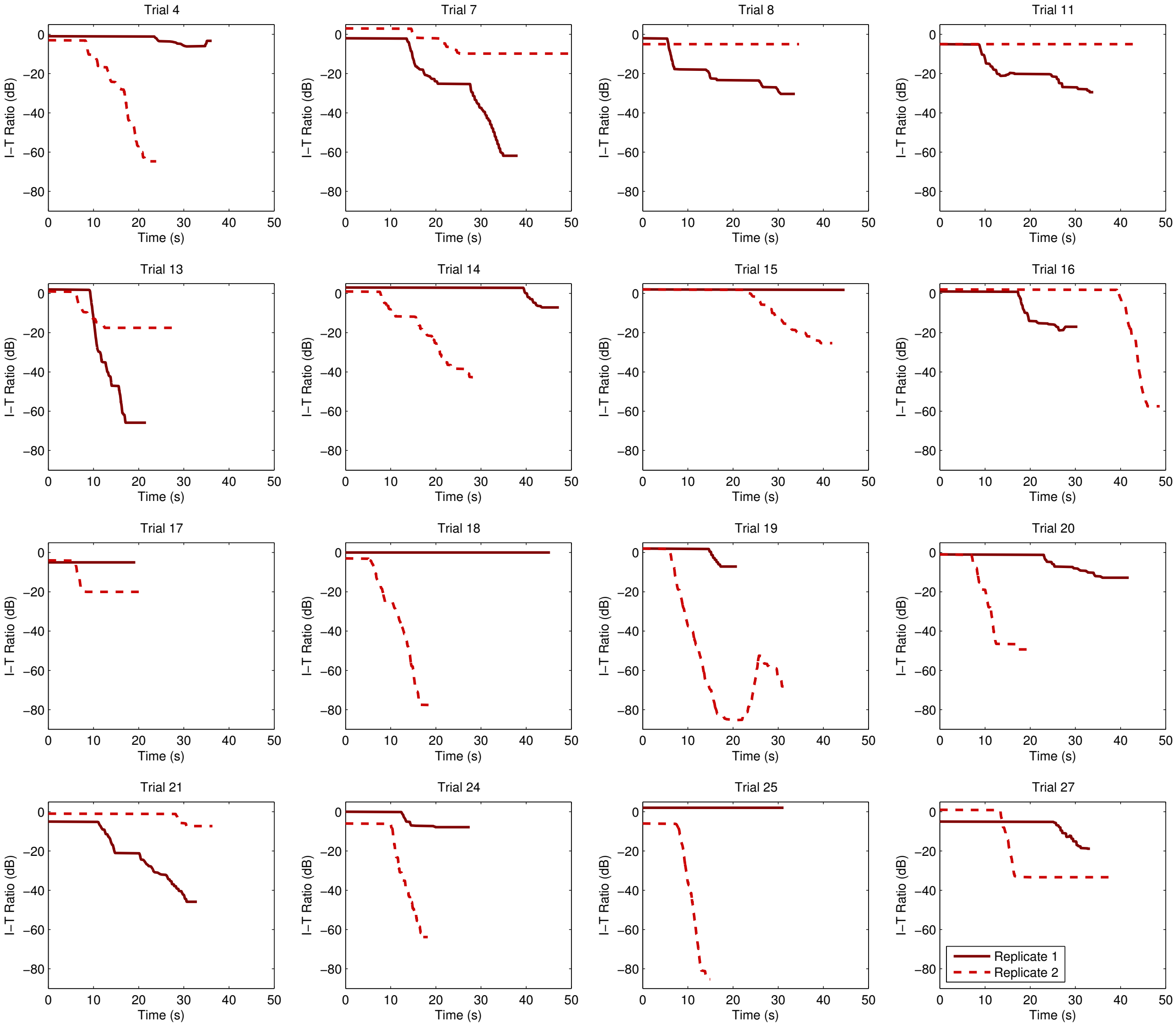


Fig. 13: Fader Use Plots for Subject 7 in Reading/Working Scenario