Implications of Object-Based Audio Personalisation Controls For Dialogue Intelligibility and Broadcast Loudness

Zeeshan Khattak
Faculty of Computing, Engineering and the Built Environment
Birmingham City University
Birmingham, England
zeeshan.khattak@bcu.ac.uk

Waqas ur Rahman
Faculty of Computing, Engineering and the Built Environment
Birmingham City University
Birmingham, England
waqas.rahman@bcu.ac.uk

Ian Williams
Faculty of Computing, Engineering and the Built Environment
Birmingham City University
Birmingham, England
ian.williams@bcu.ac.uk

***Abstract*— Dialogue intelligibility is a persistent around of concern for broadcasters of audio-visual content, where dialogue often gets lost in a busy mix of varying audio elements, and audiences must either turn up the content to catch the dialogue or turn on subtitles. Given the current channel-based approach to mixing, broadcast content cannot be altered once a mix is finalised and broadcast, meaning the dialogue component in the mix cannot be independently increased in volume by the end-listener. This work investigates the concept of dialogue intelligibility in relation to broadcast loudness and the implications of object-based audio – a technology that enables greater audio personalisation controls over the soundtrack elements. Original test data is presented from user testing, recording the personalisation of average listeners of audio-visual content, in comparison with individuals with audio engineering backgrounds. The findings reveal that users on average set the level of the content 2.2 LUFS higher between the original and reduced-dialogue versions, emphasising the importance users place on dialogue for setting their base-level volume preferences. In the object-based phase, the audio engineer test group set the loudness higher on average by 2.8 LUFS, with the dialogue mixed 2.1 LUFS lower than the original. By contrast, the average listener group mixed the loudness very close to the original source material, with very similar loudness separation between dialogue and background content. The work concludes that the disparity between audio engineer and average listener loudness and mix preferences is a likely factor in creating inadequate mixes, where object-based audio may pose a solution to such situations.**

Keywords*— *Object-based audio, dialogue intelligibility, broadcast loudness, audio personalisation, volume surfing.

I. Introduction

Dialogue intelligibility in audio content is a persistent area of concern for broadcasters, where traditionally a single master mix of channel-based audio is transmitted alongside the video content, in which the mix engineer must compromise between a 'cinematic', multi-layered mix, and one that has better speech intelligibility [1]. Although there are other components of audio that are important for narrative intelligibility of broadcast content such as non-speech Foley and environmental sounds, the dialogue component is often expected by audiences to be prioritised over other elements in the mix, as intelligible dialogue is the key to our understanding of narrative progression [2]. Complaints to broadcasters concerning dialogue intelligibility of TV programmes and films are still commonplace, with a regular complaint relating to the dialogue levels in relation to background music or sound effects, which once turned up to improve audibility causes louder peaks to become over-amplified [3].

In the traditional channel-based audio paradigm, sounds are locally fixed in place relative to the listener via the relative amplitude and panning of audio elements in the left/right or surround-sound speaker layout, and broadcast as a final, unalterable mix. Many techniques have been developed to improve dialogue intelligibility based on the channel-based paradigm, including the transmission of multiple audio mixes of varying dialogue amplitudes, to metadata control methods using algorithms to locate and enhance dialogue content according to its relative loudness [4]. The emergence of object-based audio (OBA) is seen by some as a key proponent for improving dialogue reception, which moves the assembly of individually transmitted audio elements to the end-user of the broadcast chain, thus making possible the use of personalisation controls over the final audio mixes [5]. The level of OBA personalisation made available to the end-user is set by the broadcaster, who then decide the maximum and minimum thresholds of gain and panning to control separate elements such as dialogue and music [6].

II. Similar Studies and Contribution

Ongoing studies relating to personalisation of audio content afforded by OBA have mostly focused on the benefits it may bring to the hard-of-hearing. A paper on personalisation advancements for hearing-impaired listeners by Shirley and Ward provides a useful overview of current developments and highlights the complexity of intelligibility for broadcast content with categories of speech-to-noise ratio, spatial separation and redundancy [5]. The study by Shirley et al. on assessing personalisation controls for object-based audio for hearing-impaired viewers provides the valuable categorisation of sound elements as experienced by average listeners, to determine the optimal controls for OBA personalisation [7].

The novel approach of this paper is in addressing the link between the sound design of broadcast audio as completed by audio engineers, and the experience of average listeners who ultimately listen to the content. While Shirley et al. provided controls of 4 separate elements for volume control of the categorised elements, this project's method assesses the use of personalisation of the dialogue channel only in relation to all other background audio, in comparison with conventional channel-based volume control. The reasoning being that users may prefer more limited controls over the end mix, while focusing on what is arguably the most important sound element for narrative understanding, dialogue. The contributions of this paper are as follows:

- The importance of dialogue as the anchor for setting personal audio levels is demonstrated in the channel-based paradigm, and its implications for object-based personalisation are discussed.
- Exploring the impact of personalisation controls on end-users with OBA, the testing reveals the divide between audio engineer and average listener groups, while emphasising the unique benefits personalisation could bring.

- A novel approach to assessing 'volume surfing'; the act of turning up and down audio content to catch dialogue is presented, in which the findings reveal that reduced dialogue mixes increase the likelihood of volume surfing.

III. DIALOGUE INTELLIGIBILITY AND BROADCAST LOUDNESS

The perception of dialogue intelligibility is closely linked to the overall loudness of audio-visual (AV) content, where inconsistencies in loudness continues to be problematic for modern broadcasters [8]**.** The BBC defines the continual adjustment of volume of AV content in its guide for best practice for broadcast mixing as volume surfing (VS), in which the user increases the volume in order to catch dialogue, while later turning down music and ambient effects that become too loud as a result [9]. Dialogue is commonly labelled the 'anchor' of broadcast audio, which the viewer will listen to first in order to set their volume accordingly for comfortable listening [10].

LUFS (Loudness Units relative to Full Scale), is a measurement of the average loudness of the entire programme mix, with the aim of defining a consistent integrated level between all AV content. LUFS is calculated by averaging the short-term loudness unit (LU) readings over a period of time, while ignoring quick fluctuations in amplitude by the use of a gating threshold. The EBU standards R128 recommendation on loudness normalisation for broadcast audio signals states that the overall loudness reading of an entire mix should equal -23 LUFS [11]. This ensures a normalised reference point for broadcast mixes that allows adequate 'headroom'; space for content to peak below the maximum peak level [12]. Peaks in TV audio are important for dramatic impact, and as long as they stay within the 'comfort zone' as specified by the ITU-R BS.1770 standard of +3 to -5 LUFS around the anchor, they will be tolerated by the listener [13]. As the BBC points out, the -23 LUFS target could still be achieved with an overly dynamic mix, increasing the risk of VS and worsening the user quality of experience (QoE) [9]. As a result, there is no one solution to fit all scenarios, and often mix engineers must use a combination of LUFS target reading and subjective impressions to predict the relative loudness of dialogue in the final broadcast output, a practice that can occasionally lead to inadequate broadcast mixes [14].

IV. MASTER AUDIO MIX SPECIFICATIONS

The disparity between studio environment room acoustics where broadcast audio is being mixed compared with the spaces the content is received in may create inadequate mixes with overly high dynamic ranges [2]. Studio mix environments, especially those for high-end film and TV programmes, are often much larger spaces with low reverberation, which affects our overall tolerance to perceived audio loudness, and as a result higher dynamic mixes often occur [15]. In addition to this, there is an integral difference between mixing for cinema and mixing for home broadcast viewing. Cinema environments are suitable for wide loudness ranges due to their careful acoustical design, sound proofing and large space to reduce reverberation, which improves our auditory perception of the full range of dynamics [16]. By contrast, sound for everyday broadcasting must be mixed to a lesser dynamic range to reflect smaller listening environments with variable background noise and little acoustic dampening, which reduces our reception of quieter audio content [17].

V. OBJECT-BASED AUDIO FOR IMPROVED LOUDNESS AND DIALOGUE CONTROL

While the traditional channel-based paradigm features audio mapped to fixed speaker positions with set amplitudes, OBA does not fix the individual audio components to a specific layout when broadcast. Instead, metadata is sent along with the audio components, which are now described as objects, to explain to a playout system such as a set-top-box how to assemble each sound element in relation to the speaker layout in real-time. The advantage of this approach means users can be offered the ability to personalise audio content, including dialogue, individually and to their preferences by the altering of the metadata commands [18]. In this way, personalisation of audio content may heighten the user's QoE through added control, or improve dialogue reception for the hearing impaired or everyday listeners who may otherwise have to rely on subtitles [19]. Other than creating personalisation controls, a fully implemented OBA workflow would be beneficial as it can be used as the basis to recreate any existing systems of traditional audio playback standards from stereo to 5.1 surround sound, as well as future-proofing for newer standards including immersive 3D audio and binaural audio [20]. The drawback of OBA to existing systems is it requires a dedicated object renderer to create the metadata parameters, and a decoder to properly assemble the audio at the receiving end, meaning its future total adoption would require fundamental changes to broadcaster and end-user technology [18].

VI. PERSONALISATION GAIN CONTROLS FOR END USERS

A testing scheme was derived in order to measure the loudness mix preferences of average listeners in comparison to audio engineers (who traditionally mix the content) in a traditional channel-based audio scheme and object-based scheme. In total, twenty users were selected, with an even divide between those being average listeners and audio engineers. Participants were played a series of repeating AV material with accompanying personalisation controls, contrasting a typical channel-based scheme with a master fader, and an object-based scheme with dialogue and master controls separately [21]. Surround sound content was used as pre-existing material with adequate separation of audio elements in order to offer the subjects personalisation over the dialogue content. A range of four AV clips were selected based on the variety of content that may be viewed in the home with differing dynamic ranges and loudness, including a dialogue-heavy drama with accompanying music, a dialogue-heavy exchange with no music, a cinematic battle scene with occasional dialogue and a cinematic trailer with occasional dialogue.

Testing was carried out in 3 phases: stage 1 consisted of giving the users control over the master fader only, allowing them to increase or decrease the total volume of the audio in a way that mimics our control over content in a channel-based system. This stage was labelled the original mix (OM stage), as this was the original, unedited version of the audio mix content from the broadcaster. Stage 2 similarly only gave the users control over the master volume, however this time they were presented with clips of reduced-dialogue content, which was labelled the reduced-dialogue mix

(RDM) stage, in order to test the significance of dialogue as the anchor. Finally, stage 3 gave users access to both the master volume and a separate dialogue control fader, in order to mimic an object-based audio scenario and compare the responses against the previous two stages. This stage was labelled the object-based mix (OBM) stage.

To gather the results of personalisation in real-time, the digital audio workstation (DAW) Adobe Audition was used as the primary platform, with its in-built faders used to control the master gain, and individual dialogue control (See Fig. 1). The user's personalisation mix responses were recorded in real-time using the 'write' automation commands on the faders. The results of each user's personalisation were then exported as uncompressed WAV files in post-test and analysed using a loudness metering plugin (Youlean Meter Pro). The plugin produced measurement readouts including integrated LUFS, and the option to set target loudness limits such as EBU R128 metering. A user post-test survey was used to gather feedback relating to overall phase reception. Based on the speech transmission index (STI) subjective user measurement of dialogue intelligibility, the users were asked to rate their experience of overall audio reception, and separately the clarity of the dialogue on a scale of 1-10 for each phase [22]. Finally, each of the user's performances between each clip was rated in post-test on a scale of 0-3 for qualities of VS; 0 denoting 'no volume surfing', 3 denoting 'high volume surfing' (see Fig. 2).

Figure 1: Display on Adobe Audition with separate dialogue fader control (left) and overall master fader control (right).

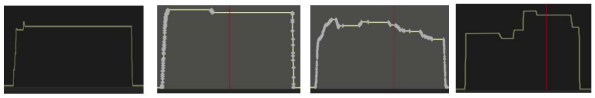

Figure 2: Visual assessment of increasing significance of volume surfing from 0 (none) to 3 (high).

The testing control measures were as follows:

- The users were first shown a reference practice clip in which they could become accustomed to the controls of the master and dialogue faders to reduce the time interval for setting faders when the clip started [23].
- Following each clip, the users were prompted to reduce all faders down to zero, with a five second count-down before each clip. The countdown phase was also put in place to provide an interlude to avoid user ear fatigue [24]. Additionally, the total test time of 16 minutes ensured that ear fatigue was not likely in users over this period of time [25].
- The sound was output through an audio interface, and a mono studio loudspeaker that was kept at a constant distance of 1m from the user's listening position. The gain level of the loudspeaker was set relative to the listening position of the user by playing the first clip and using a loudness meter to measure 90dB at peak volume, when the faders were set at 0dB digital peak level in the DAW.
- The sections of the control faders that revealed information such as short-term dB gain and visual loudness metering aids were physically covered on the monitor, to only reveal the fader tabs [26]. This prevented the participants, especially those with audio engineering backgrounds, from targeting the level of volume they thought is right for the mix based on the visual aids, rather than using their listening and preferences to set levels.
- In analysing the results, only the LUFS gain differences between the stage performances were calculated to measure the differences of user personalisation between stages, while discounting the wide variety of listening loudness base preferences of the users.

Potential limitations of the test procedure were as follows:

- Although a reduced dialogue version was created, multiple versions testing the significance of other audio elements (including Foley, sound effects and ambience) of varying loudness in the mix were not created, as this was outside the scope of the paper.
- The VS assessment was a subjective visual assessment of the severity of VS based on the visual write function automation.

VII. The Findings

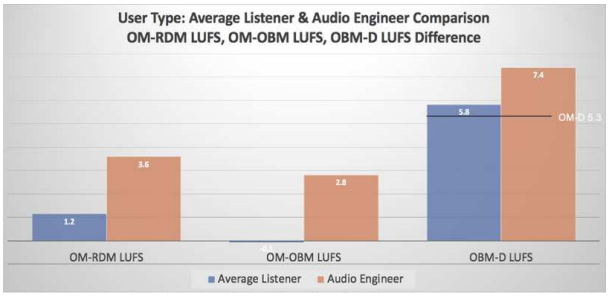

Figure 3: Comparison between average listener and audio engineer integrated loudness preferences between the mix stages. The dialogue difference in the object-based stage is marked with a black line signifying the original mix position.

As Figure 3 details, the differences between the original mix and reduced-dialogue mix (OM-RDM), and the object-

based mix and original mix (OM-OBM) were calculated between user performances, where a positive result indicates an increase in LUFS integrated loudness. Additionally, the dynamic range LUFS differences were calculated by measuring the dialogue track in isolation to the rest of the audio in the object-based mix compared with the original mix (OBM-D LUFS). As revealed, the audio engineers set their base levels for clips louder than the average listener, with a 2.4 LUFS increase in loudness between the OM and the RDM versions. When mixing the OBM, the audio engineers similarly mixed to a 2.8 LUFS increased loudness over the original mix, where the average listener mixed at a volume very similar to the OM, being 0.1 LUFS quieter. When assessing the level of dialogue in the OBM, the average listener mixed at a level much closer to the original 5.3 LUFS loudness of all the clips at 5.8 LUFS separation, while the audio engineers reduced the dialogue 2.1 LUFS from the original.

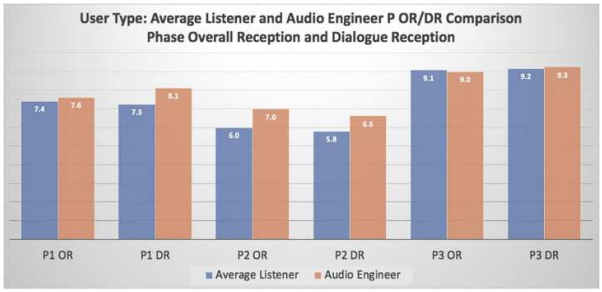

Figure 4: Comparison between average listener and audio engineer overall phase reception and dialogue reception between each stage (1-10 subjective rating of listening and personalisation experience).

In the user survey assessment of the overall mix rating and the dialogue intelligibility rating, there was a clear correlation between the ratings of the overall mix and the dialogue intelligibility. As seen in Figure 4, the average overall mix rating between testing Phase 1 (7.5), Phase 2 (6.4) and Phase 3 (9.1), shows the users accurately perceived the difference between Phase 1 and 2 reduced dialogue and rated it accordingly, while once given controls over both the dialogue and master control found the experience more satisfactory. Although similar, on the whole average listeners tended to rate the quality of audio experience as lower than that of audio engineers.

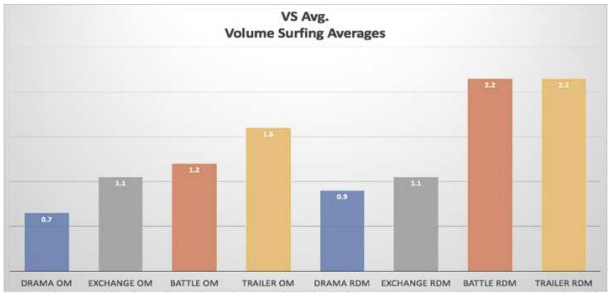

Figure 5: Mean average results of volume surfing relative levels for each of the clips, for both the original and reduced dialogue versions (0 denoting none, 1 low, 2 medium to 3 high).

As revealed by Figure 5 detailing the mean averages of VS, there is a notable correlation between the increase in VS and the version between the OM and RDM stages. Similarity these findings indicate which type of AV content is most likely to result in VS. The 'Drama' clip had an integrated loudness of -25 LUFS and was therefore accepted to be suitable for home broadcasting based on EBU and BBC broadcasting standards. This correlated with the findings with this clip featuring the least amount of VS. Additionally, the significance of integrated loudness on the amount of VS is evident from the findings. The 'Battle' and 'Trailer' clips when analysed had high integrated loudness' of -9 and -13 LUFS; well above the -23 LUFS target EBU R128 recommendations. These two clips featured the most amount of VS of 2.2, corresponding with 'medium' volume surfing.

VIII. Discussion

The results of the testing of channel-based gain controls in comparison with object-based controls for end-users proved successful in revealing overall trends in user gain personalization between average listeners and audio engineers. By measuring the difference in loudness preferences between the three stages: the original mix (OM) stage, the reduced-dialogue mix (RDM) stage and the object-based mix stage (OBM), some clear correlations between user preferences have been revealed. The OM-RDM difference stage revealed the twenty users for the variety of four clips increased the integrated loudness between the stages by a 2.2 LUFS average. Since the only variable was a drop in dialogue content in the centre channel of the 5.1 surround sound audio, this serves to confirm the principle that users base the loudness of AV content on the anchor of dialogue, which once reduced to worsen intelligibility caused them to increase the overall loudness as a result. This affected poorly on the user QoE for the RDM stage, especially by the average listener who rated it on average 16% less satisfactory. It can therefore be concluded that dialogue with poor intelligibility increases the need for VS and creates less end-user satisfaction overall.

In assessing the object-based personalisation phase, another noteworthy finding was in the loudness increase between the OBM and OM stages. Users on average increased the overall loudness of the OBM by 1.1 LUFS. This appears to reveal that with added personalisation controls to control dialogue and background content separately, users were opting to mix the content slightly louder than their original comfortable listening level in stage 1. It was however revealed in the comparison between the loudness difference of average listeners and audio engineers, that the latter preferred to mix the OBM at a level of 2.8 LUFS above the OM, while average listeners mixed very close to the OM at 0.1 LUFS below. In addition, on average mix engineers set their dialogue far below the level of the OM versions by a significant 2.1 LUFS, while average listeners mixed the dialogue at a level very close to the original. These findings highlight the inherent divide between audio engineer loudness mix preferences against the average user, who in the end makes up a much larger portion of the population who will ultimately be listening to the broadcast content.

IX. Conclusions

Given the difference of mix preferences between the two groups of audio engineers and average listeners as a whole, the findings of this paper reinforce the importance of loudness limits set by broadcasting bodies such as the EBU R128 -23 LUFS loudness standardisation, and dialogue separation of -4 LUFS as recommended by the BBC. It is concluded that the discrepancy between audio engineer loudness mix preferences compared with average end-users is a significant cause of inadequate mixes with poor dialogue intelligibility for home broadcasting scenarios. However, LUFS targets are not a complete failsafe, as can be seen in the varying instances of volume surfing depending on the content of genre of the AV material. To avoid the need for re-mastering already broadcast content that has been inadequately mixed, the benefits of an object-based approach in which average users can conveniently alter the mix to their own preferences, has been revealed.

References

[1] H. Fuchs and D. Oetting, 'Advanced Clean Audio Solution: Dialogue Enhancement', *SMPTE Motion Imaging J*, vol. 123, no. 5, pp. 23–27, 2014, doi: 10.5594/j18429.

[2] M. Thornton, 'Loudness and Dialog Intelligibility in TV Mixes - What Can We Do About TV Mixes That Are Too Cinematic?', *Pro Tools Expert*, 2019. https://www.pro-tools-expert.com/home-page/2018/8/9/loudness-and-dialog-intelligibility-in-tv-mixes-are-tv-mixes-becoming-to-cinematic (accessed May 01, 2019).

[3] G. Sergi, 'In defence of Vulgarity: the place of sound effects in the cinema', 2005.

[4] K. Lopatka, A. Czyzewski, and B. Kostek, 'Improving listeners' experience for movie playback through enhancing dialogue clarity in soundtracks', *Digital Signal Processing: A Review Journal*, vol. 48, pp. 40–49, 2016, doi: 10.1016/j.dsp.2015.08.015.

[5] L. A. Ward and B. G. Shirley, 'Personalization in object-based audio for accessibility: A review of advancements for hearing impaired listeners', *AES: Journal of the Audio Engineering Society*, vol. 67, no. 7–8, pp. 584–597, 2019, doi: 10.17743/jaes.2019.0021.

[6] M. Evans *et al.*, 'Creating Object-Based Experiences in the Real World', *SMPTE Motion Imaging J*, vol. 126, no. 6, pp. 1–7, 2017, doi: 10.5594/JMI.2017.2709859.

[7] B. Shirley, M. Meadows, F. Malak, J. Woodcock, and A. Tidball, 'Personalized object-based audio for hearing impaired TV viewers', *AES: Journal of the Audio Engineering Society*, vol. 65, no. 4, pp. 293–303, Apr. 2017, doi: 10.17743/jaes.2017.0005.

[8] Dolby, 'Broadcast Loudness Issues : The Comprehensive Dolby Approach Loudness Inconsistencies : Multiple Causes', *Dolby Laboratories, Inc.*, 2011.

[9] BBC, 'Best practice guide - sound mixing for BBC programmes', *BBC Technical Delivery*, p. 6, 2018.

[10] R. Orban, 'Using the ITU BS.1770 and CBS Loudness Meters to Measure Loudness Controller Performance', *Orban White Paper*, pp. 1–12, 2014, [Online]. Available: http://www.orban.com/white-papers/White Paper-BS.1770 vs CBS meter V3.pdf

[11] EBU, 'LOUDNESS NORMALISATION AND PERMITTED MAXIMUM LEVEL OF AUDIO SIGNALS Status: EBU Recommendation', *Ebu - R 128*, no. June, pp. 1–5, 2014.

[12] E. Tozer, *Broadcast Engineer's Reference Book*. New York: Taylor & Francis, 2012.

[13] ITU-R, 'BS SERIES: BROADCASTING SERVICE (SOUND). Method for the subjective assessment of intermediate quality level of audio systems', *ITU-R Recommendation*, vol. 1534–3, 2015.

[14] B. Shirley, L. A. Ward, L. Ward, and E. T. Chourdakis, 'Personalization of Object-based Audio for Accessibility using Narrative Importance Broadcast Accessibility using Narrative Importance View project FASCINATE View project Personalization of Object-based Audio for Accessibility using Narrative Importance', 2019. [Online]. Available: https://www.researchgate.net/publication/344786618

[15] M. Thornton, 'Are TV Mixes Becoming Too Cinematic?', *Pro Tools Expert*, 2017. https://www.pro-tools-expert.com/home-page/2016/12/21/are-tv-mixes-becoming-too-cinematic (accessed May 05, 2019).

[16] M. C. Ward, 'The Soundscape of the Cinema Theatre', *Music, Sound, and the Moving Image*, vol. 10, no. 2, pp. 135–165, 2017, doi: 10.3828/msmi.2016.8.

[17] S. Errede, 'Acoustics of Small Rooms, Home Listening Rooms, Recording Studios', *UIUC Physics 406*, pp. 1–35, 2017, [Online]. Available: https://courses.physics.illinois.edu/phys406/lecture_notes/p406pom_lecture_notes/p406pom_lect10_part2.pdf

[18] C. R. Landschoot and J.-M. Jot, "Binaural externalization processing method for object-based audio rendering," *J Acoust Soc Am*, vol. 153, no. 3_supplement, pp. A126–A126, 2023, doi: 10.1121/10.0018389.

[19] C. Cieciura, M. Glancy, and P. J. B. Jackson, "Producing Personalised Object-Based Audio-Visual Experiences: an Ethnographic Study," Association for Computing Machinery (ACM), Jun. 2023, pp. 71–82. doi: 10.1145/3573381.3596156.

[20] C. Pike, 'Presentation - Object-Based Audio In Programme Making', *BBC R&D*, 2015.

[21] B. Shirley and R. Oldfield, 'Clean Audio for TV broadcast: An Object-Based Approach for Hearing-Impaired Viewers', *Journal of the Audio Engineering Society*, vol. 63, no. 4, pp. 245–256, 2015, doi: 10.17743/jaes.2015.0017.

[22] H. J. M. Steeneken, 'The measurement of speech intelligibility', *Institute of Acoustics*, pp. 69–76, 2001.

[23] ITU-R, 'Method for the subjective assessment of intermediate quality level of coding systems (MUSHRA). Rec. ITU-R BS.1534-1 Annex 1', *Methodology*, pp. 1–18, 2003.

[24] T. Crich, Recording Tips for Engineers For Cleaner, Brighter Tracks. New York: Routledge, 2017.

[25] J. Franze, *Mixing a...z*. Nashville: Franze Music, 2008.

[26] N. Schinkel-Bielefeld, N. Lotze, and F. Nagel, 'Audio quality evaluation by experienced and inexperienced listeners', *J Acoust Soc Am*, vol. 133, no. 5, pp. 3246–3246, 2013, doi: 10.1121/1.4805210.