Personal and Ubiquitous Computing
https://doi.org/10.1007/s00779-018-1125-6

ORIGINAL ARTICLE

Exploring object-based content adaptation for mobile audio

Tim Walton[1] · Michael Evans[2] · David Kirk[3] · Frank Melchior[2]

Received: 20 February 2017 / Accepted: 14 February 2018

Abstract
Mobile devices enable users to consume media with audio content in a wide range of contexts, with environmental noise being present in many of these. Several methods exist that aim to improve the experience of mobile listening by utilising information about the environmental noise, such as volume and dynamic range adaptation. This paper explores a fundamentally different approach to improving the mobile listening experience by using the object-based audio paradigm, where individual audio sources are mixed in response to each specific listening context. Three experimental studies, containing both quantitative and qualitative aspects, are presented which investigate whether environmental noise influences preference of background-foreground audio object balance in a mix. The results indicate that environmental noise can influence the preferred audio mix and that the nature of the adaptations made is dependent upon both audio content and user. Additionally, qualitative analysis provides an understanding of the role of environmental noise on preferred audio mix. It is believed that the content adaptation method explored in this paper is a simple yet useful tool for adapting content to suit both the context and the user.

Keywords Object-based audio · Broadcasting · Personalisation · Environmental noise · Context · Headphones

1 Introduction

Mobile devices such as smartphones, tablets and laptops are playing an increasingly prominent role in the consumption of broadcast media with audio content [17]. One consequence of this is that such media is being consumed in a wide range of contexts. Whether it be in a café or on a train, the characteristics of the context of use are likely to influence the quality of experience for the consumer. In a traditional channel-based broadcast chain, the same audio content is delivered to all users regardless of factors such as environment, preference and device. In an object-based broadcasting chain, however, there is the possibility for content to be adapted at the receiver end so as to account for such factors. This is achieved by representing audio content as separate objects with corresponding metadata and then rendering these objects at the receiver end [11] (see Section 1.2). With object-based audio, it may therefore be possible to adapt an audio mix at the point of consumption so as to improve the listening experience when in noisy environments. The work presented in this paper explores this idea through three experimental studies. More specifically, these studies investigate the relationship between environmental noise and preferred background-foreground audio object balance for headphone listening.

The first of these experiments is a laboratory-based listening test in which 22 participants were required to adjust the mix of audio content to their preference, whilst in the presence of reproduced environmental noise conditions. This work was previously presented in [26]. The second experiment is of a similar nature to the first, a laboratory-based listening test with 22 participants. In addition to the adjustment task (with different experimental conditions to the first experiment), semi-structured interviews were conducted to probe the participants' quantitative results. Finally, the third experiment is a web-based listening test in which there were 50 participants. This experiment consisted of both listening tasks and qualitative survey questions. By considering the qualitative and qualitative data spanning the three experiments, an understanding of both how and why users adjust audio mixes in noisy environments has been developed.

✉ Tim Walton
t.walton3@ncl.ac.uk

[1] Open Lab, Newcastle University, Newcastle upon Tyne, UK

[2] BBC Research and Development, Salford, UK

[3] Northumbria University, Newcastle upon Tyne, UK

Published online: 27 February 2018

1.1 Audio adaptation for mobile listening

Adapting audio to improve the listening experience from mobile devices is not a novel concept in itself and, indeed, a range of approaches addressing this topic have previously been presented. On the one hand, there are methods that aim to improve the audio quality from loudspeakers incorporated into mobile devices. Such methods include low-frequency enhancement [4, 25], spatialisation [4, 6], dialogue clarity improvement [5], dynamics processing [5, 25] and linearisation of frequency response [4, 5, 25]. These methods attempt to compensate for the constraints of the devices themselves and, with the exception of [5], do not aim to account for the listening environment.

Other approaches to improve the listening experience from mobile devices focus on headphone listening and, more specifically, how to reduce the impact of environmental noise on the audio experience. It is these methods that are most relevant to this study. Active noise-cancelling headphones are perhaps the most well-known technology to address this issue and are commonplace in the aviation industry [15] as well as more recently appearing in consumer devices for recreational mobile listening. The complexity of noise-cancelling headphones, however, means that they are generally expensive and, furthermore, noise cancellation may not be desirable in all mobile listening situations. A simpler approach can be found with volume-based and dynamic range-based methods. Such methods utilise inbuilt microphones on mobile devices to monitor the environmental noise level of the listening environment and adjust the volume or dynamic range of the audio content accordingly (for example, see [10, 13, 19]).

The concept presented in this paper is a fundamentally different one to those discussed in the previous paragraph. In the previous approaches, the audio content is modified as a whole; the entirety of the content is processed with the same volume-based or dynamic range-based algorithms. This has the advantage of simplicity—no information is needed about the content being reproduced. A drawback of such methods, however, is that there is limited flexibility with regard to accounting for personal preference, listening context and the content. In the approach presented here, it is the audio mix that is modified at the point of consumption to account for environmental noise conditions. Such an approach requires more metadata from the audio content, but in return a higher degree of flexibility can be achieved. It is the object-based audio paradigm that allows for such an approach.

1.2 Object-based audio

Object-based audio is a method of representing audio content as separate elements (or ‘objects’) with corresponding temporal, positional and other/semantic metadata, which are then rendered at the receiver end. This paradigm is linked to advances in spatial audio reproduction [8, 11, 14] as, unlike traditional channel-based methods, the reconstruction of a virtual sound scene can be optimised to a given reproduction setup or listening environment [23]. Along with the advantages it brings to spatial audio reproduction, object-based audio offers possibilities for greater personalisation, interaction and adaptation of content [1] [18]. Such content adaptation includes that to suit the device, adaptation to suit the environment and adaptation to suit the user.

For example, Mann et al. investigated the benefits of using object-based audio for a live football broadcast [12]. In their experiment, listeners were given a commentary feed along with audio feeds from each end of the stadium and were able to adjust the mix to suit their preference. It was found that approximately three quarters of listeners preferred the object-based experience compared to traditional radio coverage. Other examples of object-based experiences include mix adaptation for hearing-impaired listeners [22], visual content adaptation to suit the user’s profile [7] and adapting the length of audio content to suit the user’s requirements [1].

1.3 Background/foreground object distinction

It is clear that object-based audio allows for a wide range of content adaptations to be made at the receiver end. One of the simplest of these is to adjust the relative levels (balance) of sounds in a mix. Grouping audio objects into categories simplifies this process further and several categorisations of audio objects are found in the literature. In the context of spatial audio evaluation, Rumsey et al. [20] distinguish between background components consisting of diffuse or environment-related aspects of the scene, and foreground components consisting of localisable objects. In the context of television audio for hearing-impaired users, Shirley and Oldfield [22] propose three categories of audio objects—speech content whose comprehension is critical, background noise that has been shown to be detrimental to both clarity and to perceived overall sound quality, and other non-speech sounds that are considered important to comprehension and/or enjoyment of the material. In a more complex categorisation of broadcast audio objects, Woodcock et al. [27] used hierarchical agglomerative clustering to identify seven general categories, which relate to sounds indicating actions and movement, continuous and transient background sound, clear speech, non-diegetic music and effects, sounds indicating the presence of people, and prominent attention grabbing transient sounds. In the studies presented in this paper, a simple background/foreground categorisation is used; foreground objects are important to the narrative and generally localisable whereas background objects are non-critical to the narrative and generally more diffuse.

1.4 Overview

As more people consume broadcast audio content over headphones on mobile devices, it is becoming increasingly important to consider how to improve the listening experience in mobile listening contexts. Existing approaches include volume-based and dynamic range-based methods in which the audio output of the device is adapted in its entirety, in relation to environmental noise levels. The object-based audio paradigm allows for a different approach. With object-based audio, it is possible to adapt a mix of audio content on the user's device to suit the user's context. In the studies presented here, we explore this topic by investigating how environmental noise influences preference of background-foreground audio object balance for headphone listening. In the following sections, three experimental studies are presented individually before they are discussed collectively in Section 6.

2 Study 1: does environmental noise influence preference of background-foreground audio balance?

The aim of the first experiment was to investigate whether environmental noise has a significant effect on preferred background-foreground audio balance. A laboratory-based study was conducted in which participants adjusted the background-foreground balance of audio content to their preference whilst environmental noise was reproduced via a 3-D loudspeaker setup. Results from this experiment have previously been presented in [26] and the reader is referred here for further details.

2.1 Method

Participants made background-foreground adjustments for four environmental noise conditions and three audio excerpts, which were reproduced via two different methods. A brief overview of these is given below.

2.1.1 Environmental noise

In order for the results to have ecological validity, it was important to choose environmental noise clips that correspond to realistic use cases for mobile audio listening. With that in mind, two scenarios were chosen: a café-type environment and an underground train environment—both situations where mobile listening is common. Table 1 outlines the properties associated with the noise clips. As well as representing different use cases, the two clips were chosen to be spectrally different. These two environmental noise clips were used to create a total of four noise conditions: 'No noise', 'Café quiet', 'Café loud' and 'Train'. The 'Café quiet' clip was calibrated to an L_{Aeq} of 54.5 dBA, the 'Café loud' clip to an L_{Aeq} of 64.0 dBA and the 'Train' clip to an L_{Aeq} of 64.8 dBA. The calibration levels of the two café clips were chosen to be representative of realistic levels. The calibration level of the 'Train' clip was chosen to equal the level of the 'Café loud' clip; however, it should be recognised that this is possibly lower in level than in real-life situations [16]. The recordings were ambisonic B-format and were decoded to an eight-channel cube loudspeaker array.

2.1.2 Audio excerpts

Three pieces of audio content were used in this study, as described in Table 1. The genres of sport, television documentary and radio documentary were represented with excerpts that ranged in duration from 15 to 20 s so as to be long enough for judgments to be made but short enough for the constituent background and foreground levels to remain relatively constant. The foreground components consisted of dialogue and prominent sound effects whereas

Table 1 Descriptions of environmental noises and audio items used in study 1

Stimulus	Description	Background	Foreground
Café Noise	B-format recording made in a New York diner. Sounds of many conversations, distant music and occasional cutlery clatter.	–	–
Train Noise	B-format recording made on an underground train. Sounds of rumbling carriage, electric engine, screeching wheels and very distant platform announcements.	–	–
Sport	20-s excerpt of an English football broadcast	Crowd noises	Commentary
Radio Doc	15-s excerpt of a radio documentary—'The Cornish Gardner'	Music and atmospheres	Narration
TV Doc	17-s excerpt of a TV nature documentary—'Africa'	Orchestral music and effects	Narration and prominent effects

the background components included crowd noises, music and ambience.

2.1.3 Reproduction methods

The audio content was reproduced via two headphone-based methods: conventional two-channel stereo and virtual surround sound by means of dynamic binaural processing. The aim of this was to assess if preferred background-foreground balance is dependent upon headphone reproduction method. Open-back electrostatic headphones (STAX SR-207) were used. It should be noted that headphones of this brand and design are close to acoustically transparent [21]. In our own measurements, this was confirmed and it was found that average attenuation levels were similar to those of popular earbud-style headphones. A BBC R&D binaural renderer was used in combination with an optical head-tracking system (VICON Bonita) to dynamically render the virtual surround sound content.

2.1.4 Procedure

The listening test software was implemented in MAX MSP. Each page of the software represents the adjustment of one audio excerpt for one environmental noise condition and one reproduction method. Controls available to the participant were background-foreground balance, overall level and play controls for the audio excerpt. The background-foreground balance was adjustable from only background objects audible to only foreground objects audible. Both the initial background-foreground balance and overall level were randomised. The playback of the environmental noise was not controllable by the participants. It should be noted that there was no visual feedback for the adjustments so as not to influence the participants' judgments.

A repeated measures design was used so that each participant had to make adjustments for all conditions. Two noise conditions were repeated so as to enable analysis of participant consistency; therefore, a total of 36 adjustments were made per participant. The experiment was split into two sessions of 18 adjustments corresponding to each of the two reproduction methods, with a short break in between. The order of reproduction method was balanced across participants. For each reproduction method, adjustments were grouped by the environmental noise condition, that is the three audio excerpts were adjusted for each environmental noise condition in succession. When a new environmental noise began, participants had to wait for at least 20 s before playing the audio excerpts so as to familiarise themselves with the environmental noise. The order of the environmental noise and audio excerpt conditions were randomised. Before the main rating sessions, a familiarisation stage allowed participants to explore the interface and to listen to the three audio items without environmental noise.

2.1.5 Participants

A total of 22 participants (age range 19–45; mean 28; gender 11 male, 11 female) participated in this study. All participants self-reported normal hearing, were fluent in English and could be classed as naïve listeners, that is they were not professionals in the field of audio and had no or little experience of critical listening tests. When asked about their mobile listening habits, all participants reported that they listen to audio content from a mobile device with headphones at least monthly, with 77% reporting that they listen to audio content from a mobile device with headphones everyday.

2.2 Results

The recorded data consisted of foreground-background ratio[1] (FG-BG ratio) and overall level values for every condition. An initial analysis of the data revealed a number of extreme outliers (2.6% of all data), defined as values which lie outside three times the interquartile range (IQR). It was suspected that these were due to an error in the listening test software so were excluded (see [26] for a more thorough discussion of this).

Due to the removal of outliers, a repeated measure analysis of variance could not be used to analyse the interaction and statistical significance of variables. Instead, a linear mixed model analysis was conducted for both the ratio and level data. As fixed effects in the model, variables System (reproduction method), Content (audio content) and Noise (environmental noise) including all interactions were used. To account for differences between individuals, variable Participant was used as a random effect in the model, including intercepts. Visual inspection of residual plots did not reveal any obvious deviations from homoscedasticity or normality.

2.2.1 Level

Firstly, the model was calculated with level as the dependent variable. Type III tests of fixed effects revealed that the main effects of System, Content and Noise were statistically significant ($p < .05$), as well as the interaction System × Content (see Table 2). In particular, we are interested in how the environmental noise influences the level and therefore post hoc Bonferroni-corrected pairwise comparisons of the significant factor Noise were calculated. From Fig. 1, it is seen that, as expected, participants

[1] In [26], this is referred to as background-foreground ratio.

Table 2 Statistically significant type III fixed effects for dependent variable level and ratio

Variable	Source	*df*	*F*	*p*
Level	System	1 / 475.07	133.11	<.001
	Content	2 / 475.05	6.37	.002
	Noise	3 / 475.04	212.87	<.001
	System × Content	2 / 475.04	3.09	.046
Ratio	System	1 / 460.00	5.37	.021
	Content	2 / 459.30	34.78	<.001
	Noise	3 / 459.50	4.61	.003

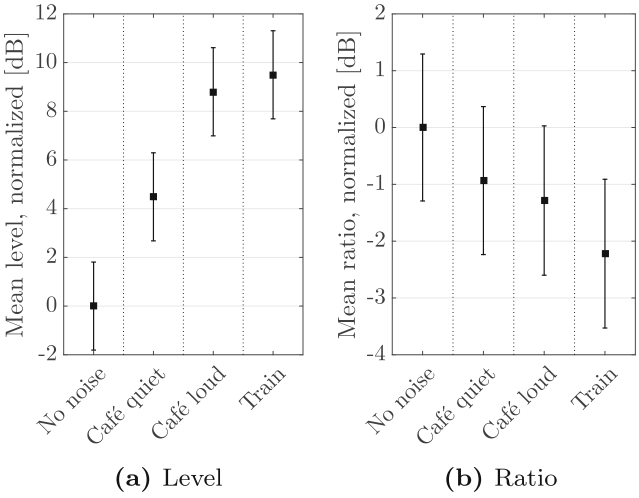
increased the level according to the environmental noise. This is consistent with other studies, e.g. [3]. The spectral characteristics of the noise did not have a significant influence on the level, as seen by the two noise items 'Café loud' and 'Train', which are not significantly different.

2.2.2 Ratio

For FG-BG ratio as the dependent variable, the main effects of System, Content and Noise were all statistically significant (see Table 2); however, none of the interactions were. This therefore suggests that participants preferred different FG-BG ratios for the different noise conditions, the different systems and also the different pieces of audio content. To investigate how environmental noise influenced the preferred FG-BG ratio, the mean ratio with respect to environmental noise was examined (Fig. 1). In other words, this is the mean ratio averaged over both content and system. Post hoc Bonferroni-corrected pairwise comparisons of the significant factor Noise reveal a significant difference between conditions 'No noise' and 'Train' ($p = .02$). Interestingly, the noise conditions have a mean ratio that is negative, which represents an increased background level in comparison to the 'No noise' condition.

2.2.3 Cluster analysis

Cluster analysis is a statistical analysis method whereby objects are classified into clusters that share similar properties (see [24] for an example in the context of audio preference tests). In order to investigate possible clustering in the ratio data, a *k*-means clustering algorithm was applied to the normalized ratio data as a whole with two clusters and a simple Euclidean distance measure. The two clusters (C1 and C2) consisted of 13 and 9 participants respectively. It should be noted that when clustering into three or four clusters, the majority of participants were still grouped into two clusters. The ratio patterns of the two clusters can be observed in Fig. 2. Cluster C1 consists of the same subset of participants for all noise types and likewise for cluster C2. It is seen that participants in cluster 1 adjusted the FG-BG ratio towards higher background levels in the presence of environmental noise, whereas participants in cluster 2 slightly increased the foreground levels or kept the ratio the same. This cluster analysis suggests that FG-BG ratio adjustment is very much down to personal preference and results should not be generalised.



(a) Level (b) Ratio

Fig. 1 Mean level (**a**) and ratio (**b**) with respect to environmental noise, normalized to the mean of the 'No noise' item. A positive ratio represents increased foreground levels whereas a negative ratio represents increased background levels. Error bars show 95% confidence intervals

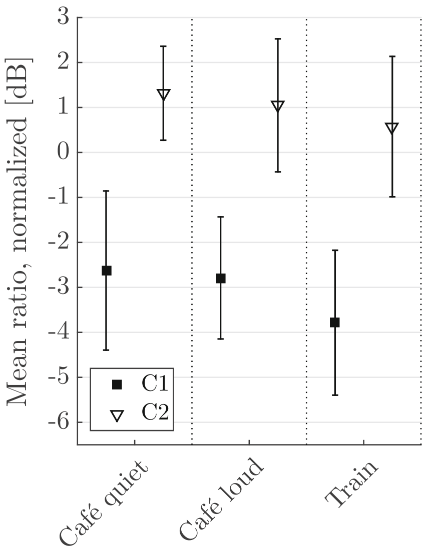


Fig. 2 Mean ratio with respect to environmental noise for each cluster of participants, normalized to the 'No noise' item. Error bars show 95% confidence intervals

2.3 Summary

The results presented in this section have illustrated that environmental noise can significantly influence preferred FG-BG ratio. It was shown that, when in the presence of environmental noise, the majority of participants adjusted the FG-BG ratio towards higher background audio levels compared to adjustments made without environmental noise. Results from the cluster analysis of the participants suggest that the mean results should not be taken as representative of the whole population. A cluster representing 41% of the participants preferred unchanged to increased foreground ratios in the presence of environmental noise.

3 Study 2: the relationship between environmental noise level and preferred background-foreground audio object balance

Results from study 1 indicate that it may be possible to adapt object-based content in order to improve the listening experience in noisy environments. The large variance in results highlights the personal nature of the adjustments and, indeed, cluster analysis indicated that participants may be grouped according to their preferences. The aim of the second study was to expand the results from study 1 with a focus on two points. Firstly, a greater range of environmental noise levels were used in order to investigate whether the trends seen in study 1 continue for higher noise levels. Secondly, a qualitative aspect was added in the form of semi-structured interviews in order to probe the clustering of participant responses.

3.1 Method

The method used was similar to study 1—a laboratory-based study in which participants adjusted the background-foreground balance of audio content to their preference whilst environmental noise was reproduced via a 3-D loudspeaker setup. Participants made adjustments for nine environmental noise conditions and three audio excerpts. Unlike study 1 where two headphone reproduction methods were compared, only stereo headphone reproduction was considered.

3.1.1 Environmental noise

As in study 1, two types of environmental noise were used. The ‘Train’ recording from study 1 was included; however, the ‘Café’ recording from study 1 was replaced with ‘Crowd’. This was due to the ‘Café’ recording sounding unnatural at high sound pressure levels. The ‘Crowd’ recording had similar spectral properties to the ‘Café’ recording but was recorded in an environment with naturally higher sound pressure levels. These two recordings were reproduced at 65, 70, 75 and 80 dBA. A ‘No noise’ condition was also included. The recordings were ambisonic B-format and were decoded to an eight-channel cube loudspeaker array.

3.1.2 Audio excerpts

The same three audio excerpts were used as in study 1: ‘Sport’, ‘Radio Doc’ and ‘TV Doc’ (see Section 2.1.2).

3.1.3 Reproduction methods

The headphones used in this study, AKG K702 open-back headphones, differ to those used in study 1. Whereas those used in study 1 were close to acoustically transparent in a similar nature to earbuds, these had attenuation properties of typical over-ear headphones. When comparing measured attenuation levels made with broadband white noise, the headphones in this study had approximately 8 dB more attenuation than those in study 1. However, when comparing attenuation differences using the specific environmental noise recordings, the attenuation differences were only 0.7 and 1.9 dB for the ‘Train’ noise and ‘Crowd’ noise respectively.

3.1.4 Procedure

The procedure was similar to that outlined in Section 2.1.4. From comments made in the first study, the interface was modified by replacing ratio and level controls with a control for background level and a control for foreground level. Again, 36 adjustments were made per participant split into two sessions of 18 with a short break in between. Three environmental noise conditions (‘No noise’, ‘Café 65’ and ‘Café 75’) were repeated in order to assess participant consistency.

After the adjustment session, semi-structured interviews were conducted. Topics for discussion included the difficulty of the experiment, how the environmental noise was perceived, how the environmental noise influenced the mix and the mixing process. These informal discussions were audio recorded and transcribed for analysis.

3.1.5 Participants

A total of 22 participants (age range 21–42; mean 27; gender 15 male, 7 female) participated in this study. All participants self-reported normal hearing, were fluent in English and could be classed as naïve listeners, that is they were not professionals in the field of audio and had no or little experience of critical listening tests. As with study 1, 77% of participants reported that they listen to audio content from a mobile device with headphones everyday.

3.2 Results

The recorded data consisted of background and foreground levels for each condition. From this, FG-BG ratio could be calculated as well as the overall level (the sum of background and foreground levels).

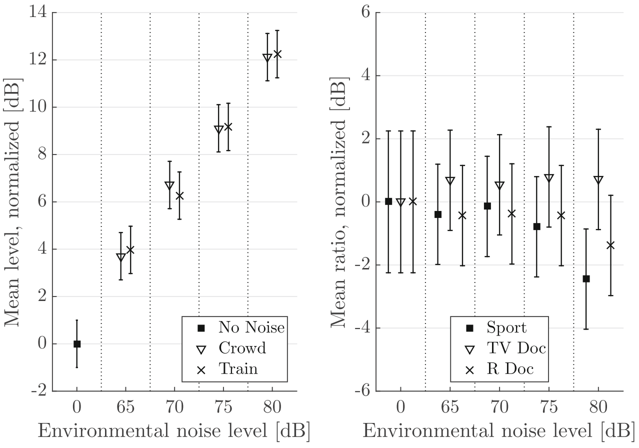
Participant consistency was checked by calculating the mean variance between repeated adjustments. For each participant, a mean variance was calculated for both the ratio and overall level data. From this analysis, it was seen that the mean level variance was 2.6 dB and the mean ratio variance was 4.2 dB. Two participants had outlying variance data and these participants were therefore excluded from further analysis. Furthermore, a third participant was excluded for adjusting the background or foreground audio to the maximum possible levels on several occasions, which could therefore compromise the preferred ratio data.

3.2.1 Level

For the analysis of the level data, a linear mixed model was used. As fixed effects in the model, variables Content, Noise type and Noise level including all interactions were used. To account for differences between individuals, variable Participant was used as a random effect in the model, including intercepts. Visual inspection of residual plots did not reveal any obvious deviations from homoscedasticity or normality. Type III tests of fixed effects revealed that the main effects of Content [$F(2486) = 15.029$, $p < .001$] and Noise level [$F(3486) = 98.849$, $p < .001$] were statistically significant ($p < .05$). The remaining effects and interactions were not significant. Therefore, the level of the environmental noise significantly influenced the listening level as did the audio item, as shown in Fig. 1. The type of environmental noise however did not influence the preferred listening level. This is consistent with study 1.

3.2.2 Ratio

For FG-BG ratio as the dependent variable in the linear mixed model, Content was the only significant effect [$F(2486) = 42.164$, $p < .001$]. When investigating this further, it was seen that all three pieces of content have mean ratios significantly different from one another. Most noticeably, content 'TV Doc' was mixed with the foreground 4–5 dB louder than the other two items. Unlike study 1, environmental noise did not have a significant influence on FG-BG ratio. This can be seen in Fig. 3.



(a) Level (b) Ratio

Fig. 3 Mean level (**a**) and ratio (**b**) with respect to environmental noise, normalized to the mean of the 'No noise' item. A positive ratio represents increased foreground levels whereas a negative ratio represents increased background levels. Error bars show 95% confidence intervals

3.2.3 Cluster analysis

To further explore the variance in the ratio data, a *k*-means clustering algorithm was applied to the normalized ratio data as a whole with a simple Euclidean distance measure, as in study 1. When using two clusters, the data was split into one cluster containing 18 participants and another containing 1 participant. With three clusters, this changed to clusters containing 11, 7 and 1 participants. The ratio data from the two main clusters from three-way clustering was further investigated (Fig. 4). 'C1' and 'C2'

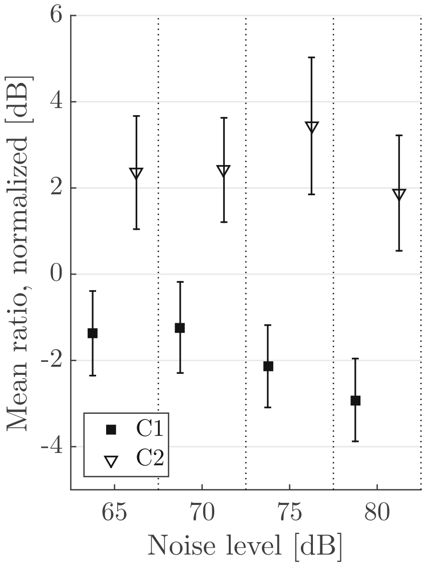


Fig. 4 Mean ratio with respect to environmental noise level for the two main clusters of participants, normalized to the 'No noise' item. Error bars show 95% confidence intervals

represent clusters with 11 and 7 participants respectively. In 'C1', a trend is seen towards increased background levels at higher environmental noise levels, whereas the FG-BG ratios in 'C2' are towards higher foreground levels. This, along with the proportion of participants found in each cluster, is consistent with the results from the first study.

3.2.4 Nature of ratio adjustments

FG-BG ratio adjustments can result from increasing or decreasing either foreground or background levels. To gain more insight into the nature of the ratio adjustments, background level and foreground level were plotted with respect to FG-BG ratio. A linear regression was then applied to these plots; the results of which are presented in Fig. 5. It is seen that foreground levels are more constant with respect to ratio than background levels. At high environmental noise levels, this is even more prominent. This suggests that ratio adjustments are primarily a result of changing background levels, especially at high environmental noise levels. The reason foreground levels are relatively constant with respect to ratio is likely due to dialogue intelligibility issues.

3.2.5 Semi-structured interviews

The qualitative data from the semi-structured interviews are considered with qualitative data from study 3 in Section 5.

3.3 Summary

The aim of the second study was to further investigate the relationship between environmental noise level and preferred background-foreground audio balance. Unlike study 1, FG-BG ratio was not statistically significant with respect to the environmental noise conditions. It is believed that this is due to the large variance in the ratio data, caused by the range of preferences highlighted by the clustering. The clustered ratio data was similar to that from study 1. From an analysis of the nature of the ratio adjustments, it was seen that at high environmental noise levels, ratio adjustments are predominantly due to changes in the levels of the background components. Semi-structured interviews were conducted and results from this are discussed in Section 5.

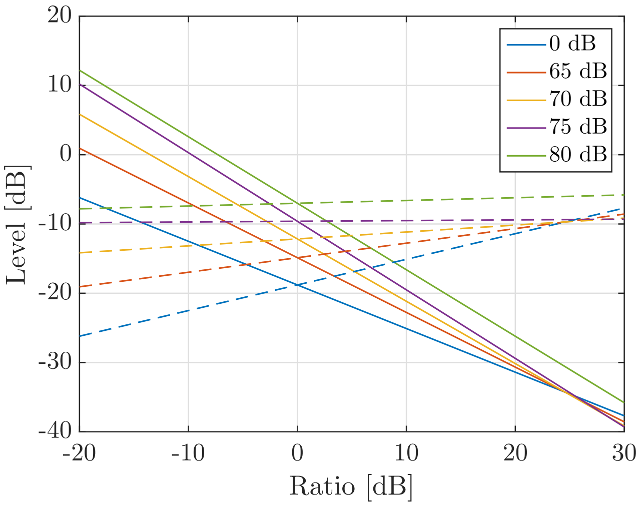


Fig. 5 Component level vs ratio for different noise levels. Dashed lines represent FG levels and solid lines represent BG levels

4 Study 3: preferred background-foreground balance in the presence of environmental noise: an online listening test

Whereas the results from study 1 indicated that environmental noise does have an influence on preferred background-foreground balance, those from study 2 are inconclusive due to large variances. This is possibly due to the method of adjustment and the relatively small sample size. In study 3, a web-based listening test with a multiple comparison method was conducted with the aim of reducing this variance. As the test was web based, a larger sample size could be achieved and the method used meant that the task was simpler for participants.

4.1 Method

On each page of the web-based interface, participants listened to five mixes of the same audio content (with different FG-BG ratios) and were required to select their most preferred mix. This was done with and without environmental noise mixed into the audio files. The study consisted of two sessions—the second being optional. In each of these sessions, one piece of audio content and one type of environmental noise were used. This design minimised the duration of the study so that each session took approximately 10 min to complete. Specifics of these variables are outlined in the following sections.

4.1.1 Audio excerpts

Two audio excerpts were used in this study: 'Sport' and 'Doc'. These were of a similar nature to the previous sport and documentary clips described. The duration of these were 24 and 16 s respectively. Mixes of these two items were made with FG-BG ratios of ± 9, ± 4.5 and 0 dB. These ratios were decided upon from a combination of examining the previous results from studies 1 and 2 and choosing ratios that the majority of untrained listeners should be able to differentiate between. It should be noted that these ratios were achieved by keeping the level of the foreground components constant and adjusting the level of the background components. The reasoning behind this was that at high environmental noise levels in the previous study, the change in FG-BG ratios came predominantly from changing background component levels, as discussed in Section 3.2.4.

4.1.2 Environmental noise

The two environmental noise clips used in study 1 were also used in this study. The ‘Café’ clip was used in combination with the ‘Sport’ content and the ‘Train’ clip was used in combination with the ‘Doc’ content. The environmental noise was trimmed to match the duration of the audio excerpts. As participants’ listening levels could not be calibrated, instead of a fixed absolute level, the environmental noise level was set as a fixed signal-to-noise ratio (SNR) in relation to the audio excerpts. The appropriate SNR values were calculated from the mean listening levels for the 80-dB environmental noise condition in study 2, taking into account headphone attenuation of the environmental noise. The SNR used was 2.3 dB.

Whereas the previous two studies delivered the environmental noise through loudspeakers, the web-based nature of this study meant that the environmental noise was delivered through headphones. In order to reproduce the spatial information from the environmental noise, the ambisonic reproduction of the noise clips was recorded binaurally using a Neumann KU 100 dummy head. Additionally, AKG K702 open-back headphones were placed on the dummy head in order to include the attenuation effects of the headphones, which were present in the previous studies. The binaural environmental noise was embedded into the audio content files at the relevant SNR.

4.1.3 Procedure

The study was implemented using the Web Audio Evaluation Tool [9], a browser-based listening test environment based on the HTML5 Web Audio API. After reading the introduction and instructions, participants were presented with a familiarisation page in which a sample of the foreground content (i.e. dialogue) was played. Participants were asked to adjust their listening level to a comfortable volume and it was additionally stated that, if possible, they should keep this level constant throughout the test. The subsequent four pages were the rating pages: without environmental noise and with environmental noise pages plus repeats. Instructions to the participants were “Imagine you are at home watching a nature documentary. Switch between the mixes below and select the mix that you would most prefer in this situation” with the content and environment descriptions changed accordingly. After the rating pages, participants were asked several qualitative questions including ‘Do you feel that the café/train noise influenced your preferred mix?’, ‘If so, in what way?’ and ‘Please enter any comments about why you think you changed your preferred mix’.

4.1.4 Participants

Fifty participants completed the first session of the study with 37 of these going on to complete the optional second session. Demographic data was not collected so as to minimise personal data collection and also to reduce the duration of the study. Participants were recruited from a range of platforms including social media and company mailing lists.

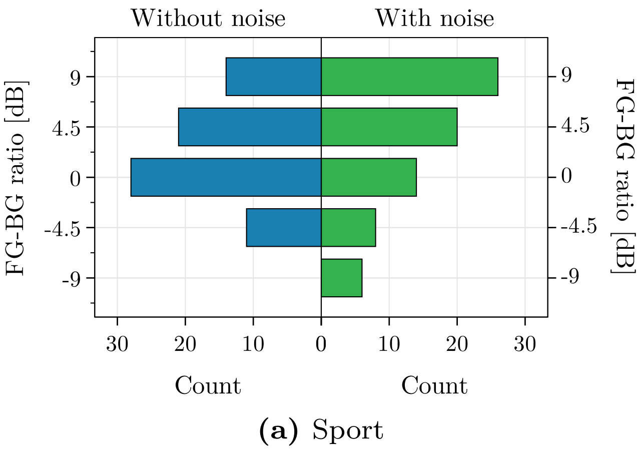
4.2 Results

Participant variance was analysed by comparing repeat choices for each condition (without/with noise). It was decided to exclude participants with a repeat variation of greater than 4.5 dB for either condition (4.5 dB being the smallest step in FG-BG ratio). For ‘Sport’, 13 participants (26%) were excluded with 37 remaining, and for ‘Doc’, 5 participants (14%) were excluded with 32 remaining.

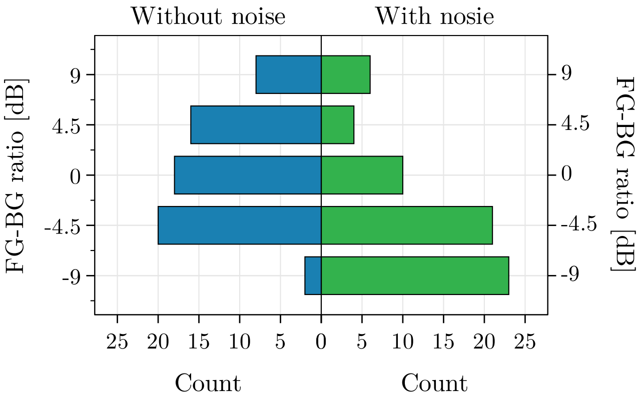
To examine the effect of environmental noise on preferred FG-BG ratio, choice histograms are plotted which show the times each mix was chosen for each condition (Fig. 6). For ‘Sport’, without environmental noise, the most chosen mix was 0 dB (FG-BG ratio) with a normal distribution around this. In the with noise condition, however, the most chosen mix was 9 dB, i.e. the mix with the lowest level of background components. It is also seen that the mix with the highest level of background components, − 9 dB, was not chosen in the without noise condition; however, in the with noise condition, it was. Therefore, the majority of participants adjusted their mix to lower background levels in the with noise condition; however, some participants increase the mix to higher background levels.

For the documentary content, without environmental noise, the most chosen mix was − 4.5 dB. Despite the preference being for high background audio levels, very few participants choose the 9-dB mix, with every other mix being more popular. With environmental noise, however, the 9-dB mix was the most preferred. This is opposite to the behaviour of the sport content which shows that the content plays a big part in how environmental noise influences the preferred mix.

To quantify the significance of the differences in the without noise/with noise choice distributions, a McNemar-Bowker Chi-square test was used. In this case, the null hypothesis is that the distribution of choices is equal in the



(a) Sport



(b) Doc

Fig. 6 Histograms of chosen mix counts for without noise and with noise conditions

without noise condition and the with noise condition. The alternative hypothesis is therefore that there is a difference in distribution of choices for the two conditions. Both the 'Sport' [$\chi^2(9, N = 74) = 18.82, p = .027$] and the 'Doc' [$\chi^2(9, N = 64) = 38.11, p < .001$] contents show significantly different without noise and with noise choice distributions.

To consider the results on an individual basis, i.e. the difference between conditions for each participant, dot plots are presented which show the mean difference between without noise choices and with noise choices for each participant (Fig. 7). As a variation of 4.5 dB was allowed between repeats and the points in Fig. 7 represent the mean between two choices, each point has a possible variance of 2.25 dB. Therefore, points within ± 2.25 dB of 0 dB have error margins spanning the three outcomes of increased background, no change and increased foreground. Despite this, the overall trends in this figure highlight what was seen in the previous choice histograms. For 'Sport', the trend is towards lower background audio levels, however, with a group of participants who preferred an increase in background levels. For 'Doc', the trend is towards higher background audio levels; however, this is not the case for everyone.

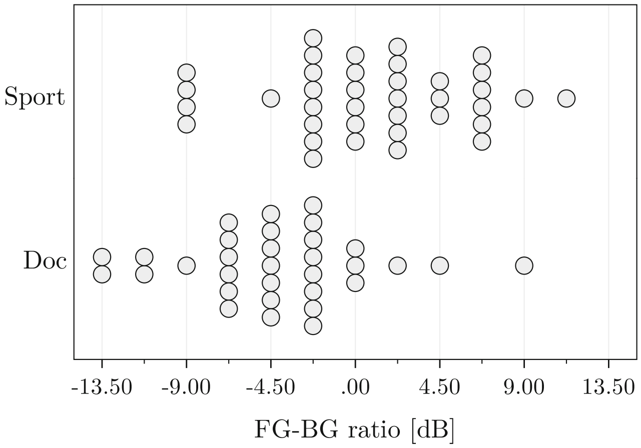


Fig. 7 Dot plots representing the mean difference between without noise choices and with noise choices for each participant

The responses from the post-test surveys are analysed in Section 5.

4.3 Summary

By conducting this study as a web-based experiment, it was possible to gather both quantitative and qualitative responses from a relatively large number of participants. This, along with the adjusted method, meant that trends seen in the ratio data were more clear than those in the previous studies, reinforcing and extending the previous results. As with the previous results, results from this study have shown that environmental noise can influence preferred background-foreground audio balance. The choice histograms presented highlight that the preferred mix in the presence of noise is very much dependent upon the audio content. As in the previous studies, it was seen that participant responses should not be generalised.

5 Qualitative analysis

In order to gain insight into why participants changed their preferred mixes in the presence of environmental noise, the qualitative data from studies 2 and 3 are now considered. In particular, we consider responses to the question 'How do you feel the environmental noise influenced your preferred mix?' from study 2 and 'Please enter any comments about why you think you changed your preferred mix' for both content items in study 3. A total of 88 responses are considered. The interviews from study 2 were audio recorded so the first step in the analysis was to transcribe these recordings. This was done using the software ‘NVivo’. In the case of study 3, the responses were typed so no transcription was necessary. Analysis of the data was based on a thematic approach [2], i.e. organising sections of the data into recurrent themes.

Figure 8 shows a schematic diagram representing the main themes identified that relate to adjusting the mix due to environmental noise. As can be seen, these themes are split into those related to increasing background components relative to foreground components and those related to increasing foreground components relative to background components.

5.1 Increased background

Thirty-five responses were related to increasing the background components of the mix relative to the foreground. A large proportion of these were comments related to increasing the background in order to mask the environmental noise.

> ‘I think sometimes I consciously went higher with the background noise to drown out the environmental noise.’

> ‘The crowd noise was more acceptable when it helped to mask the café ambient noise’

> ‘To drown out the background train noise to a greater extent.’

Several participants explicitly stated that this was due to the ‘annoying’ nature of the environmental noise. This behaviour of masking the noise is due to the characteristics of the foreground and background components. There are fewer gaps in the background components (e.g. music, crowd) than in the foreground components (e.g. speech) and as a result, an increased background can help block out the unwanted environmental noise. The results from masking the noise include increased foreground intelligibility:

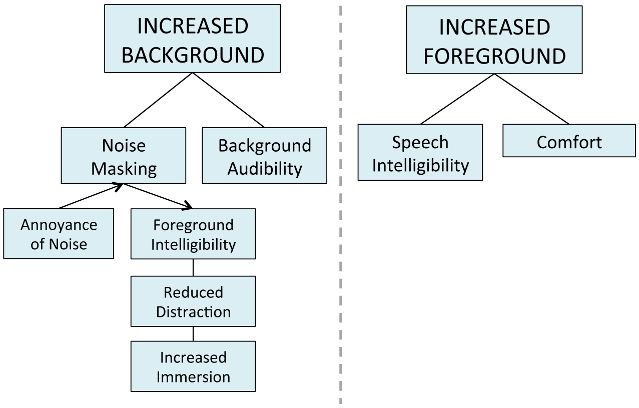


Fig. 8 Main themes identified from thematic analysis related to adjusting an audio mix due to environmental noise

> ‘It somehow actually made it easier to listen to the narration by making it more difficult to hear the train’,

reduced distraction:

> ‘The louder and more uniform and relevant background noise of the crowd prevents distraction’,

and increased immersion:

> ‘...it was definitely a case of, I feel I need to bring this up more to drown things out and to feel like I’m in the programme’.

Besides from noise masking, another prominent reason for increasing the background was audibility of the background components.

> ‘I wanted to turn [the] background music up, to be able to hear it.’

> ‘...when it’s really noisy it’s much harder to get any background so you just like get the pieces that are like kind of high, so I think I put it like higher when it was like really noisy around.’

5.2 Increased foreground

Thirty-two responses were related to increasing the foreground components of the mix relative to the background. Speech intelligibility was the primary reason for this.

> ‘In a cafe (noisy environment), sound quality and a sense of immersion is secondary to understanding the commentary.’

> ‘I wanted the narration to be clear and so I had to reduce the volume of the music...’

> ‘...when it was louder I would have the narrative on, the foreground on higher so I could concentrate more on the foreground.’

It is apparent that the environmental noise made concentrating on and understanding the foreground speech content more of a challenge.

Another reason for increasing the foreground relative to the background was due to comfort reasons. By keeping the foreground content at an intelligible level and reducing the background components, participants reduced the overall audio level.

> ‘I know there’s information being conveyed in the background sounds and I only wanna lose out on that if it’s too noisy to make it comfortable listening to the foreground sounds... in the really noisy environments

> I wanted the, I guess to keep the additional noise I'm listening to down to a minimum, so that it's comfortable.'

> 'The café... mixed with the crowd noise of the football match it becomes too noisy'

Additionally, listening comfort increased as there were less competing sounds in the mix.

> 'The background cafe noise was really irritating, and having even more noise to add into that was just even worse'

> 'The different sounds were off-putting; almost a sensory overload'

5.3 No change

A number of responses were related to participants not changing their preferred mix in the presence of environmental noise. In such responses, participants often stated that they liked a mix a certain way (e.g. '...liked the music slightly loud all the time') and that the environmental noise did not influence this. These comments were less revealing and were therefore not coded into further themes.

6 Discussion

In the three studies presented above, the influence of environmental noise on preferred background-foreground balance has been explored. In the first study, it was seen that environmental noise can significantly influence FG-BG ratio with the overall trend being towards higher background levels in the presence of environmental noise. Furthermore, the participants were clustered into two groups by their preferences. The largest cluster adjusted the FG-BG ratio to higher background levels in the presence of environmental noise whereas a second cluster preferred unchanged to increased foreground ratios.

The second study explored this further with a larger range of environmental noise stimuli and the addition of semi-structured interviews to gain a qualitative understanding of the two clusters identified in study 1. Unlike the first study, environmental noise was not seen to have a significant influence on the chosen FG-BG ratio. This lack of significance could possibly be attributed to the individual nature of the ratio adjustments, increasing variation around the mean. This could therefore indicate that the overall ratio trends seen in study 1 are not generalisable to other populations. The large variance in results was further examined with a cluster analysis and again two main clusters were identified with similar trends as in study 1. Additionally, it was seen that at high environmental noise levels, the ratio adjustments were predominantly due to adjustments of the background component levels, that is the foreground levels were approximately constant with respect to ratio.

The third study, which was web-based, aimed to reduce the variance in the data by using a simpler evaluation task and increasing the sample size. As in study 1, it was seen that environmental noise significantly influences FG-BG ratio. Moreover, the differences that audio content and environmental noise make on the preferred mix were highlighted. For the sport content, the overall trend was to adjust the FG-BG ratio to higher FG levels in the presence of noise, whereas for the documentary content, the opposite trend was seen.

The themes identified in the qualitative analysis from studies 2 and 3 revealed the different approaches taken by participants to minimise the effect of environmental noise on the overall listening experience. On the one hand, participants chose to increase the background components in order to mask the environmental noise and to ensure that the background components were audible above the noise. This noise masking from the background components in turn increased foreground intelligibility, reduced distraction and increased immersion. On the other hand, participants chose to increase the foreground components in order to improve the speech intelligibility and also the comfort of the overall experience. This dichotomy of qualitative responses is in agreement with the quantitative data.

The nature of the audio content is clearly a big factor in the responses seen. It is the continuous nature of the background components (crowd, music) that enable them to be used to mask the environmental noise. Furthermore, it is the nature of the background components that seems to influence whether the majority of participants increase or decrease their level in the presence of noise. The sport content had background components that were considered as noise and not particularly necessary by some participants. On the other hand, the documentary had background components that were considered less noise-like and more essential to the listening experience. One limitation of this study is the limited range of content used and therefore the effect of content and content-noise interaction should be further investigated.

Another point of further study could be the relationship between dynamic range adaptation and mix adaptation. Some participants mentioned that they raised the level of the background components in order to clearly hear all of the background components. Such results could also be possible by adjusting the dynamic range of the content to the environmental noise, as in [13]. The possibility of using both methods in conjunction and their relative contributions to improved listening experience should be investigated.

The findings in this paper show that with suitable object-based content, a simple background-foreground level control could be beneficial for users listening in noisy environments. As well as improving the listening experience, this could also help to prevent hearing loss; instead of increasing the level of the content as a whole, only the level of the desired components could be increased, therefore reducing exposure. Context-aware adaptation is ultimately desirable; however, the influence of content and participant on the preferred mix means that this is not a trivial task.

7 Conclusion

This paper has explored one way in which object-based audio could be used to improve the listening experience for users listening to broadcast audio content in noisy environments. Through three studies, it was shown that environmental noise has a significant influence on preferred background-foreground audio object balance in a mix and that the behaviour of the preferred adaptations is heavily influenced by content and user. In the presence of environmental noise, mix adaptations were made to both increase the background components and to increase the foreground components, depending on the content and the user. Qualitative analysis gave insight into these adjustments. Participants adjusted the FG-BG ratio towards increased background components in order to mask the environmental noise and increase background audibility, whereas participants adjusted the FG-BG ratio towards increased foreground components in order to increase speech intelligibility and overall listening comfort.

Acknowledgements The authors would like to thank John Leonard for kindly providing the environmental noise recordings and also colleagues at BBC R&D for their help with the setup.

Funding information This work was supported by an Engineering and Physical Sciences Research Council (EPSRC) industrial CASE studentship [grant number EP/L505560/1] in partnership with the British Broadcasting Corporation.

References

1. Armstrong M, Melchior F, Churnside A, Shotton M, Brooks M, Evans M (2014) Object-based broadcasting—curation, responsiveness and user experience. In: 2014 Conference on International Broadcasting Convention (IBC), pp 12.2–12.2
2. Braun V, Clarke V (2006) Using thematic analysis in psychology. Qual Res Psychol 3(2):77–101. https://doi.org/10.1191/1478088706qp063oa
3. Breinbauer HA, Anabalón JL, Gutierrez D, Cárcamo R, Olivares C, Caro J (2012) Output capabilities of personal music players and assessment of preferred listening levels of test subjects: outlining recommendations for preventing music-induced hearing loss. Laryngoscope 122(11):2549–2556
4. Cecchi S, Virgulti M, Primavera A, Piazza F, Bettarelli F, Li J (2016) Investigation on audio algorithms architecture for stereo portable devices. J Audio Eng Soc 64(1/2):75–88
5. Czyzewski A, Ciarkowski A, Kostek B, Kotus J, Lopatka K, Suchomski P (2016) Adaptive personal tuning of sound in mobile computers. J Audio Eng Soc 64(6):405–428
6. Drossos K, Mimilakis SI, Floros A, Kanellopoulos N (2012) Stereo goes mobile: spatial enhancement for short-distance loudspeaker setups. In: Eighth International Conference on Intelligent Information Hiding and Multimedia Signal Processing
7. Evans M, Ferne T, Watson Z, Melchior F, Brooks M, Stenton P, Forrester I (2016) Creating object-based experiences in the real world. In: 2016 Conference on International Broadcasting Convention (IBC)
8. Herre J, Hilpert J, Kuntz A, Plogsties J (2015) MPEG-H audio - The new standard for universal spatial/3D audio coding. J Audio Eng Soc 62(12):821–830
9. Jillings N, Moffat D, De Man B, Reiss JD (2015) Web audio evaluation tool: a browser-based listening test environment. In: 12th Sound and Music Computing Conference
10. Kean J, Johnson E, Sheffield E (2015) Study of audio loudness range for consumers in various listening modes and ambient noise levels. Online: http://www.aes.org/technical/documentDownloads.cfm?docID=523. Accessed: 22 March 2016
11. Kim C (2014) Object-based spatial audio: concept, advantages, and challenges. In: Kondoz A, Dagiuklas T (eds) 3D future internet media. Springer, New York, pp 79–84
12. Mann M, Churnside AWP, Bonney A, Melchior F (2013) Object-based audio applied to football broadcasts. In: Proceedings of the 2013 ACM International Workshop on Immersive Media Experiences ImmersiveMe ‘13. ACM, New York, pp 13–16
13. Mason A, Jillings N, Ma Z, Reiss JD, Melchior F (2015) Adaptive audio reproduction using personalized compression. In: Audio Engineering Society Conference: 57th International Conference: the Future of Audio Entertainment Technology - Cinema, Television and the Internet
14. Melchior F, Churnside A, Spors S (2012) Emerging technology trends in spatial audio. SMPTE Motion Imaging J 121(6):95–100
15. Molesworth BR, Burgess M, Kwon D (2013) The use of noise cancelling headphones to improve concurrent task performance in a noisy environment. Appl Acoust 74(1):110–115. https://doi.org/10.1016/j.apacoust.2012.06.015
16. Neitzel R, Gershon R, Zeltser M, Canton A, Akram M (2009) Noise levels associated with New York City’s mass transit systems. Am J Public Health 99(8):1393
17. Ofcom (2015) Adults’ media use and attitudes report 2016. Online: https://www.ofcom.org.uk/__data/assets/pdf_file/0026/80828/2016-adults-media-use-and-attitudes.pdf. Accessed 20 Jan 2017
18. Parmentier M (2015) Sound board: object-based audio. J Audio Eng Soc 63(7/8):659–660
19. Reis T, Carriċo L, Duarte C (2009) Mobile interaction: automatically adapting audio output to users and contexts on communication and media control scenarios. In: Stephanidis C (ed) Universal access in human-computer interaction. intelligent and ubiquitous interaction environments: 5th International Conference,

UAHCI 2009. Proceedings, Part II. Springer, Berlin, pp 384–393

20. Rumsey F, Zielinski S, Jackson P, Dewhirst M, Conetta R, George S, Bech S, Meares D (2008) QESTRAL (part 1): quality evaluation of spatial transmission and reproduction using an artificial listener. In: Audio Engineering Society Convention 125
21. Satongar D, Pike C, Lam YW, Tew AI (2015) The influence of headphones on the localization of external loudspeaker sources. J Audio Eng Soc 63(10):799–810
22. Shirley B, Oldfield R (2015) Clean audio for TV broadcast: an object-based approach for hearing-impaired viewers. J Audio Eng Soc 63(4):245–256
23. Shirley B, Oldfield R, Melchior F, Batke JM (2013) Platform independent audio. In: Media Production, Delivery and Interaction for Platform Independent Systems. Wiley, pp 130–165. https://doi.org/10.1002/9781118706350.ch4
24. Silzle A, Neugebauer B, George S, Plogsties J (2009) Binaural processing algorithms: importance of clustering analysis for preference tests. In: Audio Engineering Society Convention 126
25. Turnbull R, Hughes P, Hoare S (2008) Audio enhancement for portable device based speech applications. In: Audio Engineering Society Convention 124
26. Walton T, Evans M, Kirk D, Melchior F (2016) Does environmental noise influence preference of background-foreground audio balance? In: Audio Engineering Society Convention 141
27. Woodcock J, Davies WJ, Cox TJ, Melchior F (2016) Categorization of broadcast audio objects in complex auditory scenes. J Audio Eng Soc 64(6):380–394