Audio Engineering Society

Convention Paper 5562

Presented at the 112th Convention
2002 May 10–13 Munich, Germany

Subjective audio quality trade-offs in consumer multichannel audio-visual delivery systems.

Part I: Effects of high frequency limitation.

Slawomir K. Zielinski[1], Francis Rumsey[1], and Søren Bech [2]

[1] *Department of Sound Recording, University of Surrey, Guildford, Surrey, GU2 7XH, UK* (s.zielinski@surrey.ac.uk, f.rumsey@surrey.ac.uk)

[2] *Bang & Olufsen, Struer, Denmark* (SBE@bang-olufsen.dk)

ABSTRACT

The subjective effects of controlled high frequency limitation of the audio bandwidth on assessment of audio quality were studied. The investigation was focused on the standard 5.1 multichannel audio set-up (Rec. ITU-R BS.775-1) and limited to the optimum listening position. The effect of video presence on audio quality assessment was also investigated. The results of the formal subjective test indicate that it is possible to limit the high frequency content of the centre or of the rear channels without significant deterioration of the audio quality for some of the investigated programme material types. Video presence has small effect on audio quality assessment.

INTRODUCTION

Nowadays, it is possible to distinguish between two trends in development of audio applications. The first one aims at achieving the highest possible audio quality (for example the latest high-resolution audio applications), whereas the objective of the second one is to reduce the cost of equipment manufacturing, cost of audio broadcasting or media storage, resulting in some inevitable degradation of audio quality (one can consider audio broadcasting over the Internet, manufacturing of cheap home cinema systems, etc.). In order to achieve the best trade-offs between cost and audio quality it is necessary to optimise systems psycho-acoustically on the basis of formal subjective tests, which in general is a complicated task (the audio quality depends on many factors like: bandwidth, dynamic range, distortions, spatial characteristics, programme

material, etc.). The objective of the experiment described in this paper was to study only the effects of controlled limitation of high frequencies (HF) on subjectively perceived audio quality in a standard 5.1 multichannel audio set-up [1]. This is a companion paper to the paper describing the effects of low frequency band-limitation submitted for the AES 22nd International Conference [2]. The experiment described here was carried out in collaboration between University of Surrey (Dept. of Sound Recording), British Broadcasting Corporation (BBC) and Bang & Olufsen within a joint EPSRC-funded project investigating **subjective quality trade-offs in consumer multichannel sound and video delivery systems**.

The main research questions in this experiment were as follows:

1. What is the quantitative relationship between the high frequency limitation and audio quality?
2. How multichannel audio material may be band-limited with minimum overall subjective effect?
3. Does video presence have any effect on audio quality evaluation?

In order to answer these questions the formal listening test was carried out.

1 SELECTION OF PROGRAMME MATERIAL

The main and the most obvious criterion of selection of programme material was to choose the **most generic types** of material that are currently used and/or will be used in the future. Therefore it was decided to choose excerpts representing categories like classical music, pop music, movie and TV sport. A special excerpt with surround applause having pronounced HF content was also included in our selection.

Since surround audio material may be varied in its spatial content, it was considered to choose additional criterion of programme material selection based on **microphone and/or panning techniques** used during recording. However, there were two problems related to this criterion. Firstly, it would increase significantly the number of excerpts used in the experiment making the listening test longer and more complicated. For example, it would be necessary to select excerpts representing at least 5 basic types of multichannel microphone techniques and 5 types of multichannel panning techniques (detailed discussion on different multichannel microphone/panning techniques can be found in [3]). Secondly, the detailed information about microphone/panning of some recordings is not always easily accessible. Therefore, in order to simplify the way of programme selection, it was decided to use a criterion based on **audio scene-based paradigm** [4]. In this approach it is possible to assess the spatial characteristic of the recording just by listening to the audio content in particular loudspeakers (there is no need to analyse what microphone/panning techniques were used during recording). In other words, the actual recording is a subject of analysis, not the way in which this recording was made. The subjectively perceived spatial characteristic can be expressed in terms of audio scenes. Basically, it is possible to distinguish between 4 audio scene categories representing spatial characteristics of typical multichannel audio recordings, named *F-B*, *F-F*, *B-F* and *B-B* respectively (see Fig. 1). The first scene (***F-B***) describes the case where front channels reproduce ***F****oreground* audio content (close and clearly perceived audio sources), whereas rear channels contain ***B****ackground* audio content (reverberant sounds, not clear, "foggy", quieter than the front ones). This situation may be compared to the typical sound impression perceived by the listener sitting in the concert hall (sound stage in the front, reflections from side and back). The second audio scene (***F-F***) describes a recording in which the listener is surrounded by clearly identifiable audio sources (foreground audio content both from front and rear directions). This refers to the audio impression when the listener is surrounded by the orchestra. The third possible scene (***B-F***) represents the recordings having ***B****ackground* content in the front and ***F****oreground* content in the rear. Although this case is not common, there are some classical music excerpts with instruments in the rear and only reverberation in the front. The fourth spatial audio scene (***B-B***) describes the recording with ***B****ackground* content in all channels (for example the noise of the audience leaving the concert hall recorded using the array of distant microphones). In the authors' opinion first two spatial characteristics (*F-B* and *F-F*) are the most typical ones and therefore they were selected for this experiment (three items with *F-B* characteristic and three items with *F-F* characteristic).

Another important criterion was to select **"critical" material** (that is revealing differences of the system under test), which in our case meant material with pronounced HF content. To achieve this objective a software audio sampler was developed allowing for instantaneous access to 120 multichannel audio excerpts (see Fig. 2). Thus it was possible to make AB comparisons between different items to find out a "short list" of excerpts having the most pronounced HF content. The potential problems with mid- and long-term auditory memory were overcome due to possibility of making quick AB comparisons. Moreover, during the selection of the "short list" of "critical" items it was found useful to listen to both original and low-pass filtered versions of the selected items to check to what degree limitation of HF content would deteriorate their quality. It was accomplished by low-pass filtration of the items in real-time using built-in filters in a digital mixing console. In order to switch the filters on or off a remote MIDI controller was installed in the listening room. The reason of using the MIDI controller to switch on/off the filters instead of a direct use of the mixing console was that the latter could deteriorate the acoustic conditions if it was installed in the listening room due to its relatively large dimensions (problem with reflections). This way a "short list" of suitable excerpts was created and auditioned by the authors of this paper. Then, after discussion, the final group of excerpts was selected. Finally, the decision was verified by comparing results of objective analyses (to be described later).

Another important criterion of selection of the material was **consistency** of its characteristics. Long items having variable spectral and spatial characteristics are difficult to assess. Therefore it was decided to use relatively short (from 5 to 18 sec.), looped items with possibly time-invariant characteristic. The exception was the 'TV Sport' item in which case it was impossible to select an excerpt with very consistent characteristics. Special attention was paid to create artistically "correct" loops both from audio and visual point of view.

The last criterion of selection of programme material was **suitable audio-visual correlation**. It was preferred to use material with strong correlation between audio and visual cues. For example for music type of programme material the items with picture containing playing musicians were preferred as opposed to items with picture containing only hall interiors, landscapes or other objects not related directly to the audio content.

Finally, taking into account all the mentioned previously criteria, six items were selected for this experiment (Tab. 1). The first item ('Classical music') was a typical orchestra music recording with pronounced violin and cello sections. The spatial characteristic can be classified as the *F-B* scene (instruments in the front channels with reverberations in the rear ones). This item was not the most "critical" in the sense of HF content, but was selected because of high quality of recording and because of suitable video content (the accompanying picture contained a view of playing musicians). The low frequency effect (LFE) channel was not used in this item.

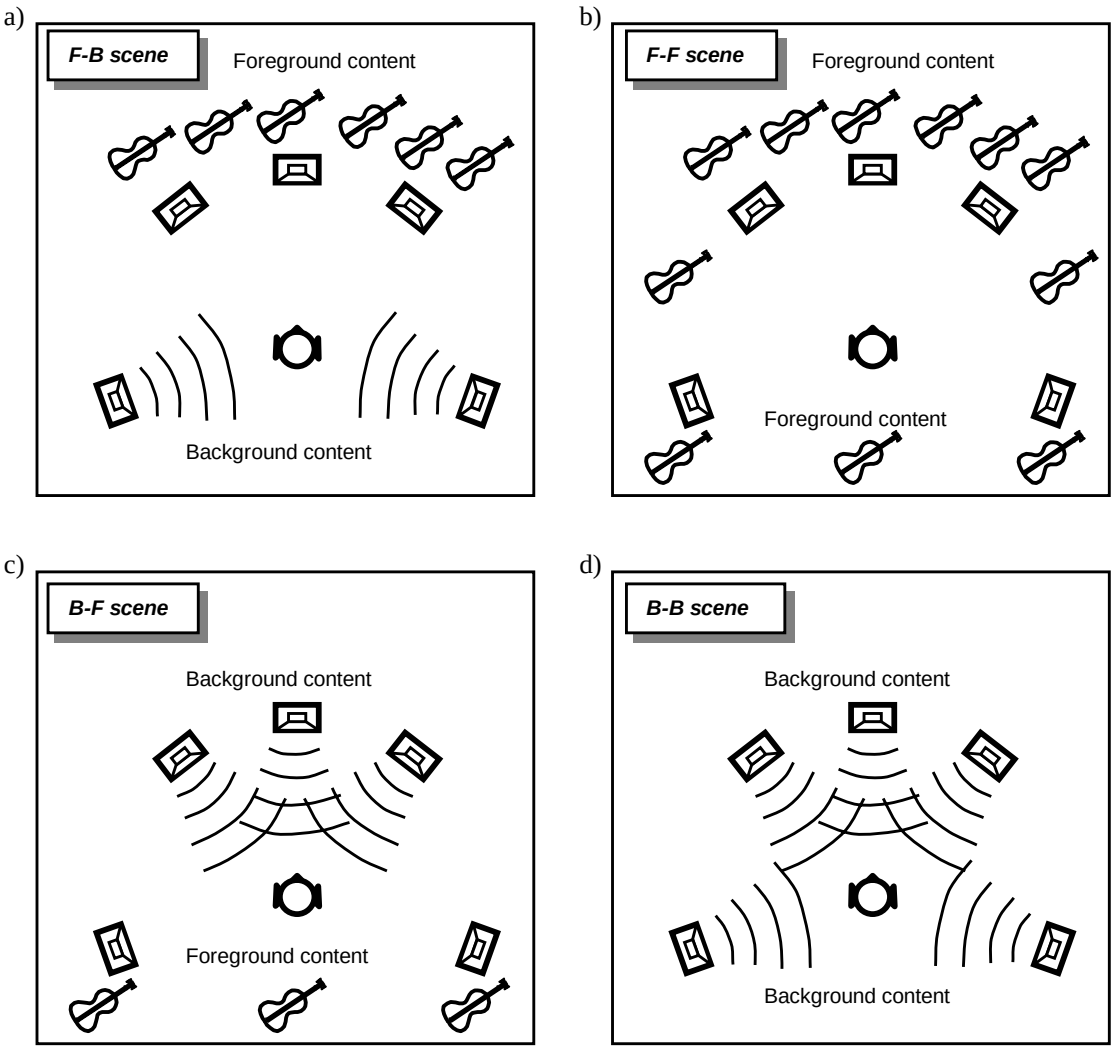


Fig. 1 Basic audio scenes representing spatial characteristics of multichannel audio recordings:
a) *F-B* scene; b) *F-F* scene; c) *B-F* scene; d) *B-B* scene.

Item No.	Programme type	Spatial characte -ristic	Du- ration
1	Classical music	*F-B*	5 sec.
2	Pop music	*F-B*	18 sec.
3	Pop music	*F-F*	18 sec.
4	Movie	*F-B*	8 sec.
5	TV Sport	*F-F*	6 sec.
6	Applause	*F-F*	13 sec.

Tab. 1 Audio-visual material selected for the experiment

Both pop music items selected for our experiment were recorded live. In the first case most of the instruments were balanced to front channels with reverb in rear channels (*F-B* spatial characteristic) whereas in the second case the instruments were mixed to all channels including percussion instruments in the rear channels (*F-F* characteristic). The LFE channel was used in both items. The accompanying picture contained a view of performing musicians on the stage.

The movie item was an excerpt containing a dialogue in the centre channel (picture contained a group of talking people). Front channels contained some special audio effects. There was also orchestral music spread around all loudspeakers. Since front loudspeakers were much louder than the rear ones, the spatial characteristic of this item can be classified as an *F-B* one. There was no LFE channel included in this item.

The sport item was a BBC recording of tennis from Wimbledon. The chosen excerpt contained crowd effects (applause) in all channels (*F-F* spatial characteristic). There was a commentary between front left and centre channels and umpire's voice between centre and front right channel. Details about this recording are described in [5]. There was no LFE channel included in this excerpt.

The last selected item was the applause. Its unique spectral characteristic (averaged over the duration of this item) was similar to the characteristic of the pink noise. Since all channels contained the

applause, the spatial characteristic of this item can be classified as the *F-F* one. The picture contained the view of the audience and the musicians. This item did not contain the LFE channel.

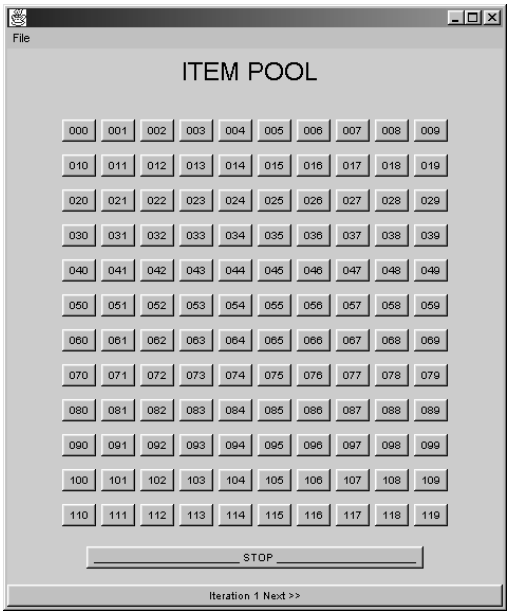


Fig. 2 "Item Pool" – a software multichannel audio sampler for critical selection of programme material

In order to quantify the level differences between channels and to check whether "truly critical" material (containing HF content) was selected two types of objective analyses were made. The first analysis was just to compare the RMS level of the audio signal (averaged over the duration of each item) between each channel (see Tab. A1). The second type of analysis was to make inter-item and intra-item (between channels) comparison of HF content. Since it was difficult to compare visually spectrograms and obtain quantitative information about differences in HF content it was decided to use a special spectral coefficient "k_{HF}" defined as the energy of the signal for frequencies higher than 10 kHz ($E_{f>10\ kHz}$) normalised to the total energy of the signal (E_{Tot}):

$$k_{HF} = \frac{E_{f>10\,kHz}}{E_{Tot}} \tag{1}$$

The frequency of 10 kHz used in definition of this coefficient is related to the highest cut-off frequency employed in the experiment (an issue of selection of cut-off frequencies will be discussed in the next section). When the coefficient k_{HF} is equal to unity (0 dB) it means that the whole energy of the signal is concentrated at high frequencies (from 10 to 20 kHz). The HF content of each item is presented in Table A2 in the appendix. According to this table the most "critical" item selected for the experiment was a pop-music excerpt with *F-B* spatial characteristic (HF coefficient ranging from -18 dB to -11 dB) whereas the less "critical" item was a classical music item (HF coefficient ranging from -51 to -33 dB).

2 PROCESSING OF AUDIO MATERIAL

In this experiment the only type of processing of audio material was just low-pass filtering of the audio signal. The following questions had to be answered before performing this task:

- What and how many cut-off frequencies should be used?
- What type of filter characteristic is most suitable for this experiment?
- How to filter multichannel audio in order to obtain a usefulresult from the listening test? All channels simultaneously? Or maybe also selected groups of channels?

2.1 Filter characteristics and processing details

According to results of the pilot experiment, literature review and estimation of the complexity of the experiment, it was decided to use 3 cut-off frequencies presented in Tab. 2.

Cut-off frequency	Comment
3.5 kHz	HF bandwidth limit corresponding to the recommendations **for control circuits** used for supervision and co-ordination purpose in broadcasting [8]
7 kHz	HF bandwidth limit for **commentary circuits** [9]; also bandwidth of the "wide-bandwidth telephony" [10]
10 kHz	HF bandwidth limit for **occasional circuits**

Tab. 2 Cut-off frequencies used for filtering of programme material according to international recommendations [6, 7, 8, 9, 10]

According to the recommendations, the filter characteristic should correspond to the characteristic of an input anti-aliasing filter. Therefore the **13th order, IIR, Chebychev I** filter was adapted to our purposes. Amplitude ripple distortions in the pass-band (including cut-off frequency) were equal to ±0.1 dB. Preliminary auditioning of the processed items showed that apart from reducing of the HF content the filter was also introducing some distortions due to a non-linear phase response and due to a very steep slope of its characteristics. In order to eliminate the problem with the phase distortions a FIR type of the filter should be used. However, it was decided to use an IIR type of filter with its inherent distortions in order to simulate the possible perceptual effects of real anti-aliasing filters.

Degrad. Type No.	Filtered channels	Cut-off frequency
1	All channels	3.5 kHz
2	*L, R*	
3	*C*	
4	*LS, RS*	
5	All channels	7 kHz
6	*L, R*	
7	*C*	
8	*LS, RS*	
9	All channels	10 kHz
10	*L, R*	
11	*C*	
12	*LS, RS*	

Tab. 3 Degradation types used in the experiment (*C* - centre channel, *L* - front left, *R* - front right, *LS* - left surround, *RS* - right surround)

There are several ways of filtering the multichannel audio. The first and the most obvious one is to filter all channels simultaneously. It is also possible to filter only selected channels leaving remaining ones unprocessed. The latter case may be especially interesting from broadcasting point of view, since it is important to know how to limit the bandwidth of multichannel audio with as small as possible loss of quality. For example, one may want to broadcast multichannel audio with full-bandwidth front channels and band-limited rear channels, in order to limit the required data rate. Therefore it is important to know

how to limit the bandwidth of all or selected channels with minimal loss of quality. In order to answer this question different ways of filtering (degradation) of audio material was used in this experiment. The detailed list of all possible degradation types is presented in Tab. 3. Only "symmetrical" configurations of filtered channels were used, since it is likely that "asymmetrical" configurations (like filtering of solely the front left channel with remaining channels unprocessed) will cause greater degradation of quality due to easily perceived distortion in spatial characteristic.

2.2 Loudness equalisation

The loudness of all stimuli used in the experiment (both original and processed) was equalised in order to minimise the experimental error due to loudness changes. The level of the audio source material was adjusted to achieve the loudness at the listening position equal to **41 sones**. This value was assessed by the authors as the most comfortable during informal listening tests. Loudness measurements were accomplished by analysing L_{eq} in 1/3 octave bands over a 32 sec. time window (audio material was looped). Loudness was calculated using Moore's loudness model [11]. Since that model was originally developed for stationary signals only, it was necessary to check its applicability to the loudness equalisation of the non-stationary, but relatively consistent audio material used in this experiment. Informal listening tests showed that the obtained results were satisfactory. Tab. 4 shows sound pressure level (*SPL*) for all items after loudness equalisation (measurements were averaged over a 32 sec. time window). Measurements were carried out at the reference listening position.

Item No.	Programme type	*SPL* (L_{eq}) [dB lin.]	Loudness [sones]
1	Classical music (*F*-*B*)	76.6	41
2	Pop music (*F*-*B*)	79.7	
3	Pop music (*F*-*F*)	78.8	
4	Movie (*F*-*B*)	76.9	
5	TV Sport (*F*-*F*)	74.4	
6	Applause (*F*-*F*)	75.5	

Tab. 4 Sound pressure level (*SPL*) and loudness measured at the reference listening position for selected programme material

3 SELECTION OF LISTENING PANEL

The listening panel consisted of **17 listeners**. They were recruited by means of a screening procedure during which a questionnaire, audiometric measurements and a special screening test were carried out to verify listener's reliability and consistency. The term "reliability" is in this paper used as a listener's ability to correctly distinguish between the hidden reference (unprocessed excerpt) and processed items.

3.1 Questionnaire

The main objective of using a questionnaire was initial identification of listener's background. Thanks to the questionnaire it was possible to divide all 25 respondents (mainly students of the Department of Sound Recording) into two groups:

- potentially experienced (23 candidates)
- inexperienced (2 candidates).

Exemplary questions used in the questionnaire are presented in the Appendix (A2).

3.2 Audiometric measurements

Candidate's hearing was examined using a standard audiometric measurements [12]. According to the obtained results more than half of candidates had normal hearing (losses ≤ 15 dB HL). Other candidates had slight losses (16-25 dB HL) or even mild losses (26-40 dB HL). Despite of the fact that the threshold hearing of some candidates differed from normal, it was decided to invite all candidates to take part in a screening test and then compare results of the audiometric measurements with results of the screening test in order to find out any correlation between hearing threshold and consistency and/or reliability.

3.3 Screening test

22 candidates agreed to take part in the screening test. Only one excerpt was used in the screening test. This item was available for the listeners during the test in 4 different forms:

- unprocessed (hidden reference)
- processed (front left and right channels filtered down to 10 kHz)
- processed (all channels filtered down to 7 kHz)
- processed (all channels filtered down to 3.5 kHz)

Listeners were asked to grade the degree of degradation of basic audio quality (in comparison with the reference). There were 6 consecutive trials in which listeners graded each item in random order. Before the proper screening test listeners had opportunity to familiarise themselves with the test and graphical interface.

In order to estimate listener's consistency (ability to give the same grades for the same stimuli) the error variance from ANOVA analysis was taken into account. Listener's reliability (more precisely "discrimination ability") was checked using *F*-statistic from ANOVA analysis. Detailed examination of obtained data showed that listener's consistency may be different for different items. For example, some subjects were very consistent when evaluating the quality of slightly impaired item (front left and right channels filtered down to 10 kHz) and when detecting the hidden reference. However, they were making significant mistakes when evaluating more impaired items (e.g. all channels filtered down to 3.5 kHz). On the other hand, other subjects were more consistent when evaluating highly impaired items and less consistent when evaluating slightly impaired item and detecting the hidden reference.

Another interesting result was obtained in case of one listener who had very good consistency (low error variance) but at the same time very poor sensitivity (small *F*-statistic). His poor reliability was caused by inability to discern between the slightly impaired item and the hidden reference. Since he was grading these two items equally his error variance was small due to his consistency in making these mistakes (in other words, he was making mistakes in a consistent way). Therefore this result seems to confirm recommendations that not only error variance but also *F*-statistic should be used in analysing data from the tests aiming at recruitment of the experienced listening panel [13].

Obtained results from the screening test allowed to select a group of **17 experienced listeners** who were used in the main experiment. It was decided to use also one inexperienced listener in the experiment in order to check his ability to "learn" how to grade the items in a more consistent/reliable manner.

It was difficult to establish any clear relation between the hearing threshold and listener's consistency/reliability. Two out of five "rejected" candidates during the screening test had mild hearing losses, but three remaining "rejected" listeners had normal hearing threshold.

4 EQUIPMENT

Five main loudspeakers were arranged according to the ITU-R BS.775 Recommendation [1] (see Fig. 3). 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. Bass management system with crossover frequency equal to 85 Hz was applied for front channels. (Signals from front channels below 85 Hz were summed with the LFE signal and fed to the subwoofer). The gain of the LFE channel in the console was set up +10 dB higher than the gain of the main channels. The technical specifications of the loudspeakers used in the experiment are presented in the Appendix A3.

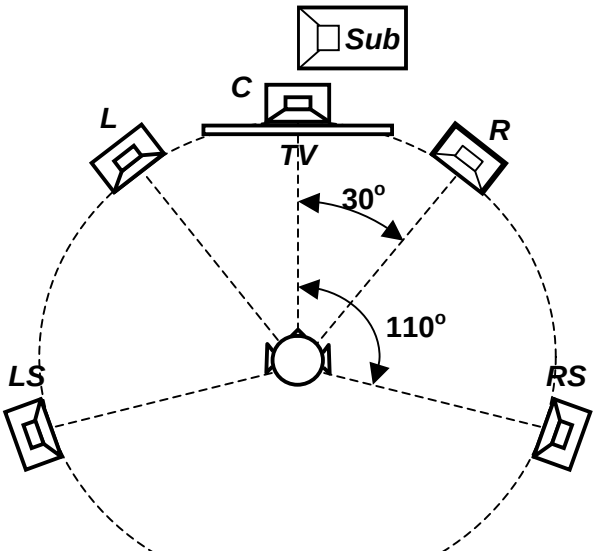


Fig. 3 Loudspeaker set-up used in the experiment

A TV monitor (42" plasma display, 16:9 aspect ratio) was used for visual presentation. 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 [14]). It was not easy to decide where to install the TV monitor with respect to the centre loudspeaker. Several options were informally tested. Eventually it was decided to set up the TV monitor below the centre loudspeaker and to fix the centre loudspeaker higher than the remaining main channels. It was the most comfortable arrangement for the listeners/viewers. To minimise the phase distortions at high- and mid- frequencies due to the different distance between the listener and the tweeters of the front loudspeakers, the centre loudspeaker was installed upside down in such a way that the tweeters were aligned at the same height (see Fig. 4).

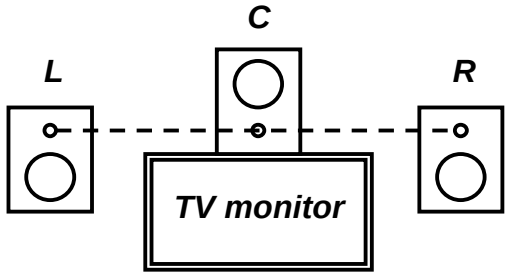


Fig. 4 Set-up of the centre loudspeaker with respect to the TV monitor and other loudspeakers

The listening tests were automated using the "Alex" software developed at the Department of Sound Recording. It was run on the SGI computer with a built in digital audio (ADAT) and analogue video extension cards. The audio items were stored using 6 channel uncompressed 'wav' audio files whereas the video material was stored in M-JPEG format using 0.85 spatial compression factor. The audio signal was transmitted digitally from the SGI to the digital mixing desk (Yamaha O2R) and then fed to the active loudspeakers using the analogue connections. The computer monitor with a mouse was set-up in a front of the listener low enough that any distortions due to acoustical "shadowing" or reflections were minimised. The computer monitor did not screen the TV monitor.

5 ACOUSTICAL CONDITIONS

The listening tests were conducted in the Listening Room of the Department of Sound Recording, University of Surrey. The acoustical parameters of this room conforms to the requirements of ITU-R Recommendation BS.1116 [15].

All main channels (*L, R, C, LS, RS*) were aligned relative to each other with a tolerance less than ± 0.25 dB SPL (measured at the reference listening position).

Phase alignment of the subwoofer was performed according to the manufacturer's instruction. The subwoofer phase offset was set to -270° at the crossover frequency (85 Hz).

The Surround Sound Forum approach was used for alignment of the subwoofer with the bass management system (this method is explained in details in [16]). Two band-limited noise signals of the same level were used for alignment of the subwoofer. The first signal was a noise filtered in the range between 25 and 50 Hz. The second signal was noise filtered in the range between 125 and 250 Hz. The spectrum of the first test signal is below the crossover frequency, whereas the spectrum of the second one is over the crossover frequency (85 Hz). During the subwoofer alignment procedure these two signals were generated alternately over the bass management system. The first signal (with spectrum below the crossover frequency) was played back by the subwoofer whereas the second signal was played back by front channels. The subwoofer potentiometer was adjusted to obtain the same sound pressure level at listening position for both test signals.

Absolute level alignment was carried using a modified Surround Sound Forum approach. The modification consisted in reducing the alignment signals by 10 dB. The reason for this modification was the high sensitivity of the loudspeakers in conjunction with a high analogue output level of the console's DACs. The alignment procedure was as follows:

- The band-limited pink noise (200 Hz-20 kHz, -30 dBFS RMS) was generated consecutively through each main loudspeaker (one channel at a time). The input sensitivity potentiometers in each loudspeaker were adjusted to achieve Sound Pressure Level (SPL) at the optimum listening position equal to 78 dBA (slow).

- The band-limited non-correlated pink noise (200 Hz-20 kHz, -30 dBFS RMS) was generated through all main channels at the same time. The sound pressure level (SPL) measured at the optimum listening position was equal to 85 dBA (slow).

During all alignment procedures the level of the main channels in the console was equal to unity (0 dB). The gain of the LFE channel in the console was set up +10 dB higher than the gain of the main channels.
Once the alignment procedures had been finished the selected audio material was auditioned. Since the subjectively perceived loudness was too high it was decided to attenuate the level of all channels in

the console by 10 dB. The resultant loudness and sound pressure level for each item used in the experiment is presented in previously discussed Tab. 4.

6 EXPERIMENTAL DESIGN

6.1 Listening test method

It was decided to use a double-blind multi-stimulus test method with hidden reference and hidden anchors (**MUSHRA** [6, 7]) as a basis for experimental design. The main reason of this choice was suitability for assessment of medium and large impairments (the quality of most of processed items used in this experiment was degraded quite considerably). Moreover, MUSHRA test allows for quick comparison and assessment of large number of stimuli which is beneficial in terms of duration of a listening test (details about number of stimuli assessed by each listener and duration of the test are discussed later in this section).

Before taking part in the tests the listeners were asked to read carefully the instructions and listen to the original (unprocessed) items and the most degraded ones (all channels filtered down to 3.5 kHz). Then they took part in the listening test. The reason that there was no any additional training phase before the test is that all listeners were already familiarised with the use of the interface and the grading scale (they had been trained before the screening test).

The graphical user interface used during the test is shown in Fig. 5. The main button ('REFERENCE') was used to play back the original (unprocessed) item. Buttons labelled as "A", "B", "C", …"H" represent processed items and a hidden reference being a copy of the unprocessed excerpt. Sliders were used to record scores given by the listeners for each item.

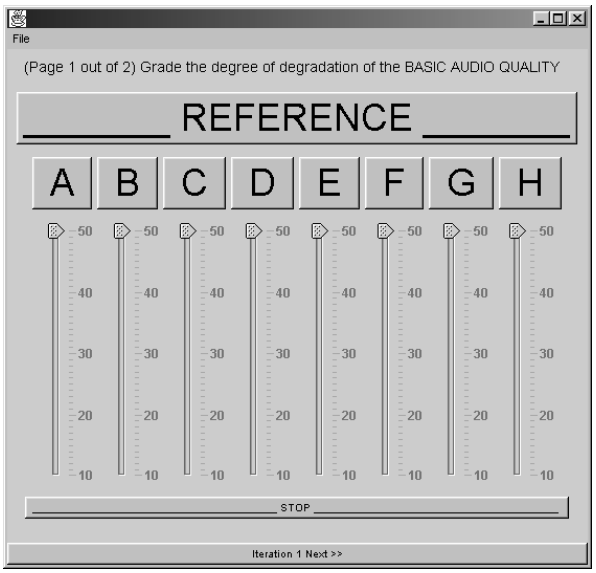


Fig. 5 Graphical user interface used during the test

Listeners were asked to grade the degree of degradation of quality of processed items in terms of **Basic Audio Quality** defined as the global attribute describing **any** and **all detected differences** between the reference and the evaluated excerpt. Subjects were asked to grade the degradation of basic audio quality using the scale adapted from the ITU-R BS. 1116 Recommendation [15]. The reason of using this scale instead of the original one recommended by MUSHRA recommendations [6, 7] was that during the pilot experiment (not described in this paper) listeners preferred 'relative' scale rather than 'absolute' one.

This scale is presented in Tab. 5. Because of some limitations of the software used for running subjective tests (problems with recording fractional values of scores) the original scale values (1 to 5) were replaced in the graphical interface by numbers ranging from 10 to 50 (see Fig. 5). Once the test has been completed the obtained scores were scaled back to the original range (1 to 5).

Impairment	Grade
Imperceptible	5
Perceptible, but not annoying	4
Slightly annoying	3
Annoying	2
Very annoying	1

Tab. 5 Scale used for evaluation of audio quality [15]

Listeners were instructed that the scale was continuous and they were free to record their scores using any number from minimum to maximum of the scale. Labels (numbers) on the scale defined only some characteristic points.

Subjects were instructed that one or more excerpts should be given the grade 'Imperceptible' because the unprocessed reference excerpt was included as one of the excerpts to be graded. Listeners could listen to the excerpts in any order, any number of times. The audio excerpts were looped. In order to make quick comparisons subjects could switch synchronously between the different excerpts.

During the screening test some listeners found it difficult to interpret the scale, especially they found it difficult to interpret the word 'annoying' since it may by context dependent. For example, some degrees of quality degradation may be 'not annoying' when somebody is listening to the personal hi-fi or listening to the audio over the Internet, whereas the same degrees of quality degradation may be 'annoying' when somebody is listening to the DVD-A and as a consequence has higher expectations. Therefore in this test listeners were asked to assume (imagine) that they were listening to an 'audio-visual home-theatre system' installed in a living room. On this basis they were asked to make their own interpretation of the word 'annoying' and be consistent in their grading.

The listeners were asked to have their eyes closed during audio-only presentation and to keep eyes opened and fixed on the TV monitor when the audio-visual material was presented (of course they had to look at the computer monitor occasionally in order to switch between the stimuli and to record their scores using a mouse).

It was emphasised that during the audio-visual presentations subjects were still expected to grade the quality of audio, not video.

6.2 Experimental factors

Following factors were used in the experiment: 'Cut-off frequency', 'Band-limited channels', 'Item' and 'Picture'. Tab. 6 shows all experimental levels and values corresponding to these factors. The last factor ('Picture') was included to the experiment in order to check whether there is any relation between evaluation of audio quality and picture presence.

Factors	No. of levels	Values
Cut-off frequency	3	3.5 kHz
		7 kHz
		10 kHz
Band-limited channels	4	All channels
		L, R
		C
		LS, RS
Item	6	Classical *F-B*
		Pop *F-B*
		Pop *F-F*
		Movie *F-B*
		TV Sport *F-F*
		Applause *F-F*
Picture	2	ON
		OFF

Tab. 6 Experimental factors, levels and values used in the experiment

6.3 Blocking and randomisation

Taking into account all the experimental factors and levels there were 144 excerpts to be graded by each listener (3 x 4 x 6 x 2). This number did not include listener's "error check" excerpts like hidden reference (HR) or hidden anchor (HA). Since there are too many excerpts for evaluation in one session (>144) it was necessary to block them into 4 separate sessions. For technical reasons the main blocking factor was 'Picture'. Another blocking factor was 'Item' (in each session 3 items out of 6 were included). Fig. 6 shows an exemplary schedule of the listening test for a randomly selected subject.

Time

Session No.	1			2			3			4		
Picture	OFF			ON			ON			OFF		
Item No.	3	2	5	4	5	3	1	2	6	1	4	6

Trial No. ***(page)***	*Trial A*	*Trial B*
Stimuli ***(buttons on the screen)***	12, 7, 5, 2, 11, HA, HR, 3	HA, 8, 6, 4, HR, 9, 10, 1

Fig. 6 Exemplary design of the listening test for a randomly selected listener (stimuli at the bottom part of this figure were coded according to Tab. 3)

Each item had 12 processed versions (see Tab. 3). According to author's experience this number of items was too large for evaluation in one trial. In other words, the evaluation task seems to be too tiresome when listeners have to evaluate more than 10 items at a time. Therefore, the processed versions of each item were blocked between the two consecutive trials (*Trial A* and *Trial B* in Fig. 6). In each trial the hidden reference HR and the hidden anchor HA were included in order to check the listener's consistency/reliability. The hidden reference HR was an unprocessed copy of the original excerpt whereas the hidden anchor HA was a copy of the most degraded item (all channels filtered down to 3.5 kHz).

During the experimental design the schedule of sessions and the order of presentation of items within each session were randomised for each subject separately. The order of assigning the stimuli to buttons on the graphical interface was also randomised (see example at the bottom part of Fig. 6).

The average duration of one session was equal to about 25 min. Breaks between the two consecutive sessions for each subject were never shorter than 1 hour (on average the breaks lasted a few hours or sometimes even a few days). The whole listening test was carried out within 2 weeks.

6.4 Experimental details

The investigation was limited to the optimal listening position, since it was assumed that degradation of quality due to the limitation of high frequencies is perceived in a similar way within a relatively wide listening area. This assumption allowed for considerable simplification of the experimental design.

The audio-visual material was looped during each trial. The subjects were able to switch between different audio items at their discretion. It is important to note that after switching the new item the play back was continued from the time-"point" the previous item has reached during switching (in other words it was synchronous type of switching). A cross-fade transition between switched audio item was used in order to avoid any problems with clicks. The looped accompanying visual excerpt was displayed synchronously with audio.

The luminance of the TV monitor, the luminance of the computer monitor and background room illumination were not measured but these parameters were kept constant during the experiment.

7 DATA ANALYSIS

In the first part of this section results of descriptive analysis of obtained data, including information about post-screening and bias effect, will be presented. Then a test of the assumptions for analysis of variance (ANOVA) will be discussed.

7.1 Post-screening

Preliminary graphical examination of obtained data using scatter plots showed that results obtained from one listener differed significantly from the results of other subjects (Fig. 7). This user for some reason decided to grade all excerpts using only two values from the top of the scale (value 4 and value 5). Therefore data obtained from this subject was post-screened. During the screening test this subject was grading similarly to other listeners, therefore he was not excluded from the listening panel before the main test.

7.2 Bias effect

Further examination of obtained data revealed the existence of bias effect due to labels (numbers) on the grading scale. Although listeners were instructed that this scale is continuous (they were free to use any value from the scale), during the listening test they tended to give grades equal to labels more frequently than other grades from the scale. The bias effect is presented on a histogram of the obtained scores averaged for all subject and all experimental factors and plotted against the grading scale (Fig. 8). Apart from the main bias effect it is also possible to note an additional slight bias effect for grades like 15, 25, 35 and 45 (values exactly between the labels). It is clear that that this effect causes distortions in distribution of scores and thus may lead to violations of assumptions of normality for ANOVA. However, the main advantage of using a scale with labels is its meaningfulness (a clear relation between scores and audio quality expressed in terms of the labels). Using label-free scales can minimise the bias effect but at the same time may increase experimental error.

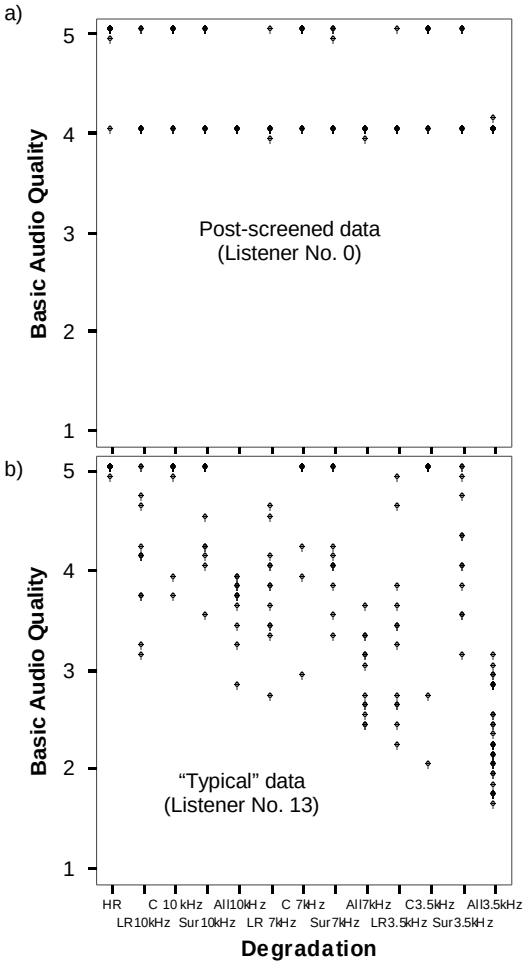


Fig. 7 Exemplary scatter plots of data obtained from two listeners:
a) "outstanding" (post-screened);
b) typical.

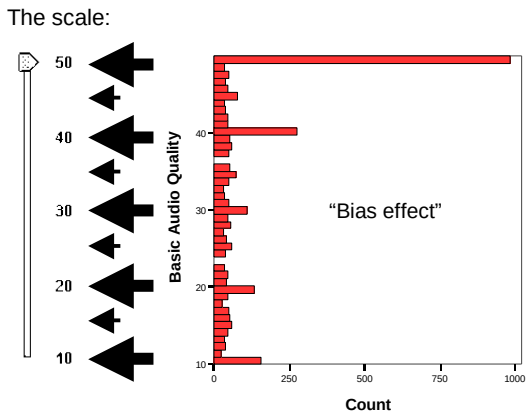


Fig. 8 Histogram of all scores obtained in the listening test (averaged for all listeners and all factors)

7.3 Test of listener's reliability and consistency

In order to check listener's reliability the hidden reference HR was included as one of the excerpts to be graded in each trial. It was assumed that that reliable listener should grade the hidden reference using a maximum value from the scale. Therefore, in order to estimate listener's reliability it was decided to calculate a **coefficient of variation *CV*** for each listener. The coefficient of variation was defined as follows [17]:

$$CV = \frac{s}{x_{\max}} \times 100\,\% \tag{2}$$

where $x_{\max} = 5$ is a maximum value of the scale (expected value for the hidden reference) and s is the standard deviation calculated using scores given for the hidden reference, according to the following equation:

$$s = \sqrt{\frac{\sum (x_i - x_{\max})^2}{n-1}} \tag{3}$$

Calculations were performed separately for scores obtained using audio-only presentation and scores obtained during audio-visual presentation in order to check whether picture presence increased the error of audio quality evaluation due to distractions of attention. Results showed that 13 listeners (out of 16) had quite low coefficient of variance ranging from 0 to 3 % and therefore they were classified as reliable subjects. Remaining 3 listeners had greater values of this coefficient (3–10 %). However, since the values of *CV* in the latter case were still "acceptable" it was decided to include data from both groups in the main analysis of the obtained results. According to a *t* test there was no significant difference between the coefficient of variation calculated for scores obtained using audio-only presentation and scores obtained during audio-visual presentation (picture presence did not increase the error of evaluation of audio quality).

The presented method of verification of listener's reliability is different from the recommended one [15], which is based on the *t* test calculated for the differences of grades obtained for the hidden reference and the least impaired item. The main reason of using the new method is that the recommended one is not "robust" to the erroneous situation when the hidden reference is graded lower than the maximum value and at the same time graded equally to the impaired item. This situation can not happen in triple-stimulus test but may happen in multi-stimulus test employed in this experiment.

The test of consistency of listeners was carried out using a measure of error variance obtained from the ANOVA test. Only scores given for the hidden anchor HA were analysed in an ANOVA test (the remaining scores were excluded from this analysis). Obtained results showed that 8 listeners had error variance ranging from 0 to 0.05 on a 5-point scale (high consistency) whereas the error variance of the remaining 8 listeners was greater than 0.05 and less than 0.15 (results less consistent). Since no abnormal (outstanding) values of error variance were detected the results from both groups of listeners were taken into account in the main analysis of the obtained results. According to a *t* test there was no significant difference between the error variance calculated for scores obtained using audio-only presentation and scores obtained during audio-visual presentation (picture presence did not increase the error of evaluation of audio quality).

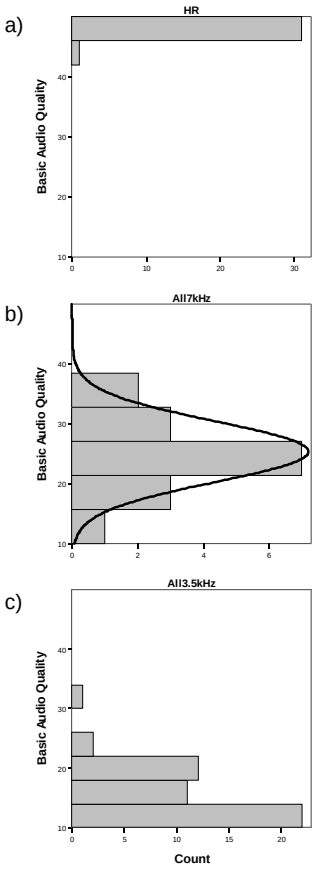
It was interesting to find out that an inexperienced listener included in the listening panel, who had very poor reliability and consistency before the test, developed his skills and in a final "ranking" of listeners was placed at a middle position in terms of consistency and reliability. It proves a view that inexperienced listeners may become experienced ones as a result of taking part in the listening tests [13, 18].

7.4 Test of ANOVA assumptions

Three main assumptions checked for ANOVA test were as follows: 1) independence of grading, 2) normal distribution of scores for each case and 3) homogeneity of variance between cases.

There are several mechanisms causing dependency of grading. For example the evaluation of quality of a given item may be affected by quality of other excerpts contained in the MUSHRA graphical interface. Another possible source of dependency of grading is visual influence of slider positions in the interface. This influence has a similar nature to the bias effect discussed previously. All these and other possible sources of dependency were minimised in this experiment due to randomisation of experimental factors.

A distribution of obtained scores was examined graphically. Fig. 9 shows 3 typical cases. The first case (shown in Fig. 9 *a*) represents distribution obtained when evaluating "high-quality" items like the hidden reference or slightly impaired excerpts. Its strong departure from normality is caused by a scale "boundary effect" (distribution is asymmetrical due to the end of the scale). Graphical inspection of different cases showed that distribution of scores given to "medium quality" items is usually normal (Fig. 9 *b*). Scores obtained for "highly impaired" excerpts are usually asymmetrically distributed due to another "boundary effect" at the bottom of the scale (Fig. 9 *c*).



"Boundary effect" at the top of the scale (non-normal distribution)

Normal distribution

"Boundary effect" at the bottom of the scale (non-normal distribution)

Fig. 9 Typical distributions of obtained data:
- a) Non-normal distribution ("high-quality" excerpts);
- b) Normal distribution;
- c) Non-normal distribution ("severely impaired" excertps).

More formal examination of distributions of obtained data using Kolmogorov-Smyrnov test revealed that about 25 % of cases with significant departure from normality.

Another assumption for ANOVA is that data in each cell come from populations with the same variance (homogeneity of variance). This assumption was tested using a Levene's Test of Equality of Error Variances. Obtained results showed that variances were not homogenous (the null hypotheses of equal error variances across groups rejected).

According to obtained results two main assumptions for ANOVA were violated (normality of distributions and homogeneity of variance). It is known that the ANOVA test is "robust" to violation of normality assumption provided the sample size is large (minimum 15 cases per group) [19]. In this experiment this condition was fulfilled since in the worst case (averaging only for subjects) the minimum number of scores obtained for each excerpt was equal to 16 (number of listeners after post-screening). Moreover, ANOVA test may still give reliable results even when variances are not equal across different groups provided the number of cases in each group is the same [20]. This condition was not fulfilled in the experiment since the hidden reference HR and the hidden anchor HA were evaluated more frequently than other excerpts (unbalanced design). Therefore it was decided to balance the obtained data (equalise the number of cases across groups) by calculating and taking into ANOVA test the mean values of scores obtained for the hidden reference HR and the hidden anchor HA ("raw" scores obtained for HR and HA were ignored in the main analysis). After this pre-processing of data the use of ANOVA in our experiment was legitimate.

There was a listener who missed one trial during one session and did not report this fact to the person supervising the listening test. As a result there were 8 missing data. It did not cause any serious problem in analysis since this value constituted only about 0.2 % of all data. However, due to that fact, it was necessary to use a Type IV model of ANOVA taking into account missing values.

As far as the post-hoc multiple comparison tests are concerned it was decided to use a 'Dunnett's C' test, which did not assume equal variance across groups.

8 RESULTS

8.1 ANOVA test

Tab. 7 shows results of ANOVA test for all experimental factors. According to the highest partial η^2 value (last column and 4th row in the table) the scores obtained in the listening test were affected mainly by the factor of groups of band-limited channels ($\eta^2 = 0.739$), not by the 'cut-off frequency' factor, which was a surprising result (it was expected the 'cut-off frequency' would have the greatest effect on the results). It is also possible to note that interaction between the band-limited channels and the items had the second significant effect on the obtained scores ($\eta^2 = 0.507$). The third important factor affecting the scores was the cut-off frequency 'FREQ' ($\eta^2 = 0.262$). Interaction between cut-off frequency and band-limited channels had also some effect on the obtained scores ($\eta^2 = 0.134$). The statistical significance of mentioned effects was at the level of $P < 0.001$. The 'Item' factor had only a small effect on the scores ($\eta^2 = 0.034$) whereas the 'Picture' influence on the obtained results was insignificant ($P = 0.468$).

Tests of Between-Subjects Effects

Dependent Variable: Basic Audio Quality

Source	Type IV Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	2667.951[a]	143	18.657	67.689	.000	.818
Intercept	30332.927	1	30332.927	110050.3	.000	.981
FREQ	210.467	2	105.233	381.795	.000	.262
CHANNELS	1679.980	3	559.993	2031.701	.000	.739
ITEM	21.056	5	4.211	15.278	.000	.034
PICTURE	.145	1	.145	.527	.468	.000
FREQ * CHANNELS	91.491	6	15.248	55.323	.000	.134
FREQ * ITEM	7.313	10	.731	2.653	.003	.012
CHANNELS * ITEM	610.197	15	40.680	147.589	.000	.507
FREQ * CHANNELS * ITEM	39.402	30	1.313	4.765	.000	.062
FREQ * PICTURE	.145	2	7.243E-02	.263	.769	.000
CHANNELS * PICTURE	.246	3	8.186E-02	.297	.828	.000
FREQ * CHANNELS * PICTURE	.297	6	4.957E-02	.180	.982	.001
ITEM * PICTURE	.577	5	.115	.419	.836	.001
FREQ * ITEM * PICTURE	2.150	10	.215	.780	.648	.004
CHANNELS * ITEM * PICTURE	2.909	15	.194	.704	.783	.005
FREQ * CHANNELS * ITEM * PICTURE	2.955	30	9.849E-02	.357	1.000	.005
Error	593.702	2154	.276			
Total	33606.770	2298				
Corrected Total	3261.654	2297				

a. R Squared = .818 (Adjusted R Squared = .806)

Tab. 7 ANOVA analysis for all experimental factors

8.2 Effects of band-limitation

According to the ANOVA test the main factor affecting the scores was the factor of groups of band-limited channels. Since the effect of band-limitation is strongly dependent on programme material, as it was detected by the ANOVA test, it is necessary to inspect the results of band-limitation separately for each type of programme material (Fig. 10).

For classical music ('Classical *F-B*') simultaneous band-limitation of all channels ('All') caused substantial deterioration of basic audio quality. According to Fig. 10, this effect was graded as almost 'Annoying'. Limitation of HF content of solely front left and right channels ('LR') also resulted in significant deterioration of quality, however deterioration of quality was less than in the previous case (mean value between 'Slightly annoying' and 'Perceptible, but not annoying'). The most interesting results (and perhaps surprising ones) were related to perceptual effects of band-limitation of the centre channel ('C') and the surround channels ('Sur'). In both cases these effects were very small. It is possible to observe in Fig. 10 that for these cases the mean values and 95 % confidence interval limits are located higher than a "threshold of annoyance" represented by the dashed reference line. These results mean that band-limitation of the centre or of the rear channels for classical music caused only small deterioration of quality. For example, the perceptual effect of limitation of HF content in the rear channels was almost 'Imperceptible' for classical music.

Results obtained for both pop-music items ('Pop *F-B*' and 'Pop *F-F*') were similar to the results obtained in case of classical music. Simultaneous limitation of HF content in all channels caused the highest degree of degradation of audio quality. Band-limitation of the front left and right channels also resulted in considerable deterioration of quality, however not as severe as in the previous case. However, the effects of low-pass filtering of the centre channel were almost 'Imperceptible'. Band-limitation of the rear channels caused 'Perceptible but not annoying' effect even for the item with foreground content in the rear channels ('Pop *F-F*').

Simultaneous band-limitation of all channels for the 'Movie *F-B*' item' caused considerable degradation of quality similar in magnitude to the effects of simultaneous band-limitation of all channels observed for other items. However, band-limitation of solely the front left and right channels or solely the rear channels caused small changes in quality. In contrary, band-limitation of the centre channel resulted in substantial deterioration of quality. This effect was related to easily perceivable effects of low-pass filtering of the centre channel due to its loud content (dialogue).

Results obtained for the last two items ('Sport *F-F*' and 'Applause *F-F*') are quantitatively similar to the results obtained for classical music item and both pop-music items. The only difference can be observed when comparing scores corresponding to band-limitation of the rear channels. It can be noted in Fig. 10 that limitation of HF content in the rear channels caused some annoying effects (scores below the "threshold of annoyance"). These effects were probably caused by easily perceivable effects of low-pass filtering of the rear channels due to their loud and clear content (foreground content).

It is also interesting to study the obtained results from point of view of the spatial characteristics of programme material. For example, further inspection of Fig. 10 shows that for all three items having *F-B* characteristic the band-limitation of rear channels had small effect on degradation of audio quality (scores above the dashed line) but for two items having *F-F* characteristic this effect caused more significant degradation of quality (scores below the "threshold of annoyance").

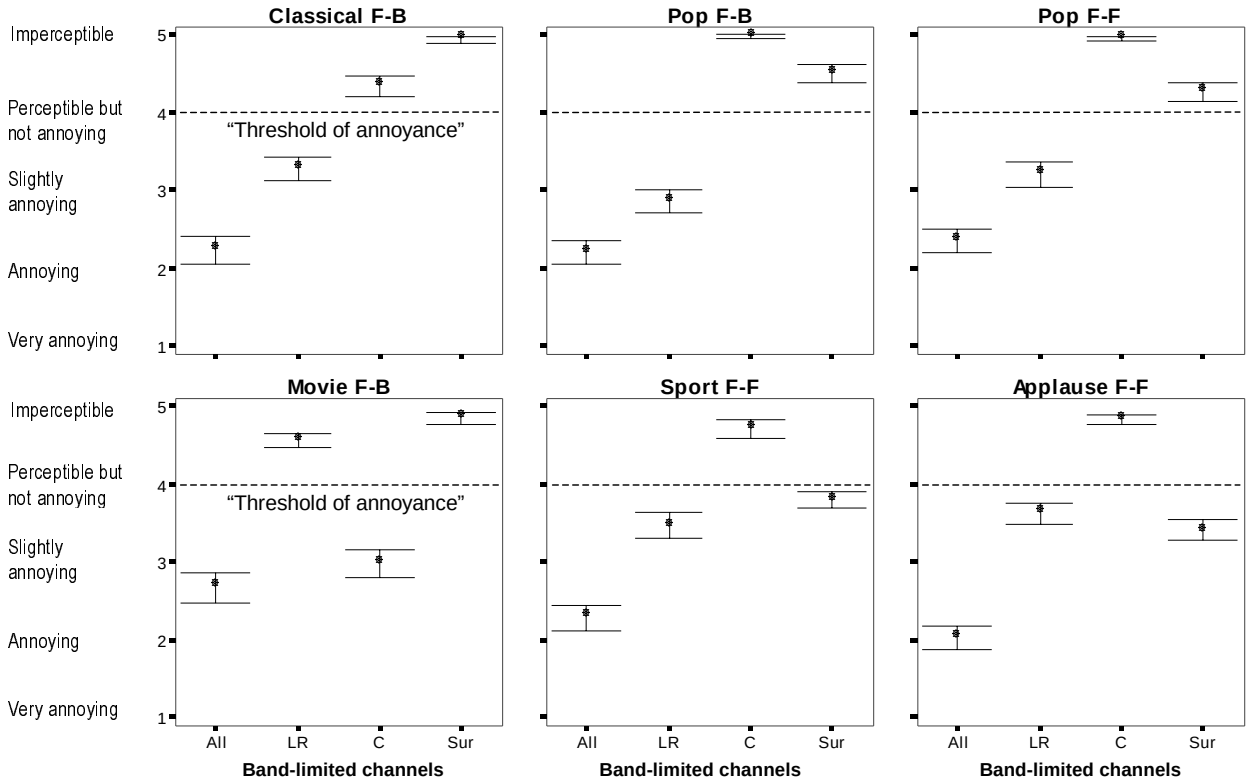


Fig. 10 Degradation of Basic Audio Quality for different programme material (mean values and 95% confidence intervals)

The centre channel is "robust" to band-limitation both for *F-B* and *F-F* items except the 'Movie *F-B*' item. This result shows that it might be useful to introduce a new sub-category of *F-B* characteristic, namely a "*F-B* with prioritised centre channel". This category might be used to represent not only movies with the dialogue, but general types of *F-B* recordings with the loud centre channel (for example classical music with a loud soloist in the centre channel). The results of HF limitation of different channels on audio quality of programme material having different spatial characteristic are clearly presented in Fig. 11. It is interesting that band-limitation of all channels and band-limitation of front left and right channels caused similar effects for material having both *F-B* and *F-F* characteristic. However, in case of the centre channel and surround channels these effects are different. The difference between results for the centre channel is clearly related to the previously discussed effect of degradation of quality in the centre channel of the Movie *F-B* item. The difference between results for surround channels may be explained by the fact that in case of *F-F* material any effects of limitation of bandwidth in the rear channels are more easily perceivable due to the relatively high loudness and clear audio content of the rear channels as opposed to *F-B* material.

According to results of the ANOVA test reported in the previous section 'cut-off frequency' and interaction between 'cut-off frequency' and 'groups of band-limited channels' are important factors affecting the obtained results. Fig. 12 shows the scores obtained during the listening test presented separately for different cut-off frequencies and different groups of band-limited channels. These results basically confirm conclusions drawn from plots discussed previously that regardless of the cut-off frequency band-limitation of all channels or band-limitation of front left and right channels caused significant deterioration of audio quality.

It is also possible to note that band-limitation of the centre channel or band-limitation of the surround channels even down to 3.5 kHz did not, in general, cause any annoying effects. This effect was independent of the value of the cut-off frequency. More detailed plots showing the effects of cut-off frequency for each type of programme material are presented in the Appendix in Section A4.

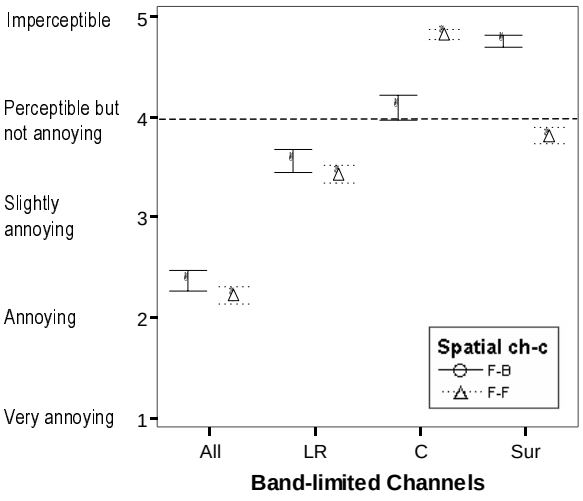


Fig. 11 Degradation of Basic Audio Quality avaraged for programme material of different spatial characteristic

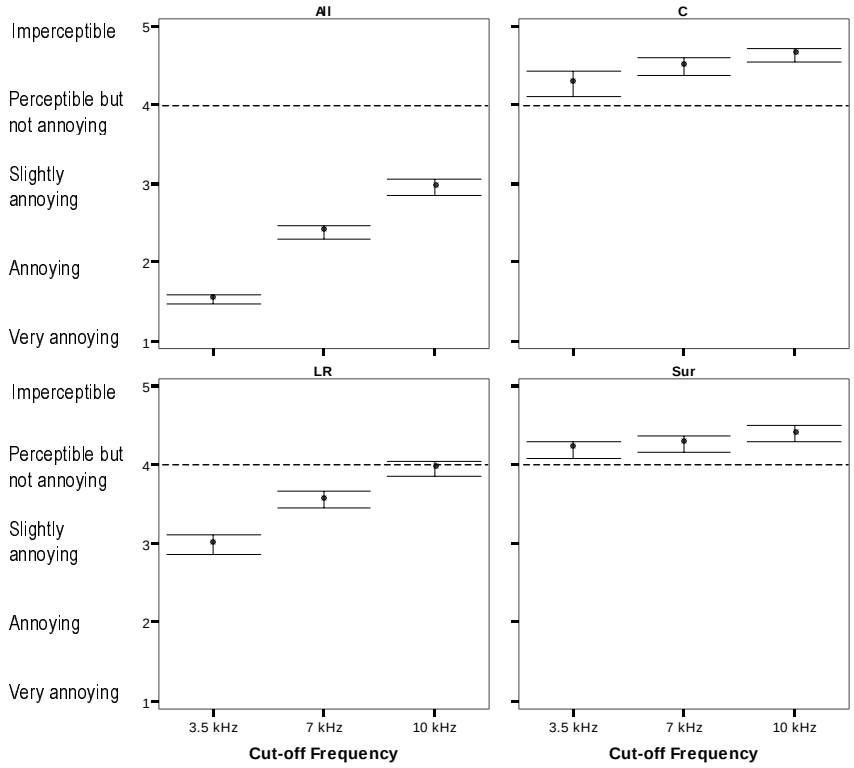


Fig. 12 Degree of degradation of Basic Audio Quality for different groups of channels at different cut-off frequencies

8.3 Picture effect

Although ANOVA test showed that there was no global effect due to picture presence the obtained scores were analysed again, this time for each subject separately. Fig. 13 shows differences between scores obtained during audio-visual presentation and audio-only presentation for each subject. Positive mean values show improvement of audio quality due to video interaction whereas negative values indicate deterioration of audio quality caused by video presence (zero represents no audio-visual interaction). Asterisks in this figure indicate mean values significantly different from zero according to the result of the *t* test. It was found that some of the listeners were more susceptible to video influence than others. For example, subjects No. 9 and 15 had tendency to grade audio quality slightly "better" for audio-visual presentation than for audio-only one. An opposite interaction was found for subject No. 4, 7 and 11. In all extreme cases video presence "shifted" the scores up to ±6 % only, which shows that mentioned effect is small, however statistically significant. This observation is in line with results obtained by Beerends *et. al.* [21].

8.4 Feedback from listeners

Each listener was interviewed after the test. Standard questions concerning the difficulty of the test, the difficulty of using the grading scale, loudness, audio-visual synchronisation and picture compression artefacts were asked. Most of the listeners had difficulty with using the grading scale, mainly because of the word "annoying". They claimed that it was easy to use the top part of the scale, but the difficulty was related to the bottom part of the scale ("Difficult to decide what is annoying"). They found it difficult to establish their own "subjective reference" for "annoying" and be consistent with it. This information indicates how important it is to use an appropriate scale and appropriate labels (if any) in the test. Because of mentioned problems with interpretation of the word "annoying" it was decided to use another scale in the complementary experiment described in [2].

Listeners confirmed that loudness of audio excerpts was at comfortable level. They did not report any problems with audio-visual synchronisation or any distraction caused by video compression artefacts.

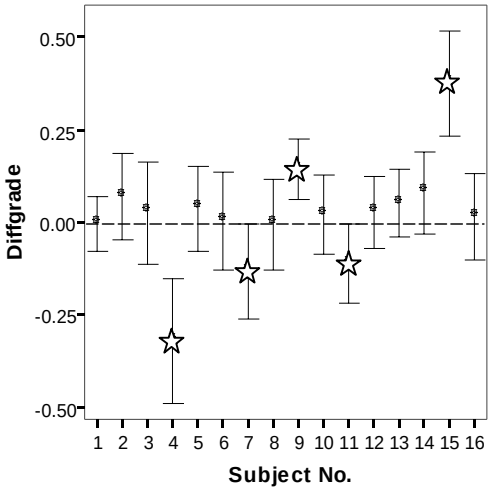


Fig. 13 Differences between Basic Audio Quality scores obtained with picture presence and without picture for **different subjects**

9 DISCUSSION

The most important outcomes of this experiment can be summarised graphically by means of Fig. 14 which shows global effects of limitation of HF content for different groups of band-limited channels. For clarity this figure does not show interactions with other experimental factors, and therefore to some extent "over-simplifies" results. However, it gives a clear indication of what the "average annoyance" was across a range of programme material, which might be especially useful for broadcasters.

It is clear that the worst result (highest degradation of quality) occurred when all channels were filtered simultaneously ('All'). Band-limitation of the solely front left and right channels ('LR') also gave rise to significant deterioration of audio quality but not as large as in the previous case. However, band-limitation of solely the centre channel ('C') or band-limitation of solely the surround channels ('Sur') caused perceptible but not annoying effects for most of programme material (scores above the "threshold of annoyance").

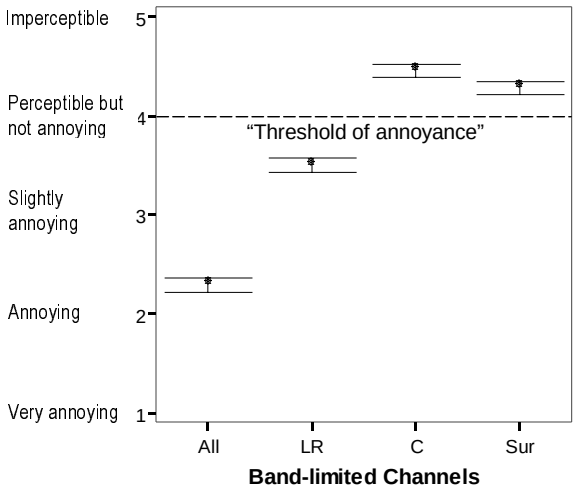


Fig. 14 Degradation of Basic Audio Quality

The main question that arises from analysis of the obtained results is: "Why did limitation of HF content in the centre channel or in the rear channels cause such a small deterioration of the audio quality?". It would be also interesting to find out why this effect is almost cut-off frequency independent. It is clear that this is not caused by inappropriately selected programme material, since according to spectral analyses most of excerpts used in the listening test contained significant HF content (except for the 'Classical music' item which was "less critical"). Therefore it is believed that this effect is caused either by masking or by perceptual streaming or by both mechanisms together. For example, most of excerpts used in the experiment had audio sources spread across front loudspeakers. Therefore, band-limitation of the centre channel was not annoying because signals from unprocessed channels masked any loss of HF content in the centre channel. The nature of masking ("spectral upward spread") may explain the fact that audio quality deterioration of the centre channel did not depend on value of cut-off frequency.

A similar mechanism of masking and/or perceptual streaming may be responsible for the small deterioration of quality in the case of band-limitation of rear channels for *F-B* material. High-frequency components in front channels may "compensate" lack of HF components in rear channels.

According to the results obtained in the listening test, the video presence had a marginal effect on evaluation of audio quality. However, the experimental procedure was limited to a passive way of watching the video (listeners were not asked to do any particular task related to video while evaluating the audio quality). Therefore, one can not exclude the case, in which video may have greater effect on audio evaluation than the effect observed in this experiment.

10 CONCLUSIONS

Effects of high frequency band limitation in a standard (5.1) multichannel audio system on subjectively assessed basic audio quality were investigated. Simultaneous limitation of bandwidth at high frequencies in all main channels of the system caused significant deterioration of basic audio quality at the optimum listening position. High frequency band limitation of solely front left and front right channels also resulted in significant deterioration of audio quality. However, it was found that limitation of bandwidth at high frequencies in solely the centre channel caused small deterioration of quality for most programme material (except for a movie item). It was also found that band-limitation of solely the rear channels caused small changes in audio quality for items having *F-B* spatial characteristic (foreground audio content in front channel and background in rear channels). Mentioned effect of insignificant deterioration of quality in case of band-limitation of the centre or rear channels was almost independent of investigated values of cut-off frequency (3.5, 7 and 10 kHz).

The obtained results indicate that for typical programme material it might be possible to limit the bandwidth of the centre channel without significant deterioration of basic audio quality. The exception is group of items having loud centre channel like movies with a dialogue, opera with a solo singer in the centre channel, etc. According to the obtained results it might also be possible to limit the bandwidth of rear channels with small deterioration of audio quality for items having *F-B* spatial characteristic. Therefore in applications where limitation of high frequencies is unavoidable (for example because of technical and/or economical constraints) it is suggested that one might choose to "sacrifice" the centre channel or rear channels. Band-limitation of front left and right channels or band-limitation of all channels would result in substantial loss of audio quality.

It was identified that video presence may have a small but statistically significant influence on the audio quality evaluation for some subjects.

ACKNOWLEDGEMENTS

The authors would like to express their gratitude to David Meares (BBC, R&D Dept.) for his comments on the results of the experiment described in this paper and for stimulating discussion. The authors are also grateful to Russell Mason (Dept. of Sound Recording, Univ. of Surrey) for his help in statistical analysis. This project was carried out with the financial support of the Engineering and Physical Sciences Research Council, UK. Some of the A-V excerpts used in this experiment was kindly supplied by BBC, R&D Dept. (used by permission).

11 REFERENCES

[1] ITU-R Recommendation BS. 775-1 *Multi-channel stereophonic sound system with or without accompanying picture*. International Telecommunications Union (1992-1994).

[2] S. K. Zielinski, F. Rumsey and S. Bech, "Subjective audio quality trade-offs in consumer multichannel audio-visual delivery systems. *Part II: Effects of low frequency limitation*" AES 22nd International Conference on Virtual, Synthetic and Entertainment Audio, Espoo, Finland (2002) (to be presented).

[3] F. Rumsey *Spatial Audio*. Focal Press. Oxford (2001).

[4] F. Rumsey *Spatial quality evaluation for reproduced sound: terminology, meaning and a scene-based paradigm*. *J. Audio Eng. Soc.* (2002, pending publication).

[5] D. G. Kirby *et al*, "Program Origination of Five-Channel Surround Sound" *J. Audio Eng. Soc.*, **46**, 4, (1999) pp. 323-330.

[6] MUSHRA – EBU *Method for Subjective Listening Tests of Intermediate Audio Quality*. European Broadcasting Union (2000).

[7] ITU-R Recommendation BS. 1534 *Method for the subjective assessment of intermediate audio quality (Draft)* International Telecommunications Union. (2001).

[8] ITU-T Recommendation G. 712 *Transmission performance characteristics of pulse code modulation channels.* International Telecommunications Union. (1996).

[9] ITU-T Recommendation G. 722 *7 kHz Audio – Coding within 64 Kbits/s.* International Telecommunications Union. (1993).

[10] ITU-T Recommendation P.311 *Transmission characteristics for wideband (150-7000 Hz) digital handset telephones* International Telecommunications Union. (1998).

[11] B. C. J. Moore, B. R. Glasberg and T. Baer, "A model for the prediction of thresholds, loudness and partial loudness" *J. Audio Eng. Soc.*, **45**, (1997) pp. 224-240.

[12] ISO 389, *Standard Reference Zero for the Calibration of Pure Tone Air Conduction Audiometers.* International Organization for Standardization, Geneva, Switzerland (1997).

[13] S. Bech, "Selection and Training of Subjects for Listening Tests on Sound-Reproduction Equipment" *J. Audio Eng. Soc.*, **40**, 7/8, (1992) pp. 590-609.

[14] EBU Recommendation Tech 2376-E, *Listening conditions for the assessment of sound programme material. Supplement 1 - multichannel sound*. European Broadcasting Union, Geneva (1999).

[15] ITU-R Recommendation BS. 1116. *Methods for subjective assessment of small impairments in audio systems including multichannel sound systems.* International Telecommunications Union (1994)

[16] Multichannel Universe. *Die Referrenz Demo- und Test DVD.* Surround Sound Forum, Balance and Media City. Balance München DVD, BAL-9500-3 (2000)

[17] A. Diamantopoulos, B.B. Schlegelmilch *Taking the fear out of data analysis.* The Dryden Press, London (1997)

[18] S. Bech, "Training of subjects for auditory experiments" *Acta Acustica*, **1**, (1993) pp. 89-99.

[19] S.B. Green, N.J. Salkind, T.M. Akey, *Using SPSS for Windows*, New Jersey (2000)

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

[21] J. G. Beerends and F. E. De Caluwe, "The Influence of Video Quality on Perceived Audio Quality and Vice Versa" *J. Audio Eng. Soc.*, **47**, 5, (1999) pp. 355-362.

APPENDIX

A1 AUDIO ANALYSES OF PROGRAMME MATERIAL

Tab. A1 shows RMS levels (in dB) for each item used in the experiment. The levels are presented for each channel separately.

Item	L	R	C	LFE	LS	RS
Classical (*F-B*)	-29.3	-28.8	-31.6	-	-31.7	-31.8
Pop (*F-B*)	-26.5	-26.2	-38.9	-40.8	-47.2	-38.4
Pop (*F-F*)	-24.7	-25.9	-25.1	-40.5	-33	-33.7
Movie (*F-B*)	-34.7	-37.3	-22.1	-	-43	-46.3
Sport (*F-F*)	-29.8	-30.6	-32.4	-	-36.2	-34.5
Applause (*F-F*)	-31	-31.1	-35.7	-	-30.4	-29.6

Table A1: RMS levels (in dB) for each item

In Tab. A2 there are presented results of spectral audio analyses of programme material in terms of the coefficient k_{HF} defined by equation (1). This coefficient is an objective measure of the HF content. The values of the k_{HF} coefficient are presented for each item and each channel separately.

Item	L	R	C	LS	RS
Classical (*F-B*)	-32.8	-40.2	-33.5	-50.9	-51.4
Pop (*F-B*)	-13	-10.7	-17.9	-15.1	-16.5
Pop (*F-F*)	-24.8	-25	-30.7	-19.5	-27.3
Movie (*F-B*)	-35.3	-35.1	-28.1	-32.1	-31.7
Sport (*F-F*)	-31.4	-27.6	-28.8	-29.9	-25.5
Applause (*F-F*)	-27.8	-27.4	-26.4	-29.6	-27.7

Table A2 High-frequency content coefficient k_{LF} (in dB) for each item

A2 EXEMPLARY QUESTIONS FROM THE QUESTIONNAIRE

The exemplary questions used in the questionnaire to identify listener's background were as follows:

- Do you have any experience in formal listening tests?

  (*People who had taken part in listening tests were considered as potentially experienced)*

- How many hours per day do you listen to the music in an active way (that is fully concentrated on music)?

  (*Candidates spending more than 1 hour per day were classified as potentially experienced*)

- Could you give any examples of surround sound formats of transmission and/or compression that you are familiar with?

  (*Answer to this question helped in identifying people interested in surround sound. Knowledge of mentioned formats was beneficial*)

- Please, give some examples of CDs and/or DVDs that in your opinion should be used in the listening tests. Give some short justification of you choices.

  (*This question was difficult to answer for most of the candidates, since they did not know the nature of the planned listening test. However, it was assumed that people who could recommend some recordings for the listening tests and who could justify their suggestions were potentially suitable as listeners*)

A3 TECHNICAL SPECIFICATIONS OF LOUDSPEAKERS

Specifications of the loudspeakers used during the listening test are as follows:

- Five main loudspeakers – Genelec 1032A (active monitors), Free field frequency response: 42 Hz – 21 kHz (±2.5 dB)

- Subwoofer – Genelec 1094A (active monitor), Free field frequency response: 29 Hz – 80 Hz (±2.5 dB); Short term output power: 400 W (8 Ω), Crossover frequency: 85 Hz.

A4 EFFECTS OF LOW-PASS FILTERING FOR DIFFERENT PROGRAMME MATERIAL

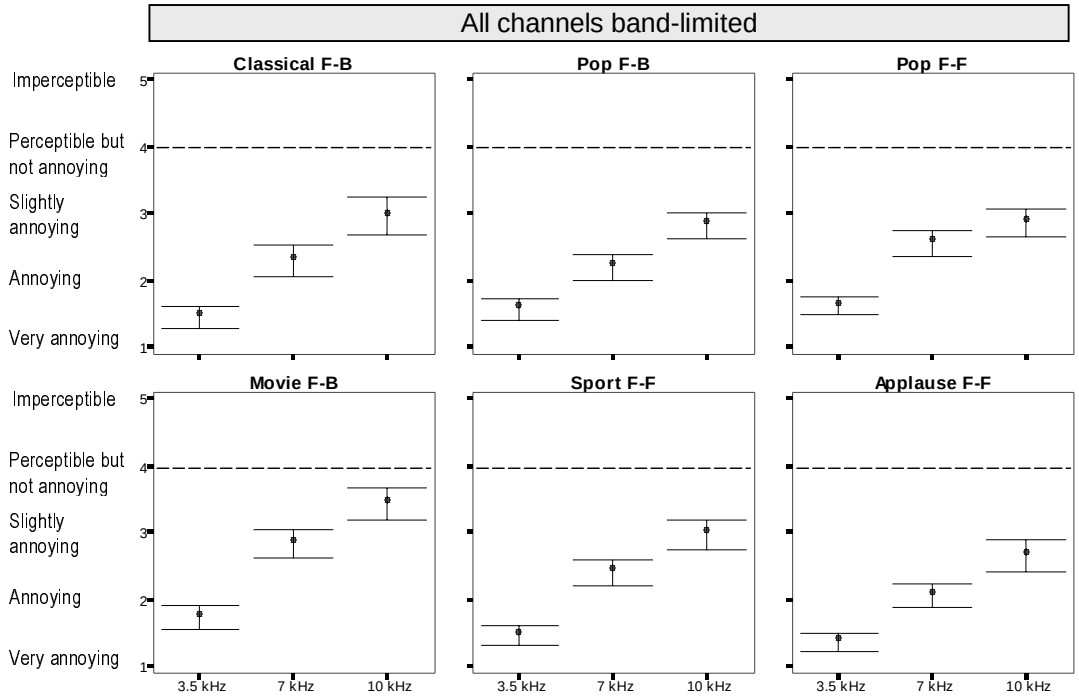


Fig. A1 Degradation of the basic audio quality caused by band-limitation of **all channels**

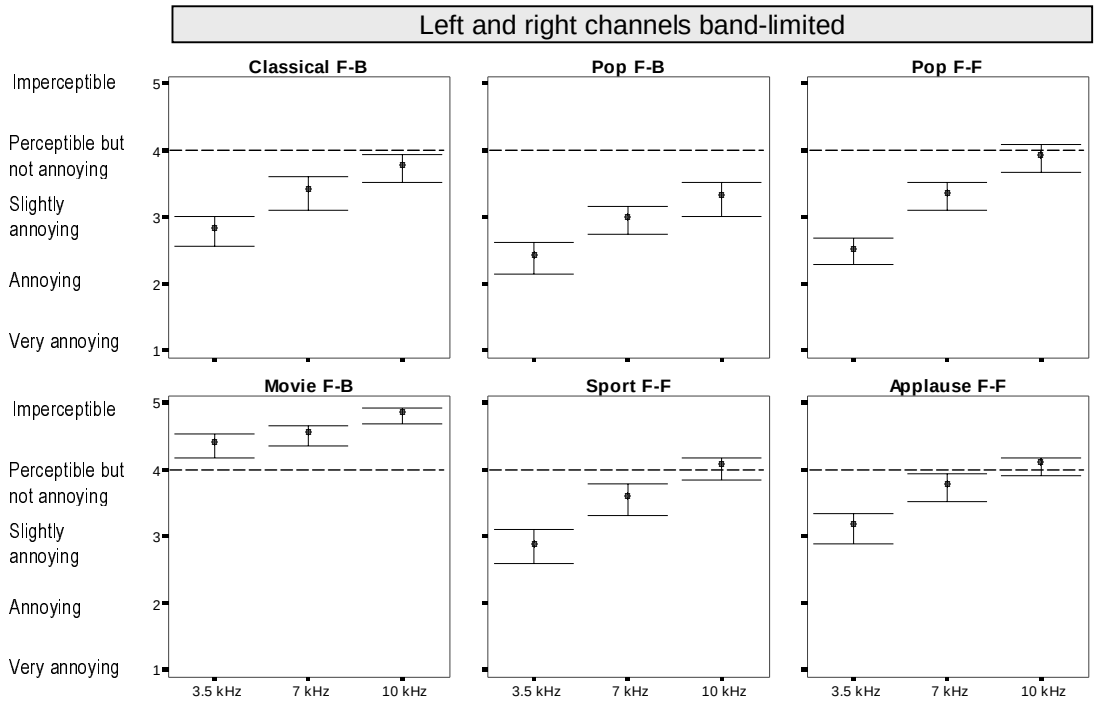


Fig. A2 Degradation of the basic audio quality caused by band-limitation of **front left and right channels**

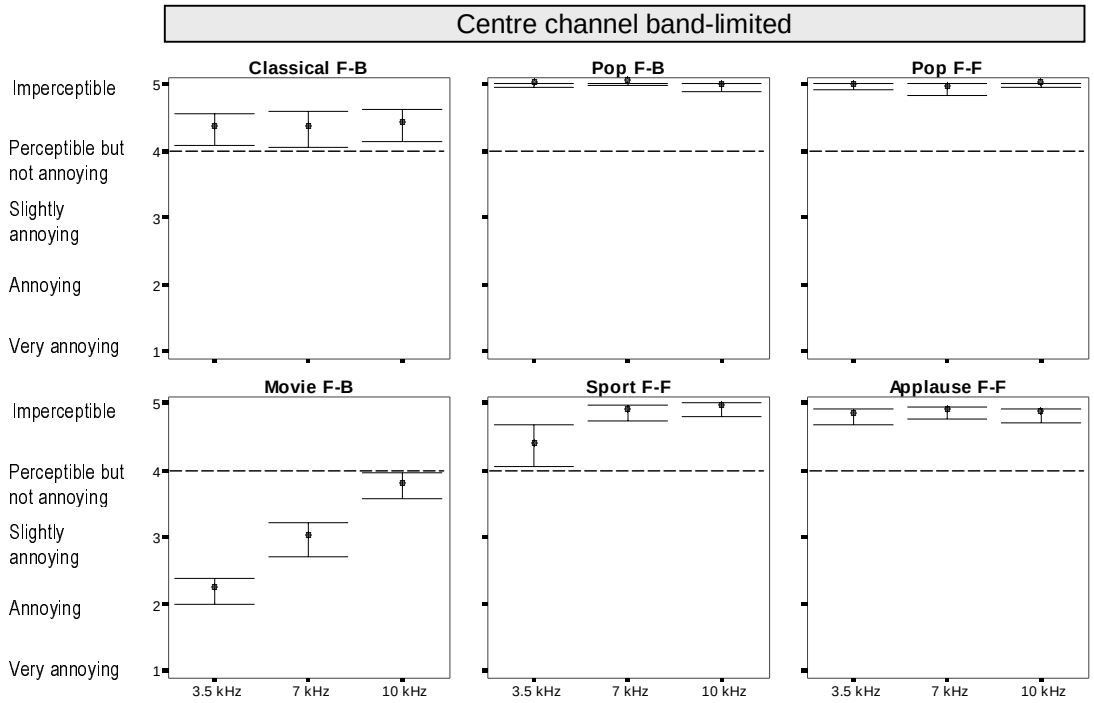


Fig. A3 Degradation of the basic audio quality caused by band-limitation of **for the centre channel**

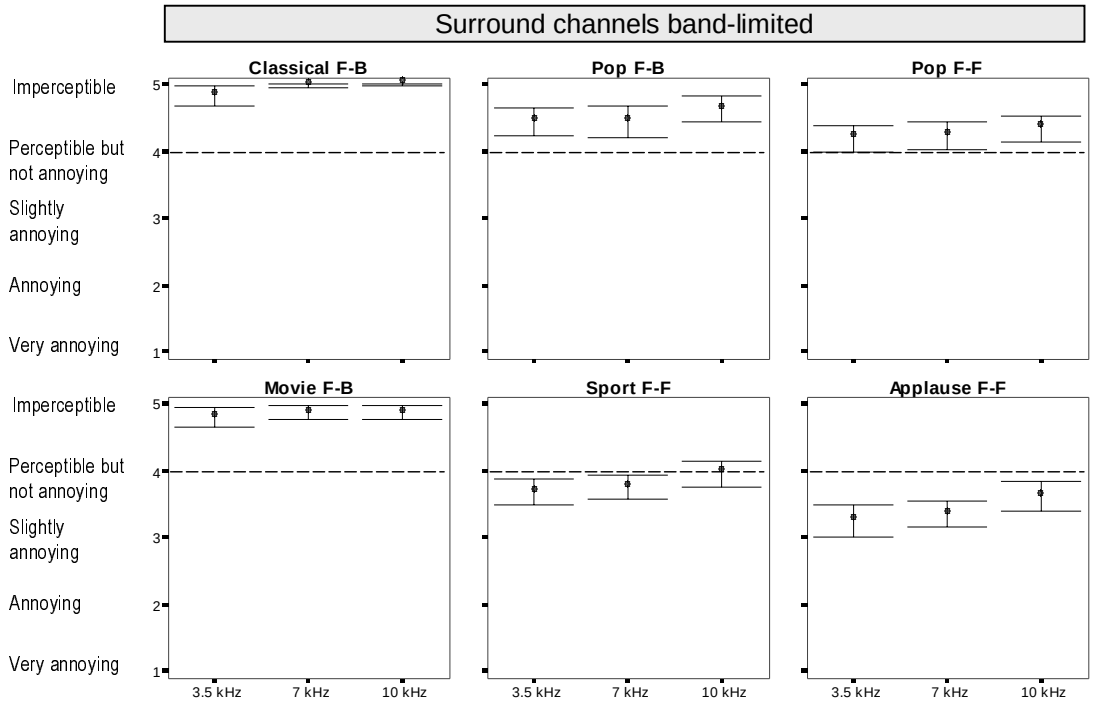


Fig. A3 Degradation of the basic audio quality caused by band-limitation of **for the surround channels**