Audio Engineering Society

Convention Paper 10598

Presented at the 152nd Convention
2022 May, In-Person and Online

# The next generation of audio accessibility

Iain McClenaghan[1], Lawrence Pardoe[2], and Lauren Ward[3]

[1] *BBC R&D, The Lighthouse, 201 Wood Lane, London, UK*

[2] *BBC R&D, Dock House, MediaCityUK, UK*

[3] *Audiolab, University of York, York, UK*

Correspondence should be addressed to Iain McClenaghan (iain.mcclenaghan@bbc.co.uk)

## ABSTRACT

Technological advances have enabled new approaches to broadcast audio accessibility, leveraging metadata generated in production and machine learning to improve blind source separation (BSS). This work presents two contributions to accessibility knowledge: first, a quantitative comparison of two audio accessibility methods, Narrative Importance (NI) and Dolby AC-4 BSS. Secondly, an evaluation of the audio access needs of neurodivergent audiences. The paper presents two comparative studies. The first study shows that the AC-4 BSS and NI methods are ranked consistently higher for clarity of dialogue (compared to the original mix) whilst improving, or retaining, perceived quality. A second study quantifies the effect of these methods on word recognition, quality and listening effort for a cohort including normal hearing, d/Deaf, hard of hearing and neurodivergent individuals, with NI showing a significant improvement in all metrics. Surveys of participants indicated some overlap between Neurodivergent and d/Deaf and hard of hearing participants' access needs, with similar levels of subtitle usage in both groups.

## 1 Introduction

Accessible broadcast audio, through the lens of the social model of disability [1], means endeavouring to make sure all listeners have an equivalent (though not necessarily identical) experience of content. For the broadcaster, this involves meeting a range of user needs, from the permanent access needs experienced by those with hearing loss (one in five people globally [2]) or neurodivergent traits (estimated to be 15% of the UK population [3]) to temporary and situational needs, such as noise induced temporary threshold loss or viewing content in high levels of background noise. Mandated access services, like subtitling and signing, address some of these barriers. However, consistent complaints about sound and speech audibility, which are not limited to a particular broadcaster, language or country [4], indicate that significant barriers remain.

Audio accessibility approaches addressing these continued barriers range from simply turning up the dialogue in the mix, to Blind Source Separation (BSS) methods [5] and production-based techniques [6]. While previous evaluations of these technologies have shown these approaches to perform poorly [7], the advent of neural network-based BSS solutions along with the potential offered by Next Generation Audio codecs (NGA), mean that a new assessment is warranted. This paper gives an overview of state-of-the-art audio accessibility approaches, followed by two comparative studies evaluating these methods and a discussion of the implications of these results for implementing NGA.

## 2 Audio Accessibility

This paper uses a definition of audio accessibility based on the social model of disability: individuals are disabled not by impairments, but by their surroundings [1]. An accessible piece of content is then defined here as one where a user can engage with the intended experience regardless of their access needs. This puts an onus on the content creator and broadcaster to ensure that the essence of the content is available in a format the user can consume with minimal barriers.

Provision of subtitling and signed content satisfy the requirements of many of those with audio access needs. However, for those with intersectional access needs (e.g. concurrent dyslexia and hearing loss) or with a preference for making some use of the audio track, the only other action currently available is to increase the media's volume [8].

Many approaches have been proposed over the last few decades to improve the ability of audiences to easily access the key information from the audio track. These often focus on enhancement of the dialogue, either from provision of separate control of audio objects [9] or through post-hoc enhancement methods [5, 10]. The remainder of this section gives a summary of the most prominent methods, where they are applied in the broadcast chain, and their potential value.

## 2.1 Format-agnostic Methods

These methods are applied after the production process, either before transmission or during playback, and require no changes to production workflows. Since identification of speech content is done post hoc by algorithm, these methods may be more prone to introducing artefacts than during-production methods.

**Frequency Based** methods emphasise speech frequencies through filtering and have become quite widespread in consumer soundbars and televisions (e.g., Samsung Clear Voice [10], and ZVOX AccuVoice [11]). These methods are frequently found to make little improvement though and, in some cases, can actively degrade the clarity [12].

**Blind Source Separation** methods take the incoming audio stream and, utilising machine learning and signal processing techniques, predict and enhance the speech component. Early approaches showed little improvement in intelligibility, though some demonstrated reductions in listening effort [7]. Recent advances in deep neural networks and the advent of object-based audio methods have reinvigorated research in these approaches [5, 13, 14].

## 2.2 Format-specific Methods

These rely on content creators generating and transmitting the requisite assets rather than deriving them from a pre-mixed stream. Their advantage is they ensure dialogue is correctly identified and they give the content creator greater control over the final enhanced or personalised mix. They are limited by reliance on either specific reproduction equipment (5.1 surround) or additional production processes (NGA metadata acquisition).

**5.1 Centre Channel** methods leverage production norms which reserve the 5.1 surround centre channel for speech. Having this clean, spatially separated dialogue track has been shown to improve clarity ratings and intelligibility in normal hearing subjects [9]. Implementation of this approach has been limited by low user adoption of 5.1 and limited standards for channel usage in production. An algorithmic approach which downmixes 5.1 to create a centre channel, called "Center Cut", has been proposed but has not been effective in improving measured intelligibility [15].

### 2.1.1 NGA Methods

Next Generation Audio codecs allow metadata, carried alongside the audio, to define aspects of the audio reproduction. This allows multiple audio tracks to be carried, which are then mixed on the playback device according to the metadata. The listener can also be permitted to control aspects of this mixing. These capabilities make a number of dialogue enhancement techniques possible. Described in the following section are both NGA codecs which offer specific proprietary accessibility solutions, and Narrative Importance (NI), which is reliant on NGA but is not codec specific.

**Narrative Importance (NI)** takes a broader approach to audio accessibility, boosting the loudness of all the audio elements relevant to the narrative of a piece of media, not just speech [16, 17]. This is achieved through assignment of NI metadata by the content creator to each audio object, over four levels of importance. The end-user can then adjust the mix to their preference across a scale from the original mix to a fully accessible mix, defined by the gains and attenuations in Fig. 1:

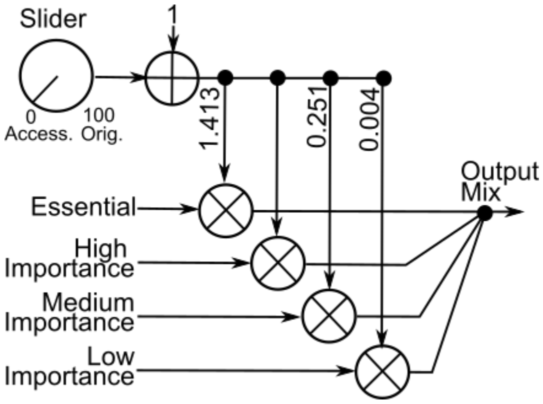


Figure 1: DSP block diagram of Narrative Importance VST implementation

This method's limitations are in its requirement of an object-based renderer and the changes to production workflows. Investigations of how this can be streamlined through semi-automation of metadata assignment using machine learning is ongoing [18].

**AC-4** is an NGA format created by Dolby Laboratories [13]. It provides a dialogue enhancement capability in which the user may specify the balance between dialogue and background during playback. This is achieved either by providing the AC-4 encoder with an object-based mix with a specified dialogue object, or through a post-hoc BSS (termed henceforth as AC-4 BSS).

**MPEG-H** is an NGA format created by Fraunhofer. In a fully object-based production, it allows the broadcaster to specify a dialogue track and the extent the user should be allowed to alter its gain. For non-object-based input Fraunhofer provide the Dialog+ tool [14], which uses BSS to separate dialogue from other audio. The output can then be used in further MPEG-H production and delivery.

Both MPEG-H and AC-4 permit gain control over multiple objects during playback when provided with an object-based input. This means they can themselves be used to deliver personalisable accessible audio using the NI method, as well as their own codec-specific methods.

**Comparing Methods**

It is evident from this overview that all state-of-the-art accessible audio methods have a trade-off between production and technological requirements and their efficacy. The remainder of this paper conducts two comparative studies comparing a subset of available methods.

## 3 Experiment One

### 3.1 Aim

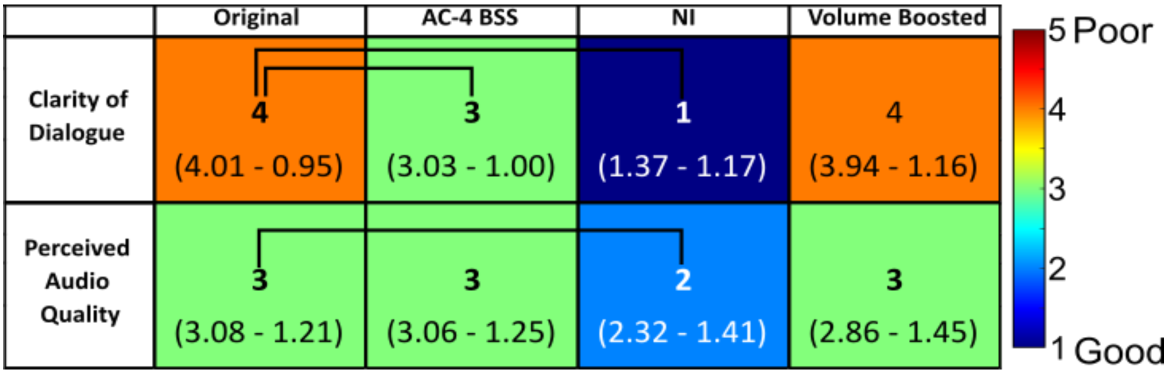
To conduct an initial study establishing whether there is an appreciable difference between a subset of available methods to audio accessibility.

Four methods were selected for comparison with the original audio, including a mix using AC-4's BSS algorithm, a NI mix and a version of the unprocessed stimuli increased by 2.4dB (based on [19], termed 'volume boosted' here). The fourth method is omitted from the reported results as it is deprecated and no longer in use. These methods were selected based on the availability of their implementation detail.

### 3.2 Methodology

Participants were asked to rank a series of unlabelled audio stimuli by 'clarity of dialogue' and 'perceived audio quality'. Participants used their own interpretation of these terms; no definitions were given. The stimuli used were 10 manually selected clips from the BBC Studios TV Drama 'Casualty' with challenging acoustic scenes. The experiment was conducted online with the end-users' own listening equipment.

Each stimulus was approximately 5 seconds long and was processed with each of the selected methods to give 5 stimuli to rank, including the original mix. In processing the stimuli with AC-4's BSS method, the pre-set setting with the largest separation between dialogue and background was selected (9dB).



| | Original | AC-4 BSS | NI | Volume Boosted |
|---|---|---|---|---|
| Clarity of Dialogue | 4<br>(4.01 - 0.95) | 3<br>(3.03 - 1.00) | 1<br>(1.37 - 1.17) | 4<br>(3.94 - 1.16) |
| Perceived Audio Quality | 3<br>(3.08 - 1.21) | 3<br>(3.06 - 1.25) | 2<br>(2.32 - 1.41) | 3<br>(2.86 - 1.45) |

5 Poor
4
3
2
1 Good

Figure 2: Median rankings for each method and metric are shown using colour and text along with mean rank and standard deviation in parentheses. Significant differences at the level of p <0.0001 are shown by braces above pairs of methods (α= 0.0167)

26 participants were recruited from BBC Research and Development staff members. No personal or demographic information was collected.

### 3.3 Results

The distribution of rankings for each method and metric were established. The median, mean and standard deviations of rankings can be seen in Fig. 2. A Wilcoxon signed rank test was then used to evaluate the null hypothesis that each method had the same distribution as the original. Comparisons are conducted pairwise with the original and as such, omission of the deprecated method has no effect on the presented results. Application of a Bonferroni correction for multiple comparisons resulted in α= 0.0167. Fig. 2 also notes which methods were significantly different to the original.
Both AC-4's BSS method and the NI method offered improvement in clarity compared to the original, with NI improving clarity by 3 ranks and AC-4's BSS by 2 ranks. One approach had a significant effect on the quality rankings, with NI improving quality by one rank.

### 3.4 Discussion

All methods which showed an improvement in clarity are carried into the next experiment. Boosted volume provided no appreciable difference in rankings of dialogue clarity compared to the original (subsequently it is omitted from the following experiment).

AC-4 BSS has a more modest effect on clarity as compared with the NI method and does so whilst retaining same quality ranking as the original audio. NI improves rankings of both quality and clarity, emphasising the advantage gained by retaining clean assets for both dialogue and other objects where possible.

## 4 Experiment Two

### 4.1 Aim

This experiment aims to evaluate the effect of two methods, AC-4 BSS and NI, on word recognition, perceived quality and listening effort of broadcast audio. In addition, it aims to assess whether there is overlap in the access needs of neurodivergent and d/Deaf and hard of hearing individuals.

### 4.2 Target populations

Participants from two cohorts were specifically recruited in addition to normal hearing listeners. The first is d/Deaf and hard of hearing audiences, who have previously been identified as most likely to benefit from expansion of audio accessibility techniques (with the majority of previous research having focused on those with age related hearing loss [20, 21]).

The second is neurodivergent individuals, including but not limited to autistic individuals, dyslexic individuals, and those with ADHD. In contrast to d/Deaf and hard of hearing listeners, the media access needs of neurodivergent individuals are largely unexplored. However, an increasing body of research suggest that neurodivergent individuals' atypical sensory processing and speech in noise perception affects their media access needs [22].

### 4.3 Methodology

To evaluate the methods, a variation on a speech in noise test was conducted. In addition to word recognition rate, two other metrics were selected: perceived quality, and self-reported listening effort. This was intended to provide a more comprehensive evaluation of the methods. These were rated on a 5-point Likert scale. As the first collection of comparative data on these methods, audio only stimuli without subtitles or visual information were used to reduce the number of experimental variables.

Additionally, participants were asked to complete a brief survey based on the TV10 (which was designed to assess an individual's hearing ability as it relates to their experience of TV audio [16]). Two questions from the TV10, about signed and foreign language content, were omitted. Four questions, extending the scenarios covered and emphasising scenarios where high levels of cognitive load might be experienced, were added. This was done to better capture the experiences of the neurodivergent participants.

#### 4.3.1 Stimuli

The target speech was 40 sentences selected from R2SPIN [23]. This dataset consists of phonetically balanced sentences spoken in a Received Pronunciation accent, with a "keyword" at the end of each sentence that participants must identify. Low predictability sentences were used, in which the final word could not be predicted from preceding words, for example, "Bob didn't know about the spoon".

Beyond speech, broadcast audio contains a variety of context clues in non-speech sound. To ensure ecological validity, sound effects related to each keyword were added to the stimuli, played after each sentence. For example, if the keyword was "hen", the sentence would be followed by the sound of a hen clucking. This was then combined with a representative background audio—either real background sounds from broadcast TV, or realistic situations synthesised using the BBC Sound Effects Archive [24]. The balance between the background and foreground elements was set by a professional sound engineer, who had been instructed to mimic the level differences found in the show Wonders of the Universe (which attracted many complaints for its mixing [25]). This ensured the stimuli represented the challenging end of the range of possible broadcast content.

All stimuli were processed with AC4 BSS, with the same 9dB difference between dialogue and background speech used. For the NI stimuli, the dialogue was assigned to 'essential', the sound effect to 'high importance' and the rest of the background to 'medium importance'. All stimuli were normalised to -23 LUFS.

Participants were only allowed to hear each clip once, to reduce learning effects, and a Latin square was used to pseudo-randomise the treatment assigned to each sentence. After giving their guess for the keyword, participants were then asked to rate the effort required to hear the word, and the quality of the audio clip, both using a 5-point Likert scale.

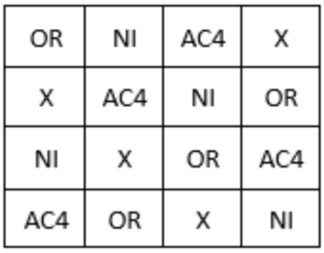
### 4.4 Respondents

Participants were recruited via the BBC, the University of York and the Leonard Cheshire charity. 30 participants took part in the listening test. 12 participants identified as Neurodivergent, 2 participants identified as d/Deaf, 9 participants identified as Hard of Hearing, and 9 participants did not identify with these terms. 2 participants declined to share whether they identified as Neurodivergent. There were 2 participants who identified as Neurodivergent as well as d/Deaf and Hard of Hearing, respectively.

### 4.5 Cleaning the Data

Word recognition responses were checked for spelling errors, with minor errors counted as a correct response (e.g. "breif" for "brief" or "oxs" for "ox"). Responses with multiple guesses were counted as incorrect.

Table 1: Latin square where OR–Original, NI–Narrative Importance, AC4–AC-4 BSS, and X–Omitted.



| | | | |
|---|---|---|---|
| OR | NI | AC4 | X |
| X | AC4 | NI | OR |
| NI | X | OR | AC4 |
| AC4 | OR | X | NI |

A Latin square (Table 1) was used to pseudo-randomize the sentence/treatment combination and their presentation order. 25% of the total sentences were allocated to each of the methods under test. The remaining 25% of sentences were used for a fourth method, whose results are omitted as the method is not publicly available. Each participant was assigned a row of the Latin square, indicating the order of the treatments, with a static order of sentences. This was to allow learning effects to be balanced between treatments and groups across the dataset.

Due to participant numbers and some incomplete responses, there was not an even distribution of combinations (all rows occurring at a frequency of 8, except row 3 which was 6). To balance the data, rows 1, 2 and 4 were under-sampled to the minimum observed frequency, reducing the effective participant number to 24. Under-sampling was carried out by discarding participants with the highest WRR on unprocessed stimuli in each group, since those responses could be considered most “saturated” on the psychometric curve.

As the stimuli utilise a fluctuating masker, the accuracy for each stimulus was investigated for ceiling effects which may mask the effects of the treatments. As many of the stimuli were found to demonstrate a very high (>90%) WRR rate, under-sampling of the sentences was conducted. For the remaining participants, the average WRR was calculated for each sentence across the treatments. Based on the Latin square allocation, each participant would receive the same treatment for every fourth stimulus. To ensure an equal number of treatments/participants were retained across the data, the sentences were split into four groups of 10 sentences, starting from sentence 1-4 and taking every fourth sentence. The three highest scoring sentences were omitted from each group, leaving a dataset of 28 stimuli with 7 iterations of each treatment for each of the remaining 24 participants.

## 4.6 Word Recognition Rate

Across each participant’s answers, Word Recognition Rates (WRR) were calculated as a ratio of correct to total answers for each method. A boxplot of the results is seen in Fig 3.

Fig. 3 shows that the distribution of WRR between AC-4 BSS and the original audio is relatively uniform. NI, for which the median is at 1.0, demonstrates a substantial increase in WRR.

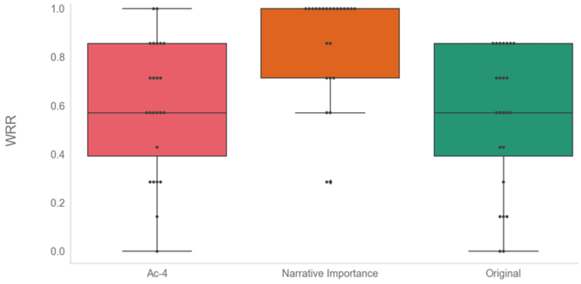


Figure 3: Boxplot of the Word Recognition Rate for each Method

The significance of the change in WRR with respect to the original audio for each treatment was calculated using a two-sided, paired t-test with a Bonferroni correction for multiple comparisons. Table 2 shows that both treatments increased the average WRR compared to the original, and the increase in WRR for NI was significant.

Table 2: WRRs of Treatments

| Method | Average WRR | Std Dev | p-value |
|---|---|---|---|
| Original | 0.55 | 0.29 | - |
| Narrative Importance (NI) | 0.86 | 0.23 | <0.0001* |
| Dolby AC-4 | 0.60 | 0.27 | 0.3733 |

* p-value < alpha = 0.025

## 4.7 Effort Rating

Participants’ ratings of the effort required to hear each sentence’s keyword are shown in the box plot in Fig. 4. Note that high ratings correspond to high levels of effort. A high level of similarity between the original audio and AC-4 BSS is seen, with a large reduction in ratings of listening effort seen for NI.

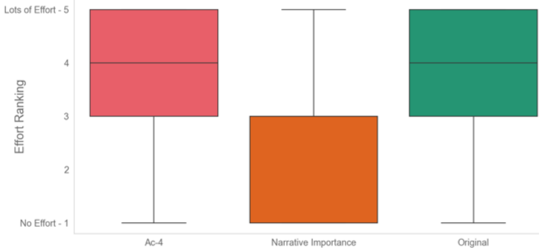


Figure 4: Distribution of Effort Ratings across Methods

A Wilcoxon Signed-Rank Test was performed pairwise between the rankings for each method and the original audio, with a Bonferroni correction applied. Table 3 shows that the decrease in effort provided by NI is highly statistically significant, showing a decrease of one median rank.

Table 3: Effort Ratings of Treatments

| Method | Median Rank | Mean Rank | p-value |
|---|---|---|---|
| Original | 4 | 3.89 | - |
| Narrative Importance (NI) | 3 | 2.50 | <0.0001* |
| Dolby AC-4 | 4 | 3.92 | 0.4587 |

* p-value < alpha = 0.025

## 4.8 Quality Ratings

Participants' ratings of the quality of each method are shown by the box plot in Fig. 5. The distribution for quality is largely similar for the Original Audio and AC-4 BSS. NI shows a substantially higher average quality rating.

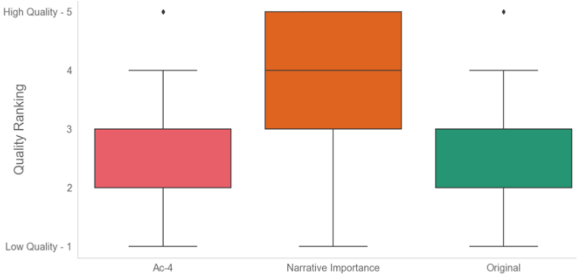


Figure 5: Distribution of Quality Ratings across Methods

A Wilcoxon Signed Rank test was again performed with the results shown in Table 4. A median increase in quality of one rank was found for NI, with a high degree of statistical significance. AC-4 BSS shows a modest but non-significant increase in mean rank.

Table 4: Quality Ratings of Treatments

| Method | Median Rank | Mean Rank | p-value |
|---|---|---|---|
| Original | 3 | 2.56 | - |
| Narrative Importance (NI) | 4 | 3.80 | <0.0001* |
| Dolby AC-4 | 2 | 2.60 | 0.6604 |

* p-value < alpha = 0.025

## 4.9 Questionnaire results

All participants responded to the survey questions (which can be seen in Table 5). A comparison of the average response values to each question can be seen in Figure 6.

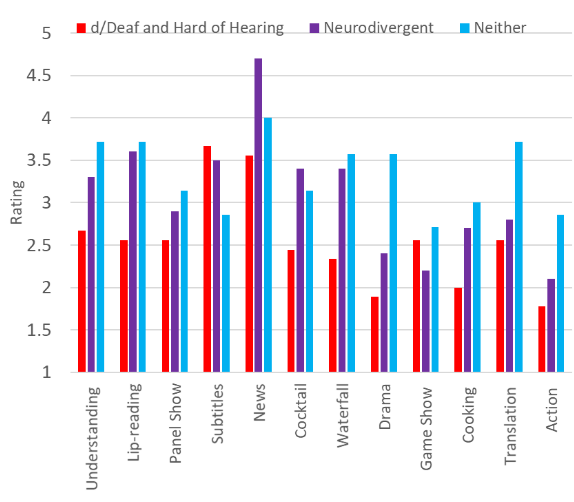


Figure 6: Average response to each of the questionnaire items grouped by Neurodivergent, d/Deaf or hard of hearing, or neither.

Four participants are omitted from this analysis: two participants who declined to respond to the neurodiversity question and two participants who identified as both neurodivergent and d/Deaf or hard of hearing. Whilst intersectional access needs are an area which require investigation, two participants deemed an insufficient sample size to provide a fair comparison with the other groups. This resulted in 10 neurodivergent participants, 9 d/Deaf or hard of hearing and 7 who identified as neither. For Understanding, Lip-Reading, Panel Show, Cocktail, Waterfall and Cooking questions the average responses from the 'Neurodivergent' group closely follow the ratings of the 'neither' group. For all these questions, the 'd/Deaf or hard of hearing' group showed consistently lower ratings (indicating greater difficulty with these scenarios). Most notable is that despite similar reported levels of speech understanding in these questions as the 'neither' group, and reporting by far the highest ease of speech understanding in quiet (News item), the 'neurodivergent' group indicate that they use subtitles almost as much as the 'd/Deaf or hard of hearing' group.

For the Drama, Translation and Action questions the difficulty reported by 'Neurodivergent' individuals is much closer to that of the 'd/Deaf or hard of hearing' group. For the Translation and Action questions, the phrasing of the question focuses the respondent more on the 'effort to follow' the overall programme rather than specifically speech understanding. Furthermore, for Gameshows the 'Neurodivergent' group report by far the greatest difficulty of any group.

A Kruskal Wallis test was performed between the three groups, with post hoc testing to determine significant pairs. Significant differences between 'neither' and 'd/Deaf or hard of hearing' at the level of $p<0.05$ were seen for Understanding, Waterfall and Drama questions. Significant differences between 'Neurodivergent' and 'Neither' at $p<0.05$ were seen for Lip-reading, News, and Waterfall. No other pairs were significantly different.

At the end of the task participants were also asked *'Would you like to be able to control the level of background sounds in TV shows?'*, with a 5-point Likert scale from 1 – Not at all to 5 – Very much so. 23 participants indicated 'Very much so', which was the median response for the groups 'd/Deaf and hard of hearing', 'Neurodivergent' and 'Neither'.

Table 5: Full survey questions

| ID | Question |
|---|---|
| Understanding | Generally, how difficult do you find it to understand speech on television? |
| Lip Reading | A character is speaking but they are not on screen. How easily can you understand the speech without seeing the character's face? |
| Panel Show | You are watching a panel show and one of the panellists is speaking whilst the studio audience laughs and cheers. How easily are you able to understand the panellist's speech? |
| Subtitles | How often do you use subtitles? |
| News | A news presenter is reporting from a quiet studio. Without using subtitles, how easily can you understand the speech? |
| Cocktail | You are watching a scene on television which has the sound of clinking glasses, music and people talking in the background. Can you make out the different sounds? |
| Waterfall | You are watching a nature documentary. The narrator is speaking with the constant sound of a waterfall in the background. Can you follow what the narrator is saying? |
| Drama | How much effort do you require to hear what is being said in a television drama? |
| Game Show | You are watching a game show which has frequent unexpected sound effects, how much effort do you require to follow what is going on? |
| Cooking | You are watching a cooking competition. There is tense background music which is getting louder and louder, how easily can you follow the show? |
| Translation | You are watching someone being interviewed on the news. They are speaking a foreign language and the English translation is playing over them. Can you follow what is being said? |
| Action | You are watching a show in which there is fast-paced dialogue, explosions, and the sound of weapons being fired. How easily can you follow the show? |

### 4.10 Discussion

The results of Section 4.6, 4.7 and 4.8 demonstrate that the NI approach benefits word recognition, perceived quality, and self-reported listening effort. This effect appears to be from both raised dialogue level and maintaining salient sound effects. This is shown by comments made by participants when giving feedback on the experiment, such as *'Context is a great help'* and *'There were words that I would never have been able to get without the audio cues'*. Overall, the usefulness of the sound effects was mentioned in 12 out of 30 participant's feedback. This suggests that methods which preserve these aspects of a sound mix may be more effective, though this would likely entail a higher production effort as a result. When clean assets are not available, these results suggest that BSS based methods offer some advantage (in mean WRR and mean quality rank). Analysis of the effect size and significance of this were limited by the size of the cohort for this experiment. However, the backward compatibility of BSS methods should motivate future work to accurately characterise the possible benefits of them.

That this experiment was conducted with users' own reproduction equipment in their homes gives confidence in the ecological validity of these results. Through the recruitment of individuals who identify as d/Deaf, Hard of Hearing and Neurodivergent as well as those who don't identify with these terms, it can be concluded that the established benefits hold for a diverse range of audiences. However, the use of audio only stimuli reduces the ecological validity of these results. Given the complexity of audio-visual processing, further research making use of audio-visual stimuli is required.

The questionnaire responses highlight that both the 'd/Deaf and hard of hearing' and 'neurodivergent' participant groups report similar levels of subtitle use and at a rate higher than the 'neither' group. For 'd/Deaf and hard of hearing individuals' the use of subtitles seems to relate to challenges with speech understanding, whilst for the 'neurodivergent' group, answers to other questions suggest that overall listening and comprehension effort form a greater part of their access needs. This questionnaire is limited in its scope, and motivates further research to understand how current and future access services might benefit neurodivergent audiences.

The responses to participant desire to control the TV audio balance mirror those from other studies [17], indicating a high level of audience desire for agency over TV reproduction. Although, the need for a range of options to be offered is highlighted by a hard of hearing participant who indicated in their feedback that they wouldn't want control and are *'comfortable with subtitles'*.

## 5 Conclusion and Future Work

This paper has summarised the advancements in audio accessibility enabled through NGA codecs and presents the first comparative investigations of these methods. The first experiment highlights the need for solutions to audio accessibility beyond just 'turning it up,' with all tested methods showing improvements in dialogue clarity, and most also in quality. Experiment two shows that the NI method significantly improves word recognition rate, perceived quality and listening effort as compared to the original mix. This result emphasises that where possible, transmission of clean assets and individual audio objects to the end user provides a better experience and is a compelling argument for NGA implementation throughout the broadcast chain.

Through the results and the review of methods, this paper shows no single approach suits all production workflows or audience members. This motivates continued research comparing these methods to allow broadcasters to make informed decisions about their implementation. A study with a larger cohort of listeners would allow more modest effects to be identified and quantified. Investigation of different genres, where the salience and complexity of non-speech sounds differs, would also help inform broadcasters.

This paper also presents initial findings on how the audio access needs of neurodivergent individuals may share commonalities with better studied audience groups like d/Deaf and hard of hearing individuals. The results highlight that subtitle usage by neurodivergent participants mirrors usage levels in d/Deaf and hard of hearing respondents, though they may be utilised for different reasons. Gaining a better understanding of the diverse range of audience access needs, as well as how audiences are utilising current access services, is an important next step in the process of developing accessible audio solutions. Finally, this work reaffirms the strongest argument for audio accessibility methods that exists – *the audience want it.*

## References

[1] Scope, "Social Model of Disability" https://www.scope.org.uk/about-us/social-model-of-disability/ (accessed: 14/2/2022).

[2] Haile, Lydia M., et al. "Hearing loss prevalence and years lived with disability, 1990–2019: findings from the Global Burden of Disease Study 2019." *The Lancet* 397.10278 (2021): 996-1009.

[3] Edinburgh University. "What is neurodiversity?": https://www.ed.ac.uk/equality-diversity/disabled-staff-support/neurodiversity-support (accessed: 14/2/2022).

[4] Mapp, P., "Intelligibility of Cinema & TV Sound Dialogue," *presented at the 141st Convention of the Audio Engineering Society* (2016 Sep.), paper 9632.

[5] Coleman P., et al. "Perceptual evaluation of blind source separation in object-based audio production." *International Conference on Latent Variable Analysis and Signal Separation*. Springer, Cham, 2018.

[6] Shirley B., et al. "Personalized object-based audio for hearing impaired TV viewers." *Journal of the Audio Engineering Society* 65.4 (2017): 293-303.

[7] Armstrong, M. "Audio Processing and Speech Intelligibility: a literature review" White Paper 190. BBC R&D, 2011.

[8] Strelcyk O, and Singh G, "TV listening and hearing aids", *PLOS ONE* 13(6): e0200083. https://doi.org/10.1371/journal.pone.0200083

[9] Shirley, B. and Kendrick, P. "The Clean Audio Project: Digital TV as Assistive Technology". *Technology and Disability*, 10.3233/TAD-2006-18105.

[10] Samsung. "What are different Sound Modes in Samsung F Series TV?" https://www.samsung.com/in/support/tv-audio-video/what-are-different-sound-modes-in-samsung-f-series-tv/ (accessed: 14/2/2022).

[11] ZVox. "What is AccuVoice?" https://zvox.com /collections/accuvoice (accessed: 14/2/2022).

[12] Which. "Sound bars that promise crystal clear speech but leave you straining your ears" https://www.which.co.uk/news/2019/02/sound-bars-that-promise-crystal-clear-speech-but-leave-you-straining-your-ears/ (accessed: 14/2/2022).

[13] J. Riedmiller et al., "Delivering Scalable Audio Experiences using AC-4," in IEEE Transactions on Broadcasting, vol. 63, no. 1, pp. 179-201, March 2017, doi: 10.1109/TBC.2017.2659623.

[14] Torcoli M., et al. "Dialog+ in Broadcasting: First Field Tests Using Deep-Learning-Based Dialogue Enhancement" *in Proceedings of IBC2021*

[15] TVC. D4.4 - Pilot-B Evaluations and recommendations. Technical report, 2016.

[16] Ward, L "Improving broadcast accessibility for hard of hearing individuals". PhD thesis. University of Salford, 2020.

[17] Ward L., et. Al. "Narrative Importance: An ecological approach to accessible audio", *New Media and Society*. 2022 [in prep.]

[18] Chourdakis, E., Ward, L., Paradis, M., and Reiss, J. D. (2019). Modelling experts' decisions on assigning narrative importances of objects in a radio drama mix., *Digital Audio Effects Conference*

[19] Riedmiller J. C. "Intelligent Program Loudness Measurement and Control: What Satisfies Listeners?" *presented at the 115th Convention of the Audio Engineering Society* (2003 Oct.), paper 5900.

[20] Straninger D. S. "Dialogue Enhancement in Object-based Audio". Bachelor's thesis. Ansbach University of Applied Sciences, 2020.

[21] Komori T. "An Investigation of Audio Balance for Elderly Listeners Using Loudness as the Main Parameter". *presented at the 125th Convention of the Audio Engineering Society* (2008 Oct.), paper 7629.

[22] Sturrock, A., et al. "Chasing the conversation: Autistic experiences of speech perception." *Autism and Developmental Language Impairments*, 2022

[23] Paradis M., Ward L., and Robinson C. "R2SPIN: Re-recording the Revised Speech Perception in Noise Test". *20th Annual Conference of the International Speech Communication Association*, 2019.

[24] BBC. BBC Sound Effects Archive https://sound-effects.bbcrewind.co.uk/ (accessed: 14/2/2022).

[25] Telegraph "BBC turns down the volume on Professor Brian Cox" https://www.telegraph.co.uk /culture/tvandradio/8379072/BBC-turns-down-the-volume-on-Professor-Brian-Cox-programme-after-viewer-complaints.html (accessed: 14/2/2022).