Statistical Inference of User Experience of Multichannel Audio on Mobile Phones

Fesal Toosy[1, *] and Muhammad Sarwar Ehsan[1]

Abstract: Mobile phones and other handheld electronic devices are now ubiquitous and play an important role in our everyday lives. Over the last decade, we have seen a sharp rise in the sophistication of both hardware and software for these devices, thus significantly increasing their utility and use. Electronic devices are now commonly used for the streaming of audio and video and for the regular playback of music. Multichannel audio has now become a popular format and with recent updates in software, the latest audio codecs that support this format can effectively be played back on most electronic devices. As a result, more audio content that is now compatible with these devices is being produced. It is important to perform subjective quality tests for multichannel audio verses traditional stereo and mono on devices like mobile phones while using headphones. This paper presents the results and analysis of such listening tests that evaluate basic audio quality and quality of experience. The results show that while multichannel audio scores higher than stereo and mono for both of these attributes, the difference in score was less for the ratings of quality of experience. The results were processed using an Analysis of Variance (ANOVA) to show the statistical significance. Calculation of the effect size showed that the audio content had little bearing on the results.

Keywords: Subjective assessment, multichannel audio, MUSHRA, listening test.

1 Introduction

Improvements on capturing and reproducing a sound scene has be an ongoing topic of research for the better part of a century. Alan Blumlein [Blumlein (2000)] made a major breakthrough in 1931 with his invention of the binaural sound now known as stereophonic sound and the Blumlein Pair stereo recording technique. Recent approaches like surround sound [Rumsey (2001)], brought setups like 5.1 (three loudspeakers in the front and two loudspeakers in the rear) and 7.1 (three loudspeaker in the front, two on the side and two in the rear). The next generation of surround sound introduced speakers in the height dimension leading to setups like 9.1 [Chabanne, McCallus, Robinson et al. (2010)], 10.2, 22.2 [Hamasaki, Matsui, Sawaya et al. (2011)] giving a more immersive effect, hence the

[1] University of Central Punjab, Lahore, 54700, Pakistan.

[*] Corresponding Author: Fesal Toosy. Email: fesal@ucp.edu.pk.

Received: 22 May 2020; Accepted: 01 July 2020.

CMC. doi:10.32604/cmc.2020.011667 www.techscience.com/journal/cmc

term “3D Audio”. Different methods have been developed to produce signals to feed multichannel audio setups. These are based on “channels”, “objects” and “scenes”.

In channel based audio, spatial sound is represented as a set of waveforms called channel signals. Each channel signal is designated to feed a loudspeaker. Each loudspeaker is in a known position relative to the listener. 2.0 (Stereo) and 5.1 formats are the most common formats in channel based audio. The drawback here is that the produced content is tied to one specific speaker configuration. Though it could be played back on other loudspeaker configurations, the resulting quality would be less than optimal. In a more recent method, “object based audio” each sound source in the sound field is represented by an “object”. The location of the sound source and the sound signal itself, defines the object. A renderer at the playback system converts these objects into suitable signals to feed to the loudspeakers. This method is independent of the speaker format. “Scene based audio” does a complete representation of the sound field by calculating the sound pressure at every point in space. This can be done using spherical harmonics and the whole matter reduces to calculating a small amount of coefficients. These coefficients are a function of time and are called higher order ambisonic signals. A renderer at the playback will use these coefficients and generate feeds for the speakers such that the resulting sound matches the sound captured at the location as closely as possible [Roginska (2017)].

Multichannel audio can effectively be implemented on a loudspeaker system in a room or theatre. Implementations on mobile phones and handheld electronic devices are moderately effective, but still in their early stages. The term binaural is now used for “sound where the two-channel sound entering a listener’s ear has been filtered by a combination of time, intensity and spectral cues intended to mimic human localization cues”. When this is done through signal processing, it is called binaural rendering. Head Related Transfer Functions (HRTF’s) are superimposing of these binaural cues on the sound before it reaches the eardrum [Roginska (2017)]. It characterizes how an ear receives sound from a point in space and gives the effect of different sounds coming from different locations in a three dimensional (3D) space. Multichannel audio cannot be achieved on a single loudspeaker but with binaural rendering, an equivalent effect can be achieved while using a pair of headphones.

Some cell phone manufacturers began with implementations of multichannel and 3D audio technologies on the embedded hardware level. Phones like the Samsung Galaxy S9, Nokia 6, Apple IPhone X, Huawei P20 and eight others [Dolby (2020)] were Dolby Atmos/Dolby Surround compatible. These phone sets either came with two loudspeakers or used the phones earpiece as a second speaker but gave a much better spatial audio effect with headphones. The VLC player by VideoLAN [VLC 3.0 Vetinari (2020)] is one of the first open source software solutions that supports 3D and multichannel audio on handheld electronic devices. The VLC player version 3.0 Vetinari [VLC 3.0 Vetinari (2020)] supports codecs like HEVC, DTS-HD and TrueHD. It also has an Ambisonics audio renderer up to the 3rd order and an audio binauralizer that supports 5.1 and 7.1 channels.

Assessment of audio can be done by both Subjective and Objective methods. Subjective assessment methods involve ‘listening tests’ in which actual users/participants listen to and rate the audio that is under test. Objective assessment usually involves a software algorithm that automatically rates the audio without having to involve actual listeners/people. Subjective listening tests are both expensive and time consuming which makes them difficult and inconvenient to repeat. Despite these drawbacks, they are still the most reliable method for assessment of audio as these are the only tests in which real users actually interface with and experience the audio material. Objective assessment techniques are evaluated based on how well their scores correlate with the scores of the subjective tests. Standards like the ITU-R BS.1387 Perceptual Evaluation of Audio Quality (PEAQ) [ITU (2018a)] and the ITU-T P.863 Perceptual Objective Listening Quality Assessment (POLQA) [ITU (2018b)] are the most current and popular standards for objective assessment of speech and audio. Both standards were originally designed for the objective assessment of speech but can be used for the assessment of audio quality.

The ITU-R BS.1543-3 [ITU (2012)] "Method for subjective assessment of intermediate quality level of audio systems" (MUSHRA) is the most suitable method for assessment of audio systems that has medium to large impairments. The experiments described and analyzed in this paper use audio excerpts with conditions that have similar impairments. Section 2 reviews similar tests performed and their results. Section 3 describes the relevant parts of the standard ITU-R BS.1534-3 (MUSHRA) [ITU (2012)] that were used in these experiments. Section 4 describes the experiments conducted that were based on the above-mentioned standard. Sections 5 describes the results of these listening tests with graphical illustrations. Section 6 first explains and justifies the statistical methods used for the analysis of the results and shows the calculations made by these methods and the subsequent inferences. Sections 7 gives a detailed discussion based on the results and statistical analysis and Section 8 concludes the paper.

2 Previous work

Shoeffler et al. [Schoeffler, Silzle and Herre (2017)] did a subjective evaluation of 3D audio while comparing it to surround and stereo formats. This evaluation used Basic Audio Quality (BAQ) and Overall Listening Experience (OLE) as attributes on a 22.2 speaker system. It also included a follow up session for Overall Listening Experience on headphones. The BAQ session used the standard MUSHRA scale of 0-100 with increments of 20 while the OLE session used a five-star Likert scale [Likert (1932)]. The results showed that the increase in perceived BAQ score was the same for stereo to surround and surround to 3D. For the OLE ratings, the increase from surround to 3D audio was lower (as compared to that from the BAQ scores).

Toosy et al. [Toosy and Ehsan (2019a)] conducted two experiments that were based on the ITU-R BS. 1534-3 (MUSHRA), one of which measured the perceived basic audio quality (BAQ) and the other recorded the users quality of experience (QoE). The results showed that 3D Audio and multichannel audio performed better than stereo and mono for both BAQ and QoE scores.

Toosy et al. [Toosy and Ehsan (2019b)] further conducted two experiments based on the ITU-R BS.1116-3 [ITU (2015a)]. In the first experiment, audio excerpts in multichannel format were rated against stereo while played back on a mobile phone using headphones. The results showed that multichannel audio outperformed stereo in terms of BAQ. In the second experiment [Toosy and Ehsan (2019c)], the same audio (in both multichannel and stereo format) was played back and rated in an ITU-R BS.1116-3 [ITU (2015a)] based listening test on the audio system of an automobile. The results of this experiment were compared to the results of the test done on a mobile phone. The same audio excerpts (in both multichannel and stereo formats) when played back in an automobile got marginally higher scores in perceived audio quality. In stereo format, the excerpts performed equally well in terms of rated BAQ when played back on mobile phones using headphones or in an automobile on loudspeakers. The participants however took more time to rate the audio in the automobile than on the mobile phone. This was due to the difference in frequency response of both sound systems.

3 ITU-R BS.1543-3 (MUSHRA) [ITU (2012)]

3.1 Test method

The International Telecommunication Union Radio Communication sector recommendation ITU-R BS.1534-3 [ITU (2012)], 'Method for Subjective Assessment of Intermediate Quality Level of Audio Systems' also known as 'Multi Stimulus test with Hidden Reference and Anchor' (MUSHRA) is suitable for evaluating sound systems with medium and large impairments. In each trial, the subject (also called listener) listens to the reference version, the low and mid anchor and all versions of the test signal processed by the systems under test. The reference version is usually an unprocessed, lossless version of the audio clip to be tested. The low and mid anchors are usually low pass filtered versions of the reference at a cut-off frequency of 3.5 kHz and 7 kHz respectively. All other versions refers to any encoded or compressed version of the reference. Other anchors can be used if the recommended anchors are not suitable for the experiment and their use should be justified in the test report. The subjects grade the signals on a continuous quality scale (CQS) from 1 to 100 with five intervals: bad, poor, fair, good and excellent.

3.2 Listening panel

The standard [ITU (2012)] emphasizes that the listeners should have normal hearing and should be able to perceive differences between the test items. Twenty participants are enough for drawing an appropriate conclusion from the test. Experienced assessors should be included in the final data analysis though results from inexperienced listeners can also be included. A training procedure is mandatory and pre-screening of assessors is recommended. This training should cover the entire test procedure and clearly explain the attributes that will be used for judging the audio quality. The training has a large impact on the results when using inexperienced listeners.

3.3 Test material

The standard [ITU (2012)] recommends that the audio excerpts should not have a duration of more than 10 seconds, with a maximum duration of 12 seconds. This is because it would cause fatiguing of the participants, which would definitely affect the results and it would unnecessarily increase the total duration of the test. In some cases a longer audio excerpt can be used if justified e.g. if the audio clip involves a slow moving trajectory of sound then it can be longer in duration to allow for the full intended spatial effect. The test material should well represent the range of material generally used in the application for which the test is being conducted for.

3.4 Attributes

The attributes used to judge the signal tested on an advanced sound system are basic audio quality, timbral quality, localization quality and surround quality. Only one attribute should be tested in a single experiment. This is because when given more than one attribute to assess, the participant can get overburdened and confused which would probably result in unreliable grading for all the signals. In the event that more than one attribute needs to be graded, basic audio quality should be graded first. The second attribute should then be graded in a separate test at least 24 hours later [ITU (2012)].

4 Experiments

4.1 Listening panel

Twenty-two participants, 14 male and 8 female aged 14 to 46, mean age 30 and standard deviation 9 took part in the experiments. All participants had taken part in ITU-R listening tests before. Two were professional musicians, one of which had a university degree in Musicology, the other a sound engineer and professional keyboard player. Five were amateur musicians and three had experience in audio coding.

Emphasis was given to the training session as the reliability of the results greatly depends on it. The participants were first explained the difference between multichannel audio, stereo and mono formats and then were given an overview of the MUSHRA test. The concepts and definitions of BAQ and QoE were clearly explained. Different attributes to look out for like timbral quality, localization quality and artifacts that could have been introduced due to down-mixing, all of which are mentioned in the ITU-R BS.1534-3 [ITU (2012)] were both explained and demonstrated.

4.2 Stimuli

Ten Audio excerpts were used for the test with range of duration five to ten seconds, mean duration was 6.6 seconds with standard deviation 1.84. Two of these ten excerpts were used for training and the remaining eight were used for the actual test. All excerpts originally had either 6 (5.1) or 8 (7.1) channels.

The two training audio excerpts had the following content:

1) **[Bee (7.1)]** sounds a bee moving from one end of the sound scene to another in a forest.

2) **[New Horizon 2 (5.1)]** excerpt from the song “New Horizon”.

The eight remaining audio excerpts had the following content:

1) **[Ashamed (5.1)]** excerpt from the song “Ashamed”

2) **[DJ Set 1 (5.1)]** excerpt from DJ Set music

3) **[DJ Set 2 (5.1)]** excerpt from DJ Set music

4) **[Forest (7.1)]** sounds of birds and insects in a forest.

5) **[Moving Bird (7.1)]** sound of a bird flapping its wings and moving from one end of the sound scene to the other

6) **[New Horizon 1 (5.1)]** excerpt from another part of the song “New Horizon”

7) **[Step Outside (5.1)]** excerpt from the song “Step Outside”

8) **[Wolf (5.1)]** excerpt from the song "Wolf"

The original excerpts were down-mixed into three versions: Stereo 1, Stereo 2 and Mono. For the rest of this paper these versions will be referred to as "conditions". For excerpts 2, 3,4 and 5, Stereo 1 was mixed according to the ITU-R BS.775 [ITU (2015b)] in which the left and right surround channels were multiplied with 0.707 and the center channel was multiplied by 1. This corresponds to giving the left and right channels (both front and back) a gain of -3 dB. For excerpts 1, 6, 7 and 8, stereo optimized versions were already available and were used instead of a down-mixed version. Stereo 2 was mixed by multiplying the left and right channels (both front and back) by 0.3 and the center channel by 1. This corresponds to giving the left and right channels a gain of -6 dB. The Mono version was mixed by combining both channels of Stereo 1. The ITU-R BS.1534-3 [ITU (2012)] requires the reference to be low pass filtered at 7 KHz and at 3.5 KHz to serve as the mid and low anchors respectively. Given the nature of the test and the material used, these anchors did not seem suitable so Stereo 2 served as the mid anchor and Mono served as the low anchor. An added benefit of Stereo 2 was also that it made the test a little more challenging as the difference between Stereo 1 and Mono would have easy to detect. Every condition of each excerpt was loudness aligned and was compliant with the standard.

4.3 Apparatus

The test was conducted with headphones in a soundproof room with background noise of NR equal to 10 that met the requirements of the ITU-R BS.1534-3 [ITU (2012)]. The Samsung Galaxy S8 phone [Samsung (2020a)] with an octa-core (4×3 GHz Mongoose M2 & 4×1.7 GHz cortex-A53) processor and 4 GB RAM, running Android 8.0 with all the latest updates was used along with the VLC player application (version 3.0.13) for Android for the payback of the audio excerpts. The first pair of headphones was the AKG EO-IG955 earbuds [Samsung (2020b)], which are tuned for Samsung were used which have a frequency response is 20 Hz-20 KHz, sensitivity is 93.2 dB, 32 ohm impedance and has a standard 3.5 mm headphone connector. For the rest of the paper they will referred to as "Headphones 1". The second pair of headphones was the Sennheiser Urbanite XL [Sennheiser Urbanite-XL (2020)]. These were over-ear (circum-aural) headphones with a frequency response of 16 Hz-22 KHz, 18 ohm impedance, and sensitivity of 110 dB. For the rest of the paper they will be referred to as "Headphones 2". Headphones 1 came standard with the phone and Headphones 2 were chosen to introduce a variety of sound in the tests due their different frequency response.

Eight playlists on the VLC player were made for the listening tests. The VLC player [VLC 3.0 Vetinari (2020)] was chosen because of its widespread popularity amongst users and for its support for a wide range of audio codecs and formats. Each playlist, consisted of the reference file, a hidden reference and the three conditions mentioned in Section 3.2. Each condition was numbered and the participants could identify the reference file but not any of the conditions. It was easy to switch between and play each condition. Every condition was numbered randomly in each playlist. Fig. 1 shows a screenshot of one of the playlists. In Fig. 1, the excerpts DJSet1 1-4 could be any of the four conditions and the participants had no way of identifying the condition based on the number assigned in the list. A similar playlist was made for each audio excerpt mentioned in Section 4.2.

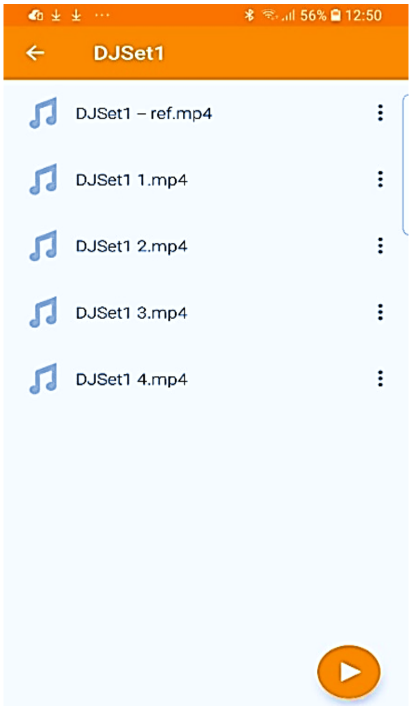


Figure 1: The GUI used for the tests. The participants could easily switch and play different audio excerpts in each playlist

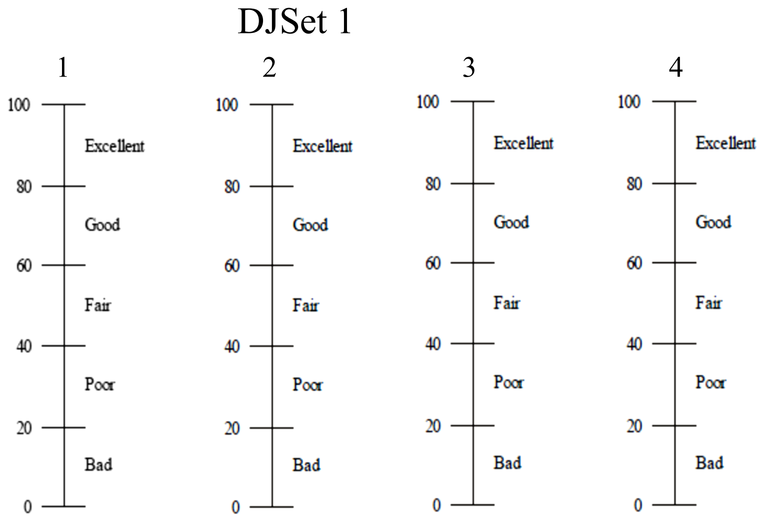


Figure 2: A scoresheet used for the tests. The conditions: Hidden Reference, Stereo 1, Stereo 2 and Mono were randomly assigned 1, 2, 3 or 4 and the participant did not know which was which

4.4 Experiment 1

In accordance with the MUSHRA recommendation [ITU (2012)], the test in which the subjects were to rate the basic audio quality (BAQ) was taken first. The standard defines basic audio quality as "the single, global attribute used to judge any and all detected differences between the reference and the object". The training session included a demonstration that showed the difference between the reference and conditions particularly in terms of timbral and spatial quality. As practice, the participants were asked to rate the second training file for perceived BAQ. After the training, they were given the main eight playlists and began the test. The original multichannel audio excerpt was used as the known and hidden reference. A score sheet was given for each stimulus with a "pen and paper scale" mentioned in the ITU-R BS.1534-3 [ITU (2012)] for each condition. Each scale was 10 cm long and the participants were asked to draw a small line on the scale which marked the perceived basic audio quality which was later assigned an integer number by interpolating between two of the marked intervals, as recommended in the standard. A sample of the scoresheet is shown in Fig. 2. The experiment was first conducted with Headphones 1 and then with Headphones 2 with a gap of at least 24 hours in between.

4.5 Experiment 2

The ITU-R BS.1534-3 recommends at least 24 hours, preferably 48 hours between listening tests for different attributes. Accordingly, the test for Quality of Experience (QoE) took place 24 hours after the BAQ test. The ITU-T P.10/G.100 [ITU (2017)] has a working definition of QoE as "the degree of delight or annoyance of the user of an application or service". The training session was repeated with emphasis on the definition of Quality of Experience. The same playlists with the same excerpts and conditions were used but this time the known reference was excluded and the participants were asked to rate the quality of experience for all four conditions in the playlist. The experiment was first conducted with Headphones 1 and then with Headphones 2 with a gap of at least twenty-four hours in between.

In accordance with the standard, only one attribute was rated per test. It was made sure that there was at least a 24-hour gap between experiment 1 and 2. Therefore, each subject performed four listening tests spread out over at least four days.

5 Results

The results of Experiment 1 and 2 with Headphones 1 and 2 are shown in Figs. 3-6. The graphs show the means and 95% confidence internals against the MUSHRA score on a scale of 1 to 100. Since the score of such listening tests do not usually make a normal distribution, the t-Test method was used for the calculation of upper and lower bound of error ("t-confidence intervals") [Moore and McCabe (1998)].

For Experiment 1 (BAQ) with Headphones 1, the trend is obvious; the multichannel audio excerpts being the reference scored the highest by a margin of at least 20. Stereo 1 came in second with Stereo 2 a close third and Mono as fourth.

Experiment 1 with Headphones 2 showed a slightly different trend. The multichannel condition still scores the highest for all items except for 'moving bird'. In the consolidated plot "All" multichannel clearly outscores the other conditions followed by Stereo 1, Mono and Stereo 2 as a close fourth.

For Experiment 2 (QoE) with Headphones 1, the trend is similar but this time the multichannel audio excerpts do not win with a large margin. The gap between Mono and Stereo 2 has also reduced. It appears that the listeners recognized the difference in BAQ for all conditions but QoE score for the multichannel audio excerpts was only 6 points higher than that of the Stereo 1 (consolidated plot "ALL"), which was the superior of the two stereo mixes.

Experiment 2 with Headphones 2, the multichannel condition outscores the rest in all items except "moving bird" and "New Horizon". In the consolidated plot "All", it follows the same trend that we saw with Headphones 1 except that mono out performed Stereo 2.

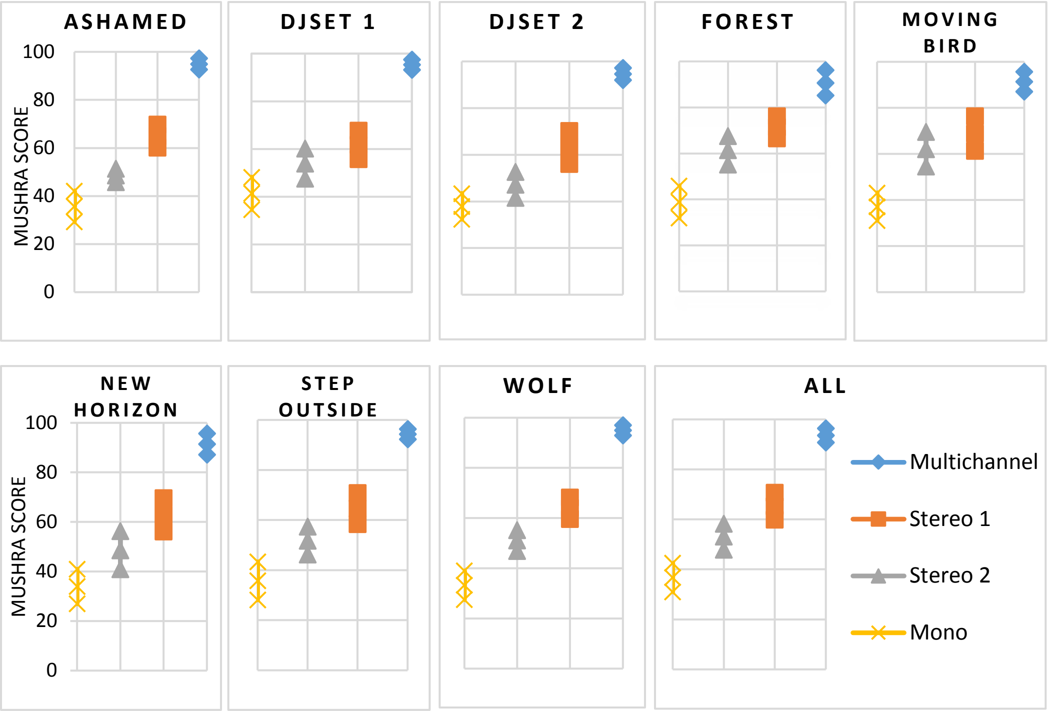


Figure 3: Results of Experiment 1 using Headphones 1 based on the 95% confidence intervals. The x-axis shows different conditions of the audio excerpts and the y-axis shows the MUSHRA score for Basic Audio Quality

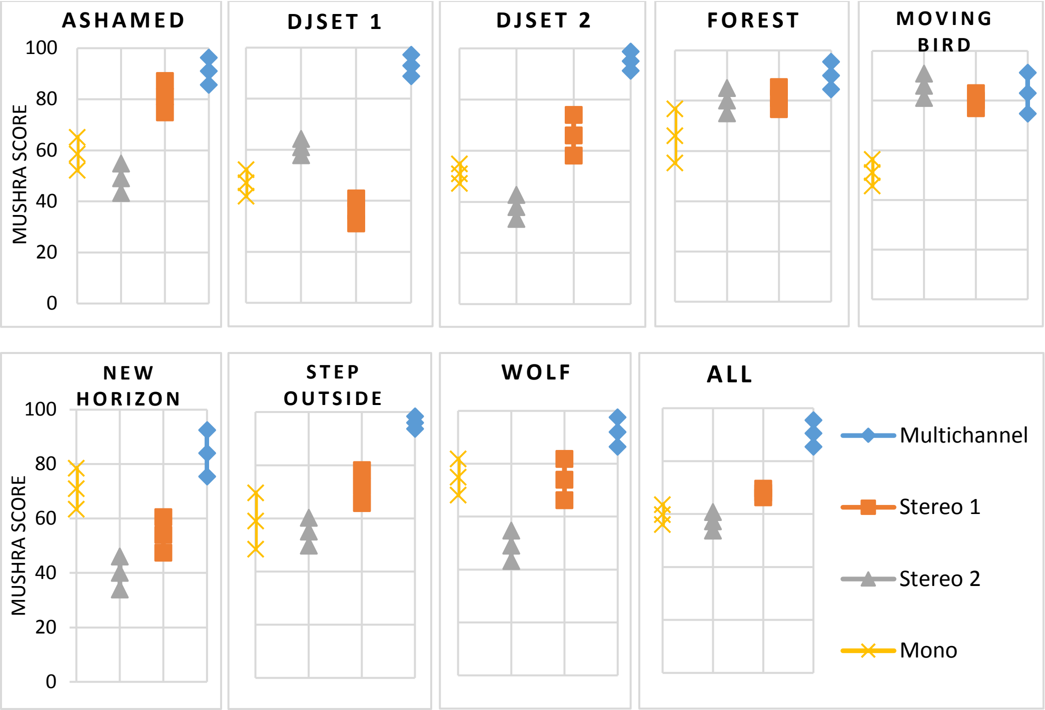
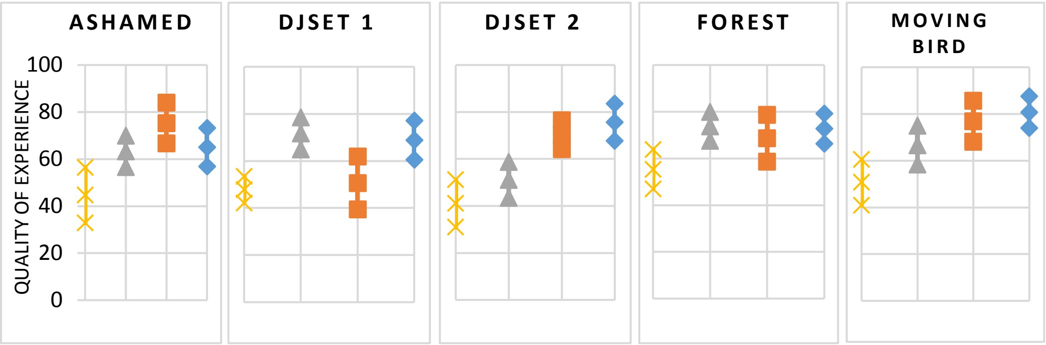


Figure 4: Results of Experiment 1 using Headphones 2 based on the 95% confidence intervals. The x-axis shows different conditions of the audio excerpts and the y-axis shows the MUSHRA score for Basic Audio Quality



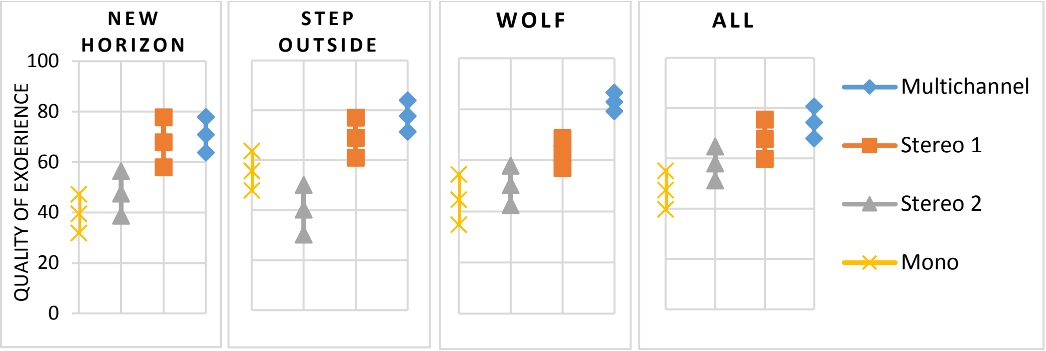


Figure 5: Results of Experiment 2 using Headphones 1 based on the 95% confidence intervals. The x-axis shows different conditions of the audio excerpts and the y-axis shows the score for Quality of Experience

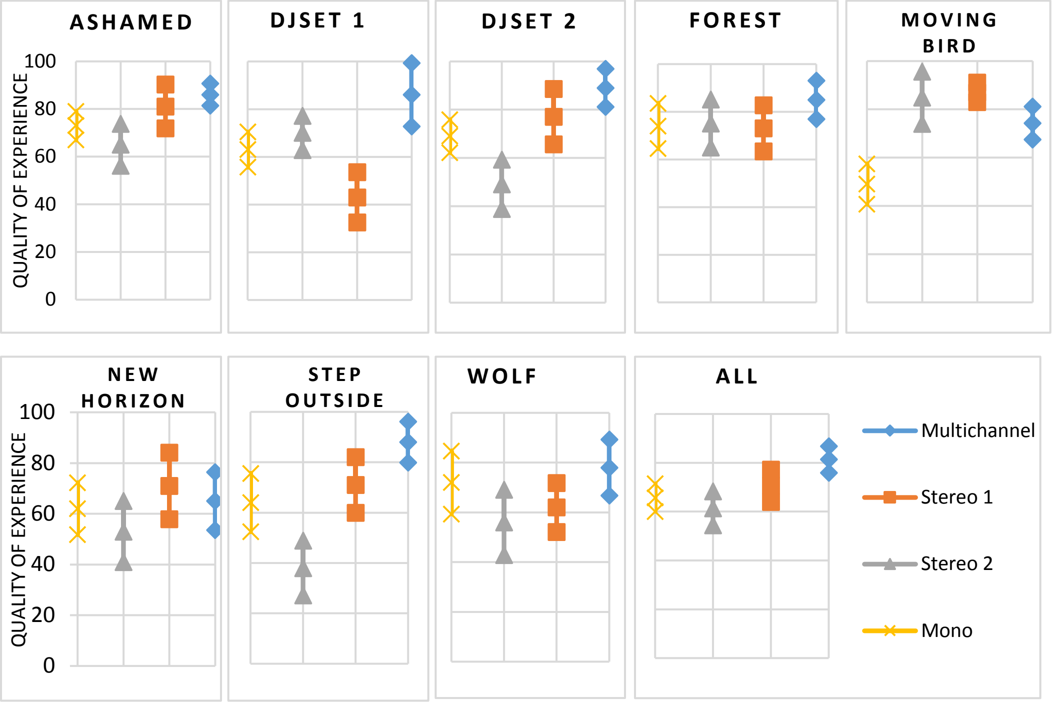


Figure 6: Results of Experiment 2 using Headphones 2 based on the 95% confidence intervals. The x-axis shows different conditions of the audio excerpts and the y-axis shows the score for Quality of Experience

6 Statistical analysis

The graphs in Figs. 3-6 show a certain trend but to gain more insights and to interpret the results we require further statistical analysis. The ITU-R BS.1534-3 [ITU (2012)] recommends parametric statistics like analysis of variance (ANOVA) for analysis of the results of the tests. The ANOVA test has some underlying assumptions. Only if the data shows severe departure from these assumptions, should non-parametric methods be used for the primary analysis. Some of these assumptions are that the data is correct and as expected, the data is unidimensional, collected independently by all assessors on a common scale, free from outliers, of normal distribution and of homogenous variance. The scores from these experiments comply with these assumptions except for distribution and variance (data from MUSHRA based tests rarely do), the ANOVA F-statistic is robust for distributions that are non-normal and that are heterogeneous in variance [Gene, Glass and Peckham (1972)]. Moreover, the data has three categorical variables (the audio conditions), making the ANOVA analysis even more suitable for these tests.

Effect size is the size of the difference between two variables [Moore and McCabe (1998)]. It is a quantitative measure of the significance of a phenomenon. The results of these listening tests show how the participants rated different conditions of eight audio excerpts, the 'items'. It is relevant to see the size of the effect each item had on the overall result. Cohen's d [Cohen (1988)] was chosen, as it is an appropriate effect size for the comparison between two means. It indicates the standardized difference between two means, and expresses this difference in standard deviation units. The next two sub-sections show the ANOVA and Cohen's d calculations of the results and discuss their implications.

6.1 ANOVA

A one way ANOVA was applied to the consolidated scores of the hidden reference and all the other three conditions (Stereo 1, Stereo 2 and Mono) for each of the experiments. Since both experiments were rating a different attribute, two different null hypotheses are defined. For Experiment 1, the defined null hypothesis is "Multichannel audio will give no improvement in perceived audio quality as rated by a user". For Experiment 2, the null hypothesis is "Multichannel Audio will show no improvement in quality of experience as rated as a user".

Tabs. 1 to 4 show the ANOVA tables for Experiments 1 and 2 using Headphones 1 and 2. In the tables, the conditions Hidden Reference, Stereo 1, Stereo 2 and Mono are abbreviated as HR, S1, S2 and M respectively. The value of α for all statistical calculations in this paper is 0.05. Each individual table shows the calculations of three one-way ANOVA's.

In each ANOVA, the F statistic is higher than the F-critical value and in each case the value of P if extremely low. So the null hypothesis can be rejected for both Experiment 1 and 2 (for both headphones). However, the difference in F and F-critical is much less in Experiment 2 and the P-values though still small, are larger than those of Experiment 1. This is because while rating BAQ, Multichannel took a larger lead in scores as compared to QoE.

Table 1: ANOVA Table for BAQ Using Headphones 1

HR & S1	*SS*	*df*	*MS*	*F*	*p-value*	*F crit*
Between Groups	5607.6	1	5607.6	135.8	8E-12	4.098
Within Groups	1073.6	38	41.292			
HR & S2						
Between Groups	11451	1	11451	409.3	2E-17	4.098
Within Groups	727.4	38	27.977			
HR & M						
Between Groups	22572	1	22572	504.2	1.5E-18	4.098
Within Groups	1163.9	38	44.765			

Table 2: ANOVA Table for BAQ Using Headphones 2

HR & S1	*SS*	*df*	*MS*	*F*	*p-value*	*F crit*
Between Groups	5062.5	1	5062.5	79.555	7.4E-11	4.098
Within Groups	2418.13	38	63.635			
HR & S2						
Between Groups	11055.6	1	11056	129.27	8.6E-14	4.098
Within Groups	3250	38	85.526			
HR & M						
Between Groups	9409.56	1	9409.6	104.9	1.7E-12	4.098
Within Groups	3408.7	38	89.703			

Table 3: ANOVA Table for QoE Using Headphones 1

HR & S1	*SS*	*df*	*MS*	*F*	*p-value*	*F crit*
Between Groups	307.23	1	307.234	6.995	0.0137	4.098
Within Groups	1141.9	38	43.9209			
HR & S2						
Between Groups	1854.5	1	1854.54	29.24	1E-05	4.098
Within Groups	1648.9	38	63.42			
HR & M						
Between Groups	5042.4	1	5042.43	44.66	4E-07	4.098
Within Groups	2935.4	38	112.9			

Table 4: ANOVA Table for QoE Using Headphones 2

HR & S1	*SS*	*df*	*MS*	*F*	*p-value*	*F crit*
Between Groups	1102.5	1	1102.5	10.48	0.00251	4.098
Within Groups	3998.1	38	105.21			
HR & S2						
Between Groups	3802.5	1	3802.5	24.52	1.5E-05	4.098
Within Groups	5893.8	38	155.1			
HR & M						
Between Groups	2363.9	1	2363.9	18.13	0.00013	4.098
Within Groups	4955.6	38	130.41			

6.2 Effect size

To calculate effect size, the Cohen’s d was used with the necessary correction applied due to the relatively small sample size. As suggested by Cohen and then updated by Sawilowsky [Sawilowsky (2009)] a d value of 0.01 is very small, 0.2 is small, 0.5 is medium, 0.8 is large, 1.2 is very large and 2.0 onwards is huge.

In Experiment 1 using Headphones 1 & 2, items “Ashamed”, “Forest” and “Moving Bird” have a large effect size for the condition Stereo 2. With Headphones 2, the conditions Stereo 1, Stereo 2 and Mono show large to huge effect sizes for many items shown in Tab. 5.

In Experiment 2 with Headphones 1 the item “Wolf” has a medium to very large effect size for conditions Multichannel, Stereo 1 & 2. The items “DJSet 1” and “Step Outside” show large effect size for Stereo 1 and very large for Stereo 2 respectively. With headphones 2 the item “DJ Set1” and “Moving Bird” show a very large effect size for Stereo 1 and “Step Outside” shows a large effect size for Stereo 2. All excerpts (across all experiments) that showed a very large or huge effect size are either of the condition Stereo 1 or 2. This could be due to artifacts introduced by the ITU-R BS.775-3 [ITU (2015b)] method of down mixing that was used for these two conditions. The effects of such down mixing have been explored in Zielinski et al. [Zielinski and Rumsey (2003)].

Table 5: Cohen's d calculated for each condition of each excerpt in Experiments 1 & 2. HR: Hidden Reference, S1: Stereo 1, S2: Stereo 2, M: Mono

		Experiment 1		Experiment 2	
		(BAQ)		(QoE)	
		Headphones		Headphones	
		1	2	1	2
Ashamed	HR	0.32	0.04	0.77	0.4
	S1	0.03	0.99	0.1	0.61
	S2	0.82	0.8	0.5	0.22
	M	0.1	0.1	0.2	0.58
DJSet 1	HR	0.4	0.3	0.5	0.21
	S1	0.35	4.1	1.1	1.47
	S2	0.09	0.5	1.2	0.6
	M	0.43	1.3	0.02	0.2
DJSet 2	HR	0.2	0.48	0.1	0.5
	S1	0.2	0.16	0.3	0.3
	S2	0.7	2.17	0.6	0.7
	M	0.1	1.1	0.4	0.2
Forest	HR	0.3	0.4	0.2	0.2
	S1	0.7	1.83	0.1	0.1
	S2	0.9	2.5	1.5	0.7
	M	0.2	0.4	0.5	0.5
Moving Bird	HR	0.3	0.51	0.62	0.5
	S1	0.3	2.3	0.7	1.4
	S2	0.8	3.16	0.7	1.2
	M	0.04	0.9	0.2	1.1
New Horizon	HR	0.3	0.44	0.32	0.86
	S1	0.2	1.38	0.01	0.02
	S2	0.4	1.63	0.8	1.19
	M	0.3	0.88	0.6	0.09
Step Outside	HR	0.2	0.66	0.36	0.44
	S1	0.1	0.41	0.12	0.02
	S2	0.15	0.2	1.2	1.19
	M	0.1	0.05	0.6	0.09
Wolf	HR	0.3	0.13	1.2	0.2
	S1	0.2	0.5	0.5	0.5
	S2	0.3	0.81	0.6	0.24
	M	0.4	1.3	0.17	0.3

7 Discussion

There are four obvious conclusions one can draw from the results and analysis of the two experiments:

1. Multichannel audio performs better than stereo or mono in terms of perceived basic audio quality and user experience. With different headphones slightly different scores are observed but a similar trend is followed.

2. The increase in BAQ score from Stereo 1 to Multichannel is significantly larger than the increase in QoE score for the same.

3. Perception of spatial effect changes with the change of headphones which is evident from the fact that mono scored higher than stereo 2 in BAQ and QoE with Headphones 2.

4. Audio content (the items) has very little effect on the overall scores but the conditions (particularly Stereo 1 & 2) seem to have a significant effect.

The first conclusion was already expected, given the design of the MUSHRA test and the definition of BAQ. In Experiment 1, multichannel audio scored the highest with a large margin despite the fact that most of the participants in the test were not expert listeners. A binaural rendered multichannel audio excerpt performs better than one mixed into stereo and mono. The second conclusion is an important find in this study. With BAQ as the attribute, the MUSHRA becomes more of a technical evaluation. Multichannel audio being the reference was bound to have an edge over the other conditions. When the participants were asked to evaluate QoE the variance, particularly the gap between the scores of Stereo 1 and Multichannel audio reduced. One can infer that a typical user while listening to music on headphones does appreciate the improved quality and spatial effects of multichannel audio but not much more than that of a properly mixed stereo version of the same audio clip. Schoeffler et al. [Schoeffler, Silzle and Herre (2017)] got similar results while evaluating 3D audio on a loudspeaker system.

The third conclusion was probably due to the different frequency responses of both headphones and the fact that Headphones 1 was tuned for the mobile phone that was used. Most participants commented later that Headphones 2 “had more bass” making it more difficult to tell the difference between each condition. The fourth conclusion though described in Section 6 requires further analysis. More experiments using a variety of content for audio items under test can provide further insights.

8 Conclusion

This paper showed and analyzed the results of subjective quality tests of multichannel audio as compared to two different stereo down-mixes and mono while listened to on a mobile phone through two different headphones. A different test was conducted for both BAQ and QoE. The perceived basic audio quality of multichannel audio was rated higher than both stereo mixes and mono. The ratings of quality of experience followed a similar trend but with less gap between the scores of each format. The gap between stereo and multichannel audio was much larger in the BAQ ratings as compared to the QoE ratings. The same tests with a different pair of headphones gave slightly different scores but showed a similar trend. The results of these experiments certainly build a case for the implementation of multichannel audio content to be played or streamed on handheld electronic devices like mobile phones. Future work could include objective assessment of multichannel audio and correlating its results to the results in this paper.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

Alexander, R. C. (2008): *The Inventor of Stereo: the Life and Works of Alan Dower Blumlein*. CRC Press. Abingdon, UK.

Chabanne, C.; McCallus, M.; Robinson, C.; Tsingos, N. (2010): Surround sound with height in games using Dolby Prologic IIz. *Proceedings of 129th AES Convention*, San Francisco. USA.

Cohen, Jacob (1988): *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates. 2nd Edition.

Dolby Website (2020) https://www.dolby.com/us/en/categories/smartphone.html.
Last accessed: May 2020.

Gene, J. R.; Glass, S. V.; Peckham, P. D. (1972): Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research*, vol. 42, no. 3, pp. 237-288.

Hamasaki, K.; Matsui, K. K.; Sawaya, I.; Okubo, H. (2011): The 22.2 multichannel sounds and its reproduction at home and personal environment. *Proceedings of AES 43rd International Conference Audio for Wirelessly Networked Personal Devices*, Pohang, Korea, Paper 3-1.

ITU (2012): Method for subjective assessment of intermediate quality level of audio systems (MUSHRA). *ITU Recommendation ITU-R BS.1534-3*. Geneva, Switzerland.

ITU (2015a): Methods for the subjective assessment of small impairments in audio systems. *ITU Recommendation ITU-R BS.1116-3*. Geneva, Switzerland.

ITU (2015b): Multichannel stereophonic sound system with and without accompanying picture. *ITU Recommendation ITU-R BS.775-3*. Geneva, Switzerland.

ITU (2017): Vocabulary for performance, quality of service and quality of experience. *ITU Recommendation ITU-T P.10/G.100*. Geneva, Switzerland.

ITU (2018a): Method for objective measurements of perceived audio quality. *ITU Recommendation BS.1387-3*, Geneva, Switzerland.

ITU (2018b): Perceptual objective listening quality prediction. *ITU Recommendation P.863*, Geneva, Switzerland.

Likert, Rensis (1932): A technique for the measurement of attitudes. *Archives of Psychology*: 1-55.

Moore, D.; McCabe, G. P. (1998): *Introduction to the Practice of Statistics*. W.H. Freeman.

Roginska, A.; Geluso, P. (2017): *Immersive Audio*, Routledge*,* 1st edition.

Rumsey, F. (2001): *Spatial Audio*. Focal Press. Abingdon, UK.

Samsung Website (2020a): https://www.samsung.com/global/galaxy/galaxy-s8/specs/ Last accessed: May 2020.

Samsung Website (2020b): https://www.samsung.com/latin_en/mobile-accessories/earphones-tuned-by-akg-eo-ig955/EO-IG955BREGWW/. Last accessed: May 2020.

Sawilowsky. S. (2009): New effect size rules of thumb. *Journal of Modern Applied Statistical Methods*, vol. 8, no. 2, pp. 467474.

Schoeffler, M.; Silzle, A.; Herre, J. (2017): Evaluation of spatial/3D Audio: basic audio quality versus quality of experience. *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 1, pp. 75-88.

Sennheiser Website (2020): https://en-uk.sennheiser.com/urbanite-xl. Last accessed: May 2020.

Toosy, F.; Ehsan, M. S. (2019a): BAQ and QoE: subjective assessment of 3D audio on mobile phones. *146th AES Convention*, Dublin, Ireland.

Toosy, F.; Ehsan, M. S. (2019b): Subjective evaluation of multichannel audio and stereo on cell phones. *147th AES Convention*, New York, USA.

Toosy, F.; Ehsan, M. S. (2019c): Evaluation of multichannel audio in automobiles versus mobile phones. *147th AES Convention*, New York, USA.

VLC Website (2020): https://www.videolan.org/vlc/releases/3.0.0.html. Last accessed: May 2020.