RNIB

See differently

RNIB Study Report: Synthetic Speech for Audio Description

Project Team

RNIB
John Paton, Media Culture and Immersive Technologies Manager
Sonali Rai Media Culture and Immersive Technologies Senior Manager
Jonathan Marshall, Research Officer, Media Culture and Immersive Technologies

Expert advisor
Dr Ben Shirley, Associate Professor, Acoustics Research Centre, University of Salford

Many thanks to BBC, ITV and Warner Bros Discovery for their content and technical support.

Audio description produced by Roz Chalmers.

Software used: Video to Voice

Table of contents

User study: Using synthetic speech for audio description

Audio description (AD) is an essential service that makes visual content accessible to blind and partially sighted people by providing spoken commentary for visual details not conveyed through dialogue. Commonly delivered as an alternate track on TV, streaming platforms, and in cinemas, AD enables those with sight loss to follow the plot and enjoy visual media. However, the quality of AD significantly impacts the overall experience, with clarity, naturalness, and emotional engagement being crucial. Research conducted by the Research Institute for Disabled Consumers (RIDC) for Ofcom underscores the importance of clear diction and emphasises that the accents and tones of audio describers should align with the actors, genre, and context of a programme to enhance the experience [1].

In recent years, synthetic voices have emerged as a practical solution for AD, driven by their affordability and scalability. AI-generated voices provide cost-effective alternatives to human narrators and offer the flexibility of personalisation. Producers can tailor synthetic voices to match specific accents, or styles, potentially enhancing the overall viewing experience by creating a more cohesive alignment between the AD and the programme’s content.

The exploration of synthetic speech in media is not new. In 2017, the European Commission funded SCRIPT, a research project aimed at developing synthetic voices for low-resource languages such as Hausa, Swahili, and Bengali [2]. The BBC, in collaboration with the University of Edinburgh's Centre for Speech Technology Research (CSTR), contributed to this project, facilitating the creation of voice-overs for videos produced in other languages [3]. In 2020, the BBC developed a prototype of the ALTO tool [4], a virtual voice-over solution that uses text-to-speech synthesis, enabling efficient reversioning of video content with multiple voices through a single editor. More recently, Project Gutenberg, in collaboration with MIT and Microsoft, developed an advanced text-to-speech system that narrates over 5,000 audiobooks, showcasing the feasibility of large-scale automated audiobook production [5].

Blind and partially sighted people have long been early adopters of synthetic speech, primarily through screen readers like JAWS, VoiceOver, and TalkBack. However, its use has largely been limited to work or study related tasks and also, uncurated tasks such as reading digital content or browsing the internet, where human alternatives are not available.

One significant challenge for synthetic voices is replicating the emotional nuances that human narrators bring. Subtle cues such as irony, enthusiasm, or emphasis are difficult for synthetic voices to convey, yet these elements are crucial to the quality of AD. According to RIDC, while many participants opposed synthetic voices for AD, some acknowledged their potential for specific programmes, provided they maintain high-quality standards and expand the availability of AD content[6].

In response to growing inquiries about the suitability of synthetic voices for AD, RNIB launched this project to evaluate the quality of synthetic voices, their suitability for AD and their impact on viewer experiences. The study used the MOS-X2 [7] scale, a widely recognised metric for assessing the quality of multimedia services, including synthetic speech. The analysis focused on categories such as intelligibility, naturalness, prosody, and social acceptance, offering insights into synthetic speech's current capabilities and its potential future role in AD.

This section of the report summarises the perspectives of focus group participants on the naturalness, clarity, and suitability of synthetic voices, with particular emphasis on factors such as emotional engagement, accents, and continuity in delivery.

1 Results from the quantitative study

1.1 Introduction

This section presents quantitative findings from a survey investigating the effectiveness of synthetic voices for audio description (AD).

A 31-question survey was made available for six weeks, featuring six sample clips with synthetic AD across various genres. The genres represented included entertainment, drama, sports, documentary, and factual content to provide a comprehensive basis for examining the performance and suitability of synthetic

voices in different contexts. Due to content licensing agreements, clips may be blocked in certain regions.

The clips used were:
- Strictly Come Dancing (entertainment)
- I'm a Celebrity (gameshow)
- Rugby match (sports)
- Riches (drama)
- Living Next Door to Putin (documentary)
- DIY SOS (factual reality)

Each clip served as a basis for exploring participant reactions to synthetic voices in AD. Participants were required to watch each clip in full before proceeding to questions, ensuring that their feedback reflected a complete viewing experience.

The modified MOS-X2 scale was adapted to analyse the data gathered in this study.

1.2 MOS-X2 Scale

MOS, or Mean Opinion Score, is a widely used measurement tool to assess perceived quality across various factors, particularly in synthetic voice applications. By obtaining user ratings across a series of categories, then averaging these scores, a single metric is produced to represent the perceived quality of the synthetic voice.

MOS-X2 builds on earlier iterations by simplifying user input to four key categories:
- Intelligibility
- Naturalness
- Prosody
- Social impression

Each is scored from 0 to 10, with 0 representing the lowest level of perceived quality and 10 the highest. These scores are averaged, then multiplied by 10 to provide a single metric out of 100, which is mapped against a curved scoring gradient for a final grade.

The 'Modified MOS-X2' used in this study adapted the scale to emphasise program-specific factors. Each category was framed in terms of relevance to the effectiveness of the AD within the

specific program genre (e.g., the appropriateness of timing and pitch for a given program). The criteria could be summarised as follows:

- Intelligibility: (0 = unintelligible, 10 = fully intelligible)
- Naturalness: (0 = extremely unnatural, 10 = completely natural)
- Prosody: (0 = completely inappropriate timing and emphasis, 10 = always appropriate)
- Social Impression: (0 = emotionally inappropriate, 10 = always appropriate)

1.2.1 Scoring Guide

The scoring translates into the following grades:

- 0 - 53.9: F
- 54 - 61.2: D
- 61.3 - 63.4: C-
- 63.5 - 66.2: C
- 66.3 - 67.6: C+
- 67.7 - 69.6: B-
- 69.7 - 70.8: B
- 70.9 - 71.4: B+
- 71.5 - 73.4: A-
- 73.3 - 79.7: A
- 79.8 - 85.2: A+
- 85.3+: A+, Human-like

The modified MOS-X2 scale used in this study followed the MOS-X2 structure but included slight wording changes to align with specifics of AD (e.g., emotional tone appropriateness and timing suited to the programme type).

1.3 Modified MOS-X2

This study aimed to objectively measure the perceived quality and appropriateness of the voices used for audio description. To achieve this, the original wording of the MOS-X2 questions was modified slightly to better align with the context of audio description. Specifically, prosody and social impression were evaluated in terms of their suitability for each programme. The modified questions were:

Intelligibility: Please rate from 0 to 10 the extent to which it was easy or difficult to understand what the voice was saying.

- 0 – completely unintelligible (hard to understand)
- 10 – completely intelligible (easy to understand)

Naturalness: Please rate from 0 to 10 how natural (pleasantly human-like) the sound of the voice was.
- 0 – extremely unnatural
- 10 – completely natural

Prosody: Please rate from 0 to 10 to what extent were elements of timing, pitch, and emphasis appropriate for the programmes.
- 0 – completely inappropriate
- 10 – always appropriate

Social impression: Please rate from 0 to 10 to what extent was the tone of the voice emotionally appropriate for the programmes.
- 0 – never appropriate
- 10 – always appropriate

All other aspects of the MOS-X2 scale were maintained. Where this report references MOS-X2 (except in the introduction), it refers to this modified version.

1.4 Demographic information

A total of 37 people responded to the survey.

1.4.1 Age

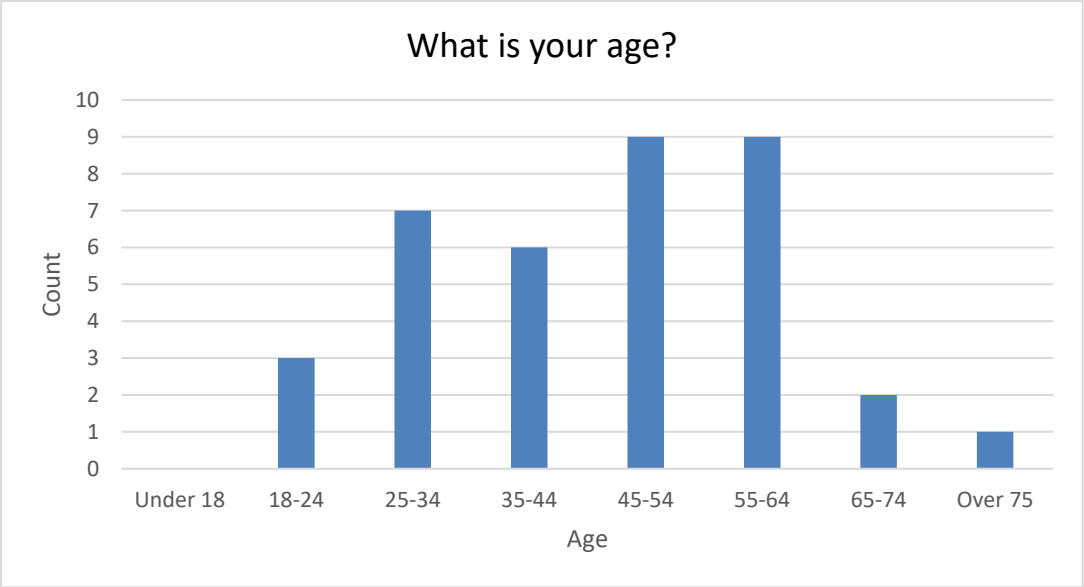

Fig 1. Ages of participants

The largest age groups in the study were those aged 45-54 and 55-64, each with nine participants. This was followed by seven participants aged 25-34, while other groups included six participants aged 35-44, three aged 18-24, two aged 65-74, and one respondent over 75.

1.4.2 **Level of sight loss**

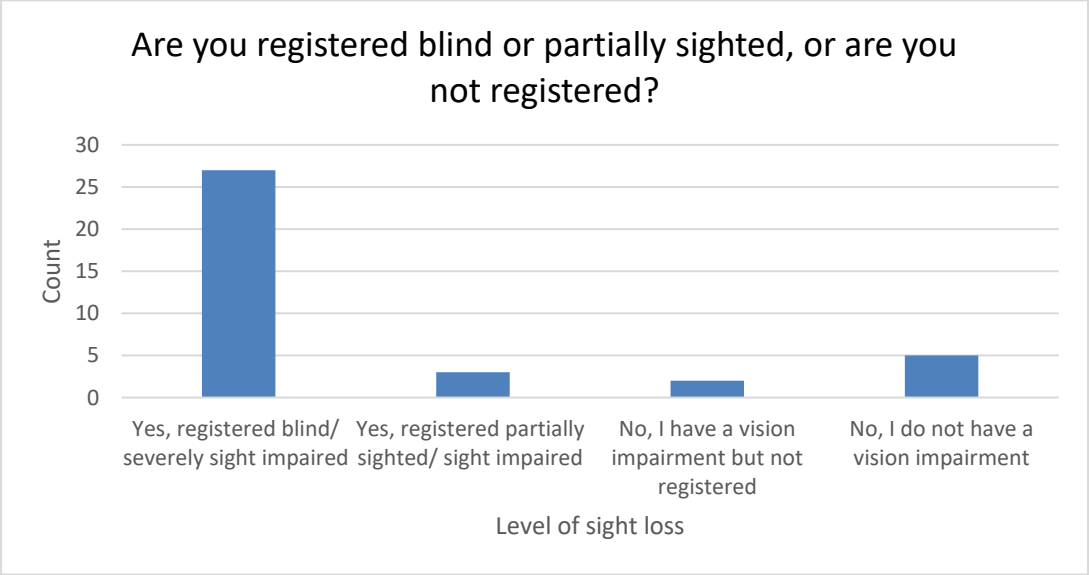

Fig 2. Level of sight loss

A majority of participants (27) were registered as blind, while three were registered as partially sighted. Two participants had sight loss but were not registered, and four participants reported no sight loss.

1.5 Use of AD

1.5.1 How often do you use AD?

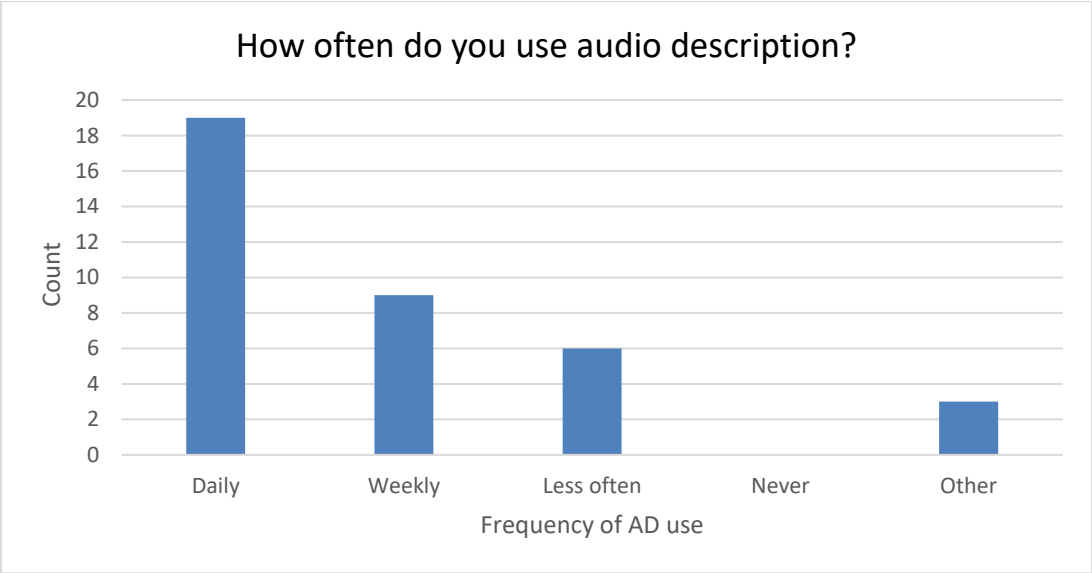

Fig 3. Frequency of AD use.

Most participants used AD daily (19), while others used it weekly (nine) or less often (six). No one indicated that they never use AD, though three selected the 'other' option. Of these, two were audio describers, and one respondent described an academic interest in AD, using it occasionally to "watch" TV or listen to audio books, podcasts, and live events.

1.5.2 Do you always switch AD on when it is available while watching TV?

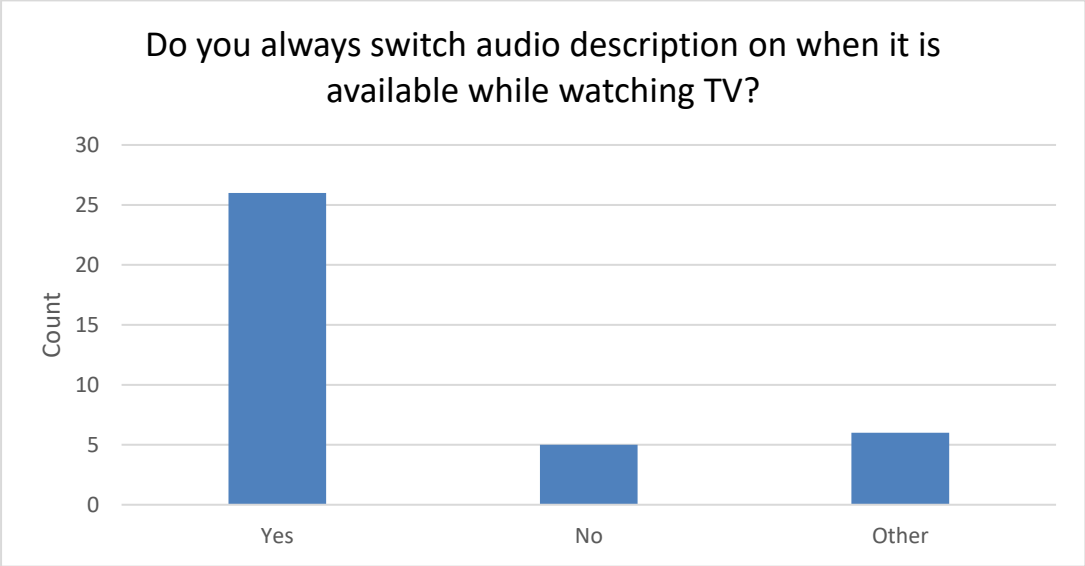

Fig 4. Do you always switch audio description on when it is available while watching TV?

Twenty-six participants reported they always switch on AD when available, while five indicated they do not always use it, and six selected the ‘other’ option. Comments from the ‘other’ responses included:

- Occasional use (“sometimes,” “only when I’m at Gran’s”)
- Context-based use, such as while cooking or when viewing alone due to family preferences or screen accessibility issues.

One respondent, although they always used AD when watching TV, noted that they rarely watched TV.

1.5.3 How often would you say you turn AD off when it is available?

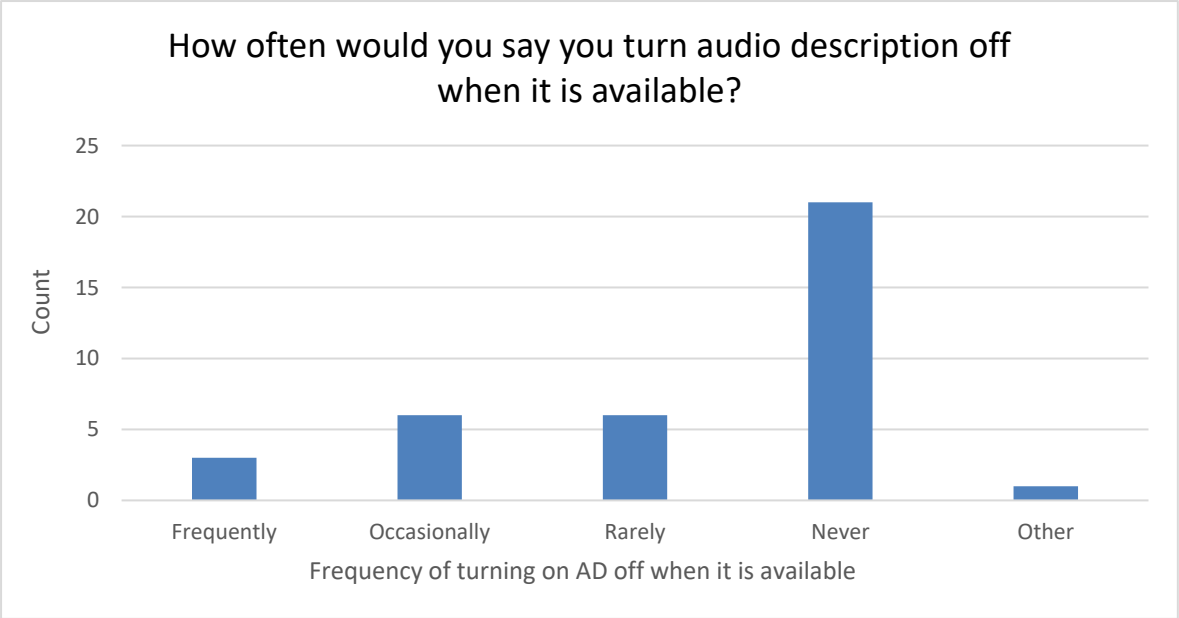

Fig 5. How often would you say you turn audio description off when it is available?

Twenty-one participants indicated they never turn off AD once it is available. Six participants said they occasionally or rarely switch it off, three said they frequently switch it off, and one person selected the ‘other’ option, noting that they provide AD services.

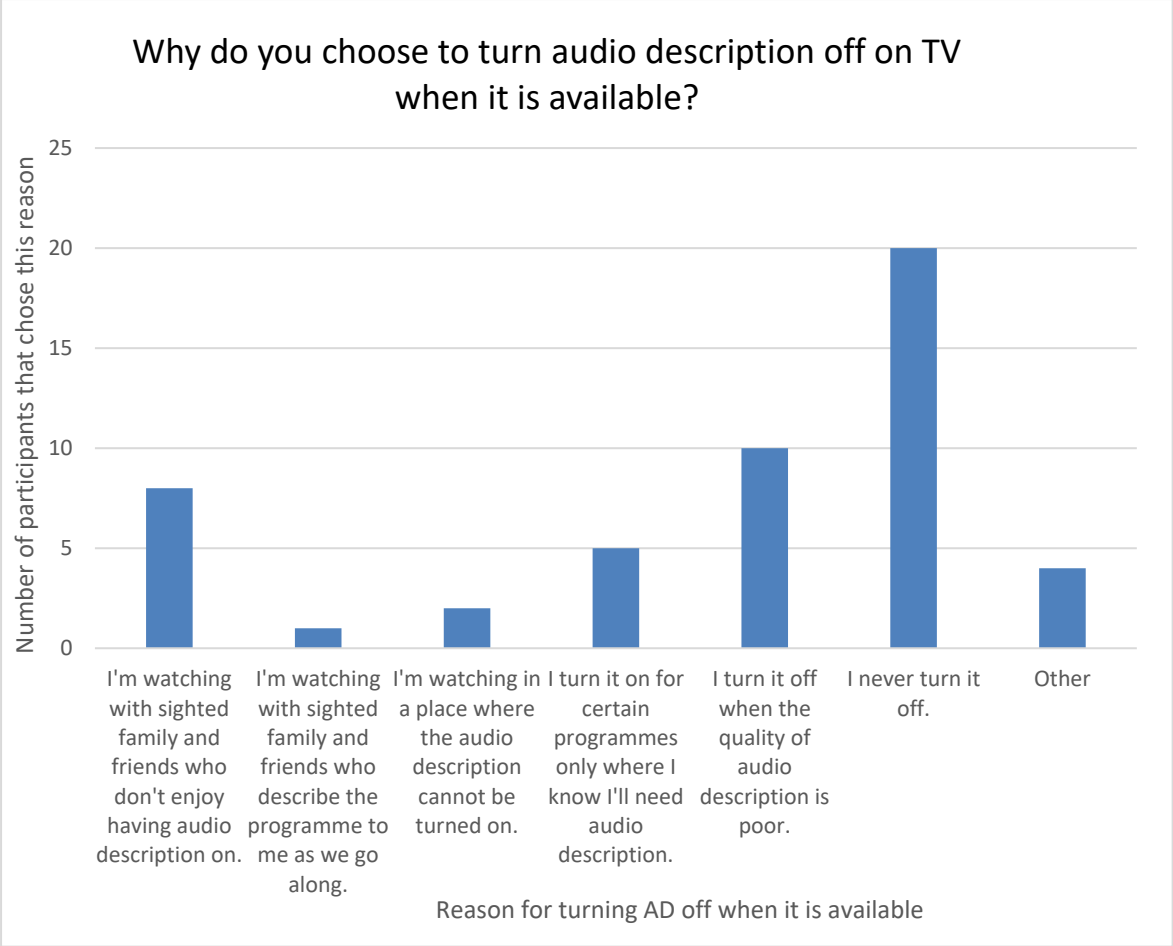

Fig 6. Why do you choose to turn audio description off on TV when it is available?

When asked why they turn off AD, responses varied:

- 20 participants said they never turn off AD if it's available.
- 10 noted they turn it off when the quality of the AD is poor.
- Eight turn it off when watching with sighted family or friends who find it distracting.
- Five use it selectively for specific programs, while two noted accessibility issues in certain locations.

Of the people who selected 'other' two said they were audio describers

A third who identified themselves elsewhere as an audio describer said:
"For work purposes"

And one said:
"Sometimes audio description is way too chatty and/or uses unnecessarily flowery language when it should do the

minimum to convey what's going on without getting in the way of the programme."

This feedback illustrates diverse habits and preferences regarding audio description usage, with accessibility, quality, and social context influencing participants' decisions to engage with AD while watching TV.

1.6 Grading of AD voices

1.7 Clip 1 - Sports

Rugby Union match between Saracens vs Leicester (1 min 42 secs)

This clip opened with commentator discussion, transitioning into play after 20 seconds. The AD on this clip received an average score of 68.6 (B- grade).

Score distribution by factor:

- Intelligibility 30%
- Naturalness 24%
- Prosody 24%
- Social Impression 22%

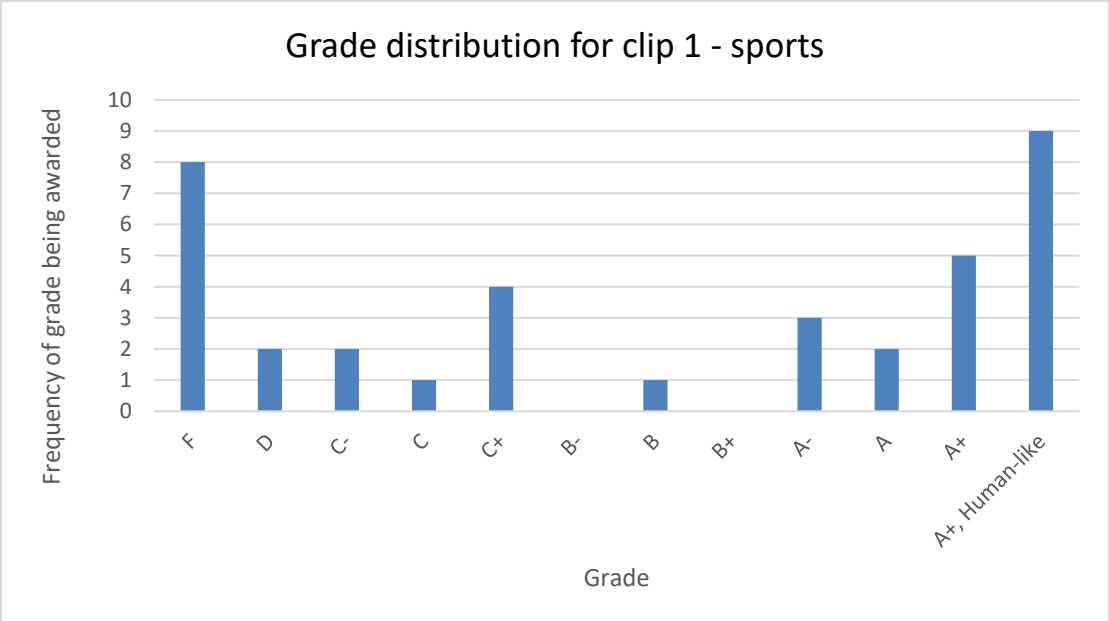

Fig 7. Grade distribution for clip 1 - sports

The response distribution shows mixed opinions, with nine participants giving an A+ (human-like), while eight rated it F. Peaks

were seen in the C+ and A- grades, with fewer participants awarding A, C-, and D grades.

Graph description: This clip received nine A+, human-like grades and eight F grades. Five people graded it A+ and there were two peaks at C+ (four) and A- (three). A, C- and D grades were all given twice, and C and B were given once.

1.8 Clip 2 – Drama

Riches 1 mins 33 secs

This clip features a black CEO of a beauty empire discussing his business challenges with a reporter. The AD on this clip received an average score of 77.5 (A grade).

Score distribution by factor:

- Intelligibility 23%
- Naturalness 25%
- Prosody 25%
- Social Impression 26%

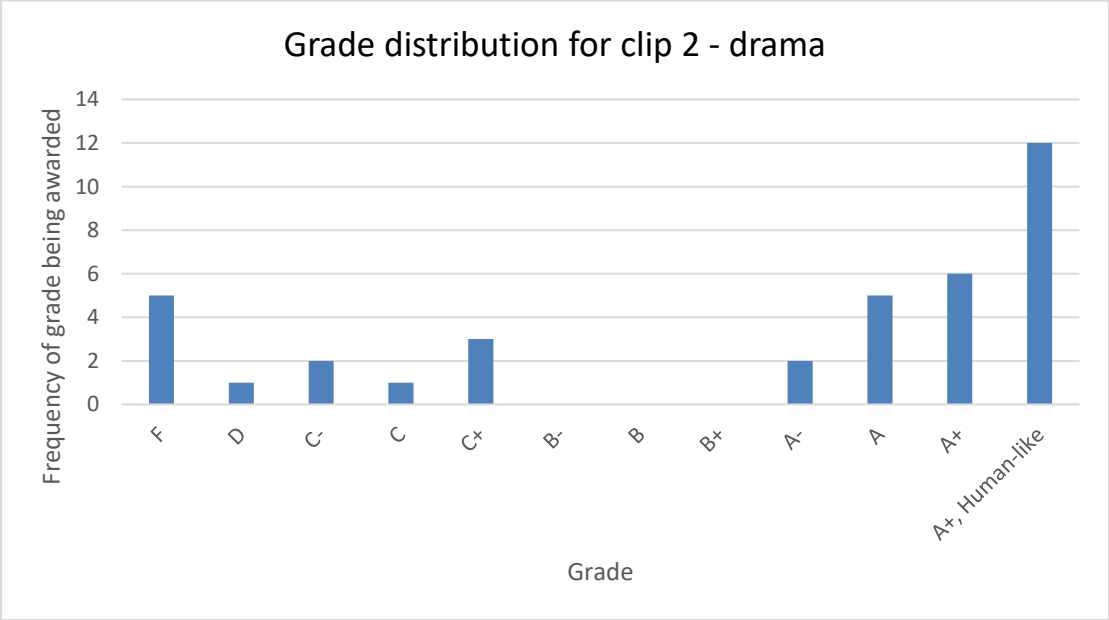

Fig 8. Grade distribution for clip 2 - drama

Participants awarded 12 A+ (human-like) grades, with some variance in other responses. Scores ranged across F, C+, A-, and C- grades, showing a strong performance with some detractors.

Graph description: 12 participants gave this clip a grade of A+, Human-like. Six gave an A+ grade and five gave grades of A and F. Three people gave a grade of C+, two people each gave A- and C- grades and D and C grades were given once each.

1.9 Clip 3 – Documentary

Living next door to Putin 1 min 48 seconds

This documentary with Katya Adler is about the response in Europe to Russia's invasion of Ukraine. The AD on this clip received an average score of 82.8 (A+ grade).

Score distribution by factor:

- Intelligibility 27%
- Naturalness 24%
- Prosody 25%
- Social Impression 25%

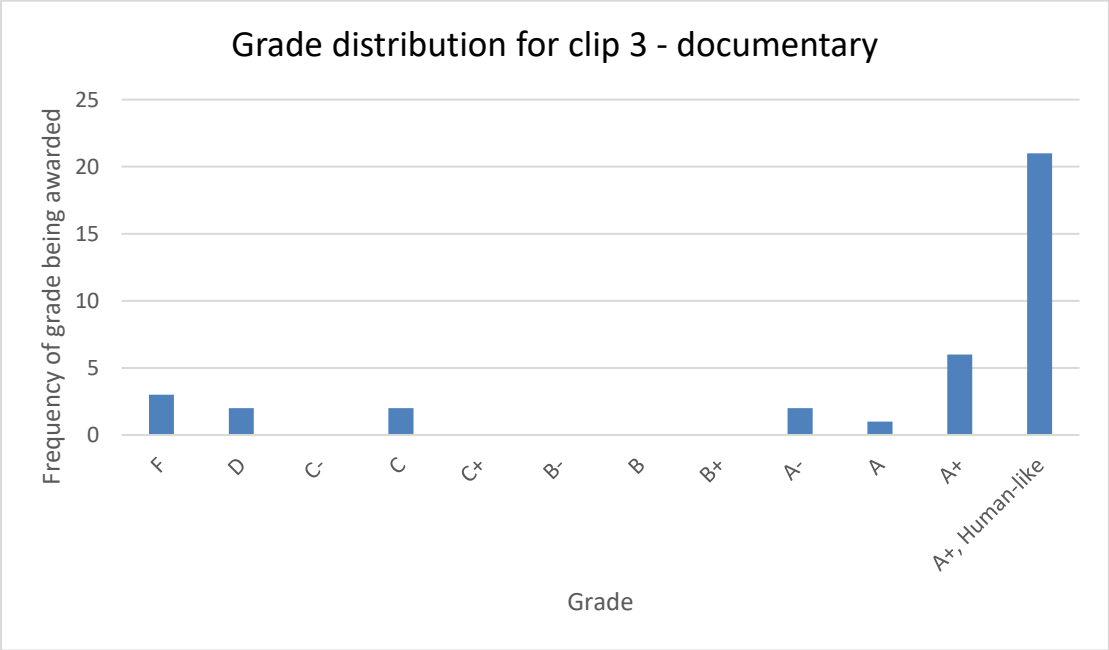

Fig 9. Grade distribution for clip 3 - documentary

A large majority (21 participants) rated this clip as A+ (human-like), with the rest spread across A+, F, A-, and C. This high rating indicates effective alignment with the documentary genre's tone and content.

Graph description: 21 participants gave this clip a grade of A+, human-like. Six gave it a grade of A+ and three gave it an F grade.

Two people each gave grades of A-, C and F and one person gave an A grade. B+, B, B-, C+ and C- grades were not given.

1.10 Clip 4 – Entertainment

Strictly come dancing (1 min 9 secs)

The clip includes the show's opening credits and descriptions of presenters Claudia Winkleman and Tess Daly. The AD on this clip received an average score of 76.6 (A grade).

Score distribution by factor:

- Intelligibility 30%
- Naturalness 22%
- Prosody 25%
- Social Impression 23%

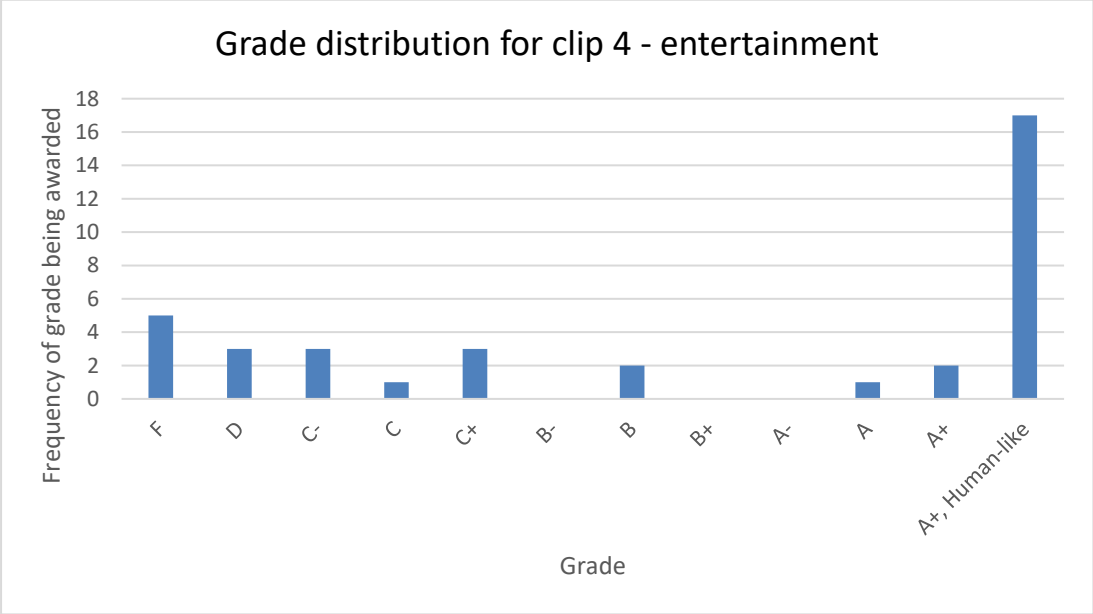

Fig 10. Grade distribution for clip 4 - entertainment

This clip was rated A+ (human-like) by 17 participants, though a few graded it F, suggesting the entertainment genre might benefit from improved naturalness and emotional alignment.

Graph description: 17 participants gave this clip a grade of A+, Human-like. Five gave an F grade. C+, C- and D grades were given thrice each and A+ and B grades were given twice each. A and C grades were given once each.

1.11 Clip 5 – Game show

I'm a celebrity… South Africa (1 min 37 secs)

The clip had the series recap and the welcome from Ant and Dec cutting off just before the opening credits. The AD on this clip received an average score of 69.5 (B- grade).

Score distribution by factor:

- Intelligibility 29%
- Naturalness 25%
- Prosody 24%
- Social Impression 23%

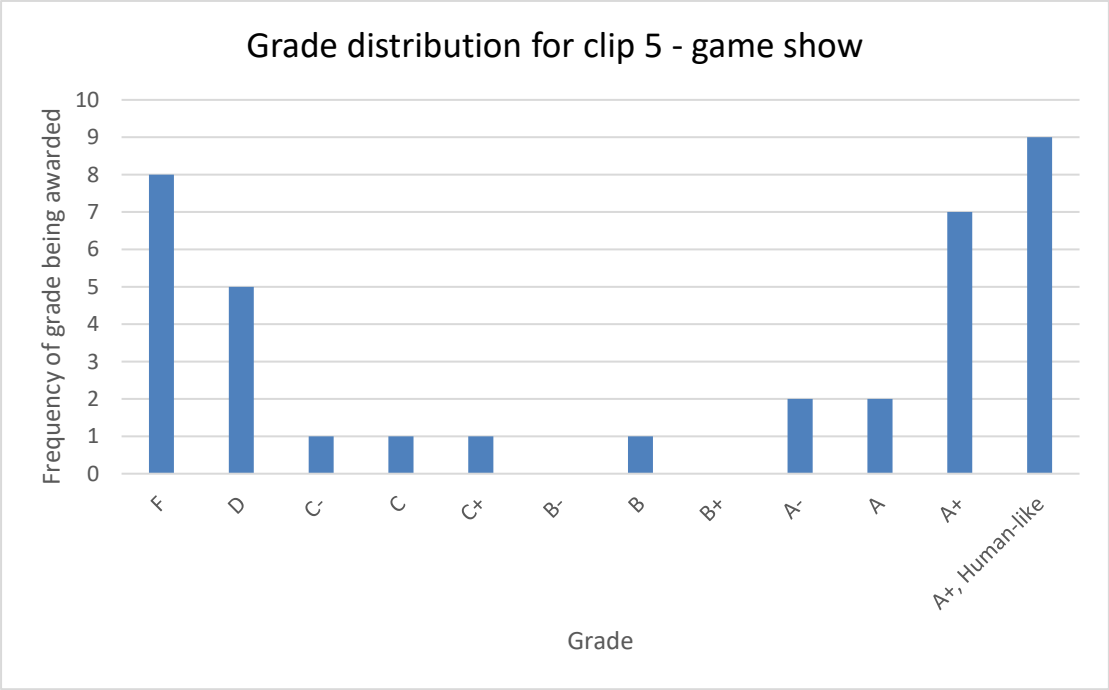

Fig 11. Grade distribution for clip 5 - game show

The distribution shows polarised responses, with nine A+ (human-like) grades but also a noticeable number of F and D grades, indicating varied perceptions of its suitability for the genre.

Graph description: Nine participants gave this clip a grade of A+, human-like and eight people gave an F grade. Seven A+ grades were awarded and six D grades. A and A- grades were awarded twice each and B, C+, C and C- grades were awarded once each.

1.12 Clip 6 - Factual reality show

DIY SOS special for Children in Need (1 min 33 seconds)

Nick Knowles and Pudsey Bear attempt to access the backstage area at Radio 2 in the Park to ask for help from Scott Mills and the Radio 2 DJs for a build. The AD on this clip received an average score of 78 (A grade).

Score distribution by factor:

- Intelligibility 28%
- Naturalness 25%
- Prosody 24%
- Social Impression 23%

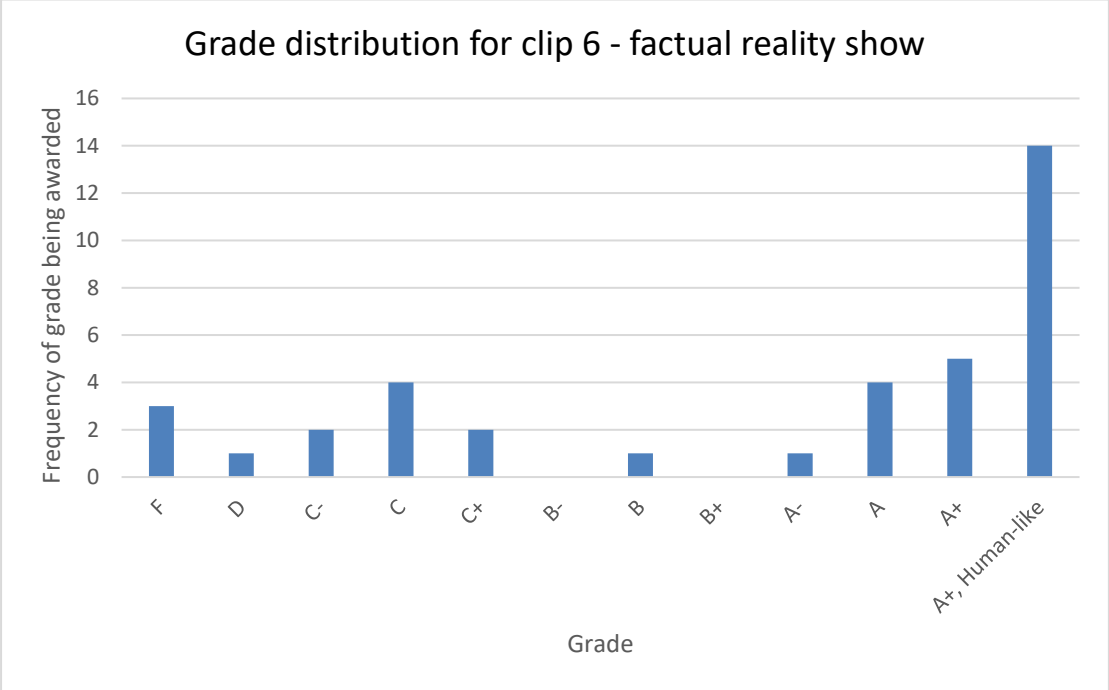

Fig 12. Grade distribution for clip 6 - factual reality show

Fourteen participants rated this A+ (human-like), with some scoring at A+, A, and C. Fewer scores were seen in the F, C+, and D categories, indicating strong performance in factual reality.

Graph description: A+ human-like was awarded 14 times. A+ was awarded 5 times an A and C were awarded 4 times each. Three F grades were awarded and C+ and C- were awarded twice each. A. B and D grades were awarded once each.

1.13 Scores across all the clips

The average across all the clips was 75.5 which corresponds to an A grade on the modified MOS-X2.

The scoring across the four factors (Intelligibility, Naturalness, Prosody and Social Impression) was fairly even with each factor making up between 22% and 30% of the average score for the clip. The largest difference for one clip was the sports clip where Intelligibility made up 30% of the score and Social Impression made up 22%.

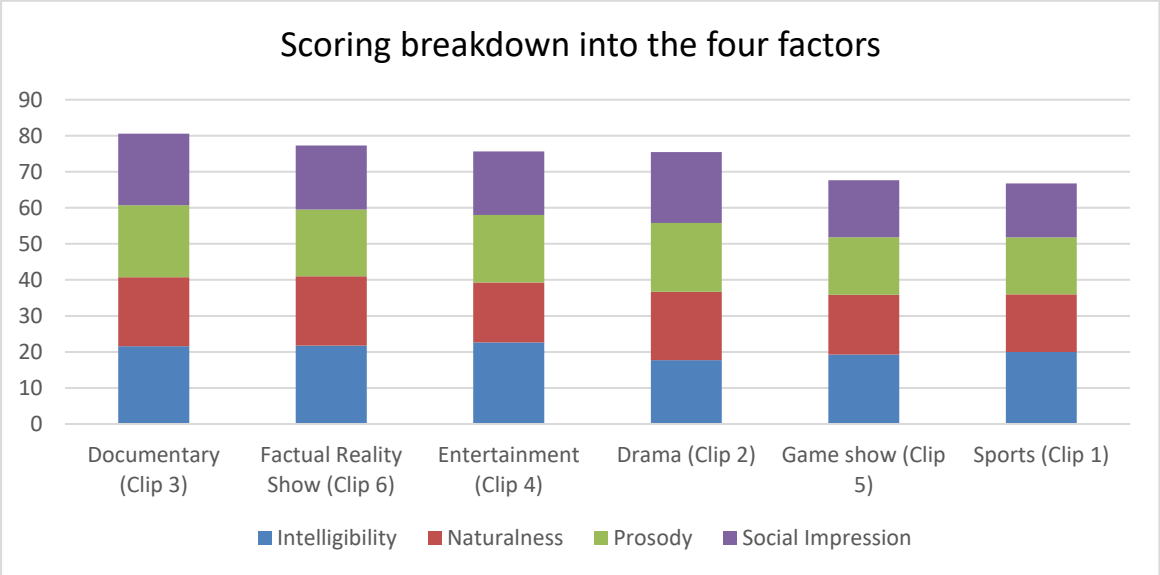

Fig 13. Scoring breakdown into the four factors

Across all the scores given the proportion attributed to each factor is:

- Intelligibility 27.7%
- Naturalness 24.0%
- Prosody 24.4%
- Social Impression 23.8%

This is fairly even and mostly consistent across the clips. Intelligibility is generally the highest segment of the score with the drama program “Riches” bucking this trend. Social Impression made up the smallest proportion of the scores being the worst scored factor in half of the clips.

1.14 The views of audio describers

Three participants identified themselves as audio describers. The average score from this group was 77.8. This is very slightly higher than participants who did not identify themselves as AD providers who gave an average score of 75.3. The scores of both groups correspond to an A grade.

The proportion attributed to each factor for the audio describers is:

- Intelligibility 27.9%
- Naturalness 23.3%
- Prosody 25.8%
- Social Impression 23.1%

This is broadly in line with the average from all participants.

1.15 Would you use synthetic speech to listen to any of the following?

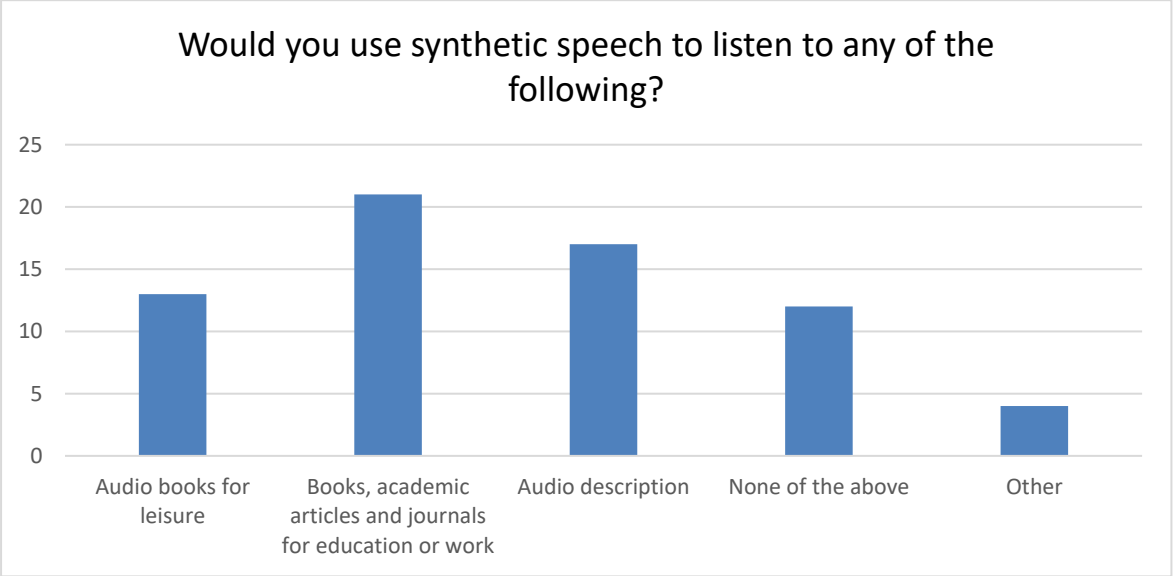

Fig 14. How do participants use synthetic speech?

Participants generally showed an openness to using synthetic speech, especially for informational content, such as educational materials and work-related texts. A majority (21) indicated they would use synthetic voices for academic reading, and a smaller number (17) for AD. However, 12 participants were reluctant to use synthetic voices at all, and four chose the ‘other’ option, with mixed opinions on its suitability.

Among the ‘other’ responses, two were firmly against synthetic speech for AD, expressing concerns that it lacks naturalness and may affect employment for narrators. Some participants showed conditional support, with one respondent suggesting they would opt for synthetic speech if no alternative existed, while another felt that improving synthetic quality could make it more acceptable.

1.16 Attitudes towards synthetic speech

In this study, the use of synthetic speech was not revealed to participants initially, allowing participants to form objective assessments of voice quality.

1.16.1 Participants who said they would use synthetic speech for AD

This group gave an average score of 77.5, corresponding to an A grade, which is slightly above the overall average for all responses.

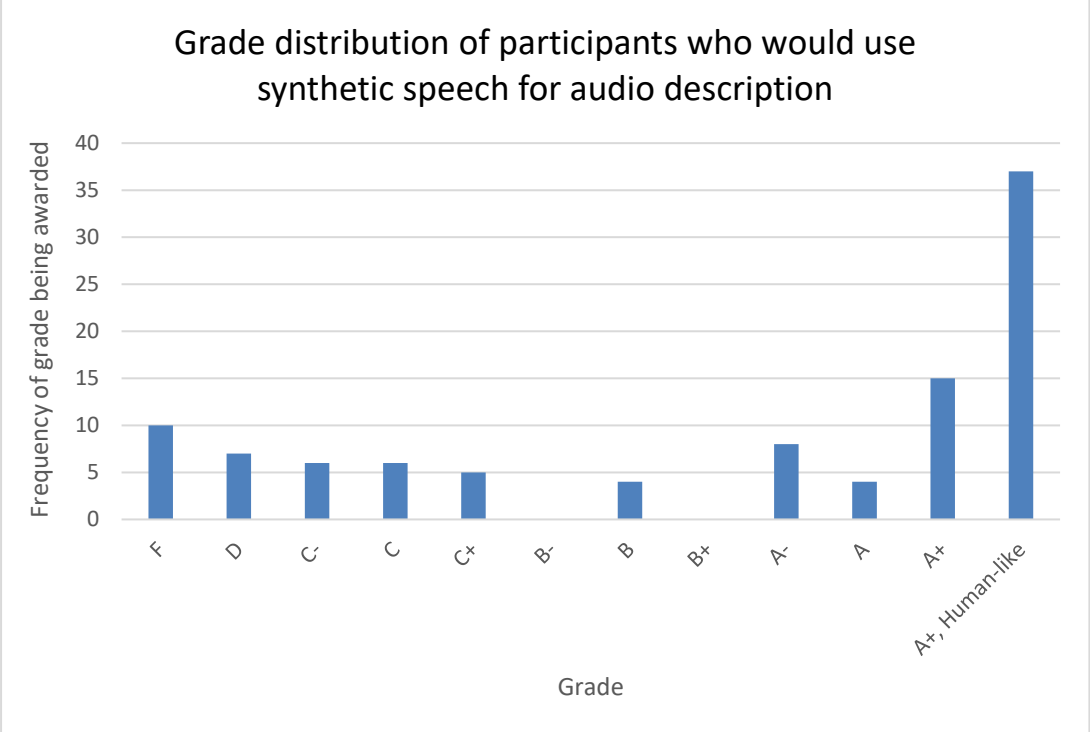

Fig 15. Grade distribution of participants who would use synthetic speech for audio description

Grade distribution:

- A+ Human-like grades were awarded 37 times, with A+ as the next most frequent grade at 15.
- Lower grades, such as F, appeared 10 times; A- eight times; and D seven times.
- C-range grades included C and C- six times each, C+ five times, and A and B four times each.
- No B- or B+ grades were awarded.

1.16.2 Participants who said they would not use synthetic speech for AD

This group awarded an average score of 73.8, which also corresponds to an A grade, though slightly below the overall average.

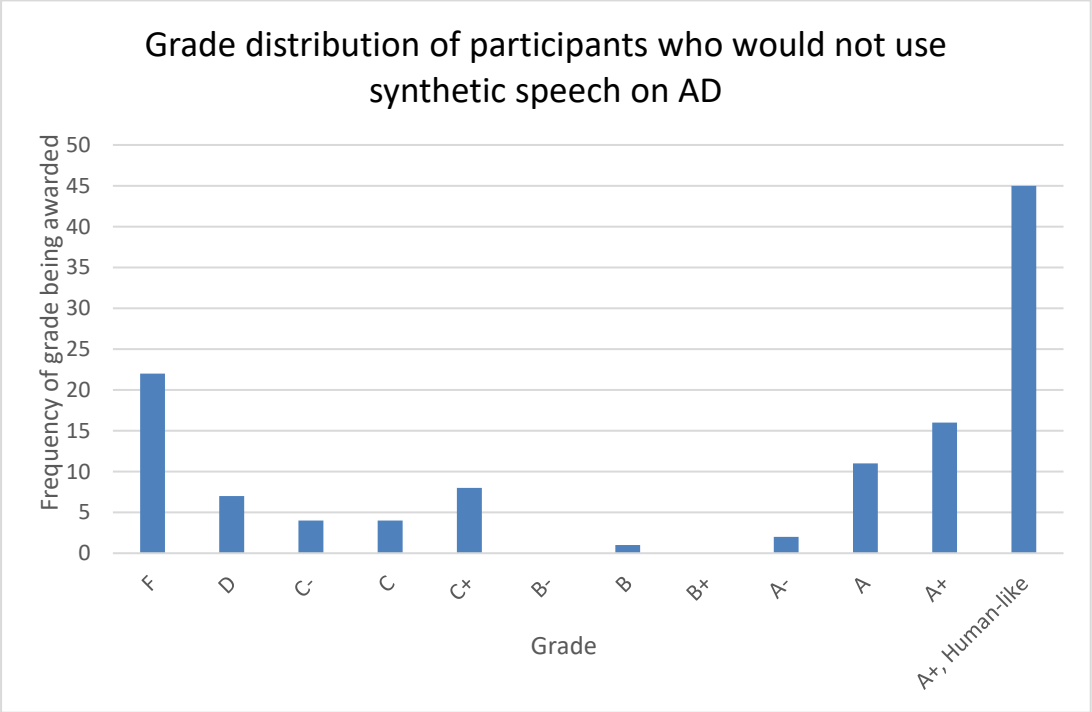

Fig 16. Grade distribution of participants who would not use synthetic speech on AD

Grade distribution:

- A+ Human-like was the most awarded grade with 45 instances, followed by F at 22.
- Additionally, A+ grades were awarded 16 times, A 11 times, and C+ 8 times.
- Lower grades included D seven times and C and C- four times each.
- Only two A- grades and one B grade were awarded; no B+ or B- grades were given.

1.16.3 Grade distribution of participants who would not use synthetic speech for any options

This subset yielded an average score of 75, equating to an A grade, which is slightly below the overall average.

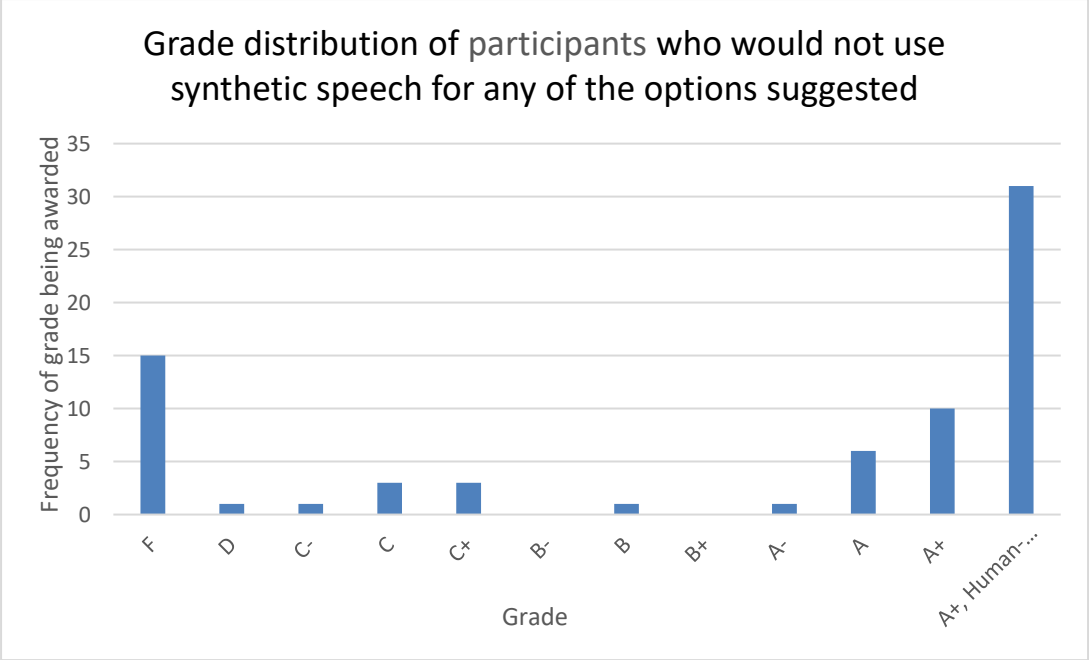

Fig 17. Grade distribution of participants who would not use synthetic speech for any of the options suggested

Grade distribution:

- A+ Human-like was given 31 times, with F as the next most frequent at 15.
- A+ appeared 10 times, A six times, C+ and C three times each.
- Lower grades A-, C- and D each had one instance.

One participant who indicated a preference for human voices rated each clip with F grades, commenting, “I can always tell when a human is reading…” This participant’s average score deviated substantially from the mean by 2.9 standard deviations from other participants and 2.3 deviations from those who would never use synthetic speech, suggesting an outlier. Their ratings may reflect a higher personal standard or suspicion that the voices were artificial.

If we take this respondent out, then the average score for participants who would never use synthetic speech is 78.5 which is higher than the average overall but still an A grade on the modified MOS-X2 scale.

Did participants who disliked synthetic speech grade the voices lower?

If a preference for “human quality” voices influences participants’ assessments of synthetic voices, one might expect that

participants who prefer not to use synthetic speech for AD would consistently rate the voices lower in their evaluations.

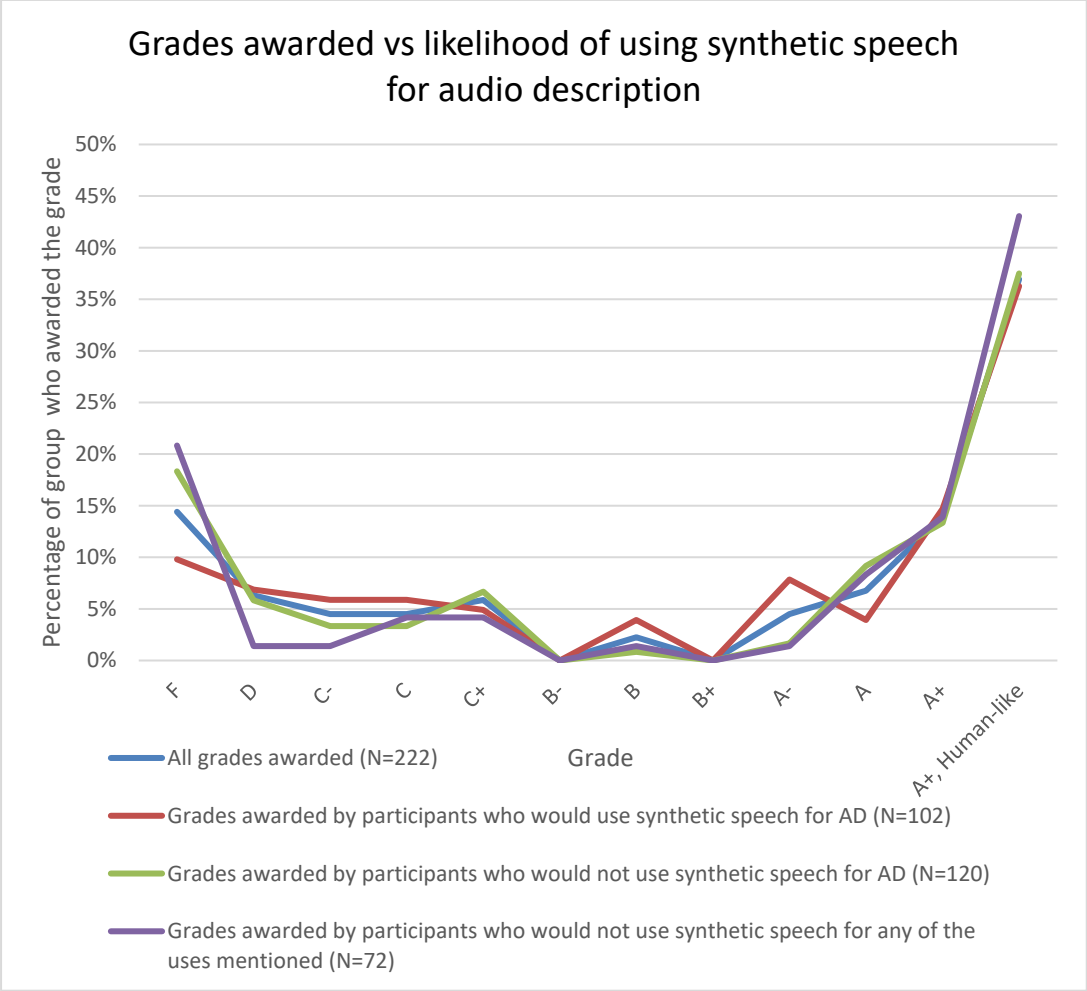

Fig 18. Grades awarded vs likelihood of using synthetic speech for audio description

The gradings do not show a significant trend of lower scores from participants who prefer human voices, suggesting a disconnect between perceptions of synthetic speech and their evaluation in the study. The synthetic voices selected for the study were high-quality samples, as evaluated by the testers, which may have influenced their evaluation. Additionally, participants' general aversion to synthetic voices could encompass a broader range of available voices, many of which are of lower quality. Another factor to consider is clip length: shorter clips reduce the risk of listener fatigue or discomfort, which may occur over prolonged exposure, as the 'uncanny valley' effect may make subtle imperfections more noticeable in longer samples.

1.16.4 Grade distribution for participants with different levels of sight loss

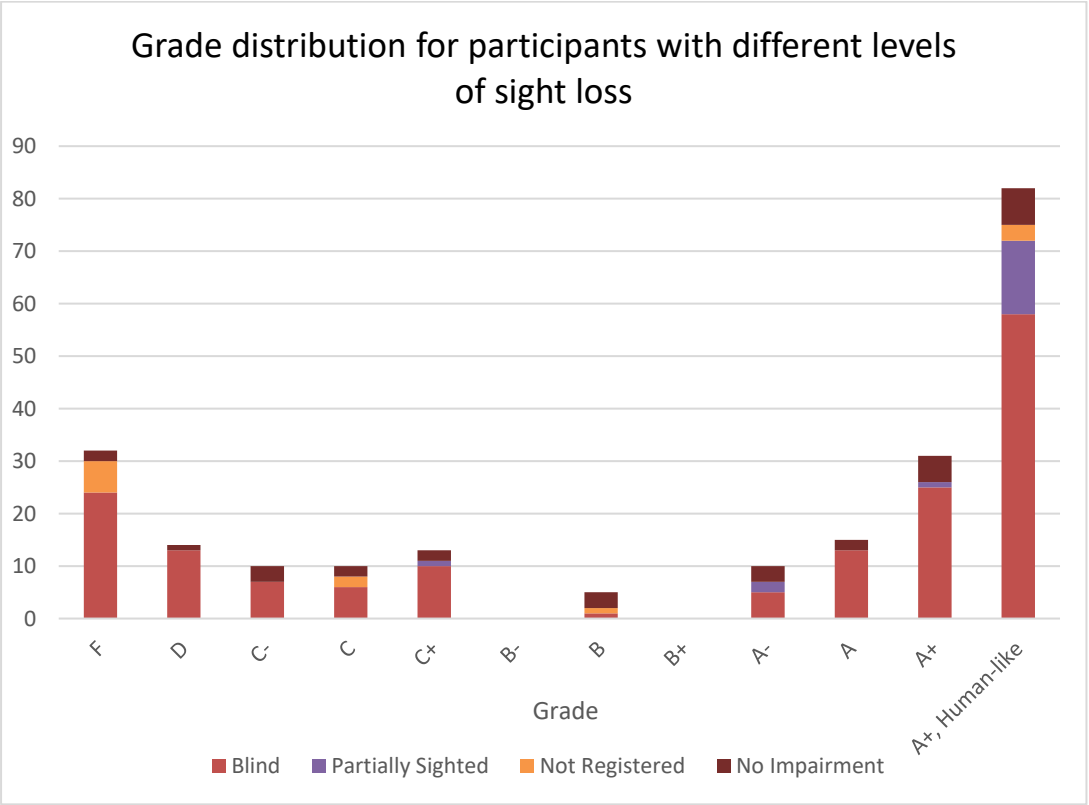

Fig 19. Grade distribution for participants with different levels of sight loss

The data across levels of sight loss was limited, with only three participants identified as partially sighted. This group scored the clips exceptionally high, averaging 93.75 (A+ human-like). By contrast, the 27 blind participants gave an average score of 74.8, slightly below the overall study average. Two participants with unregistered sight loss provided a range of scores, with grades corresponding to both A+ and F. Among the four participants with no visual impairment, the average score was 74.9, aligning with an A grade.

The difference between the scores given by blind and partially sighted participants is highly suggestive that blind AD users are more critical of the quality of the voice even though it lacks sufficient statistical significance to draw meaningful conclusions.

1.16.5 Proportion of grades given vs frequency of using AD on TV

- Daily AD Users (19 participants): This group assigned an average score of 79.6 (A). The distribution of grades showed a

lower number of grades from F to B- but a sharp increase in A+ (human-like) grades, resulting in the second-highest proportion of top grades among all groups, just slightly lower than the 'Other' category.

- Weekly AD Users (9 participants): With an average score of 66.7 (C+), this group awarded the most F grades and spread the higher grades broadly across A- to A+ (human-like) ratings.
- Occasional AD Users (6 participants): This group averaged a score of 74.5 (A) and distributed their grades relatively evenly across all levels, with an increase in A+ (human-like) scores.
- Other Category (3 participants): These participants averaged 77.9 (A). They gave a notable number of F grades but also had the highest proportion of A+ (human-like) grades, though only by a small margin.

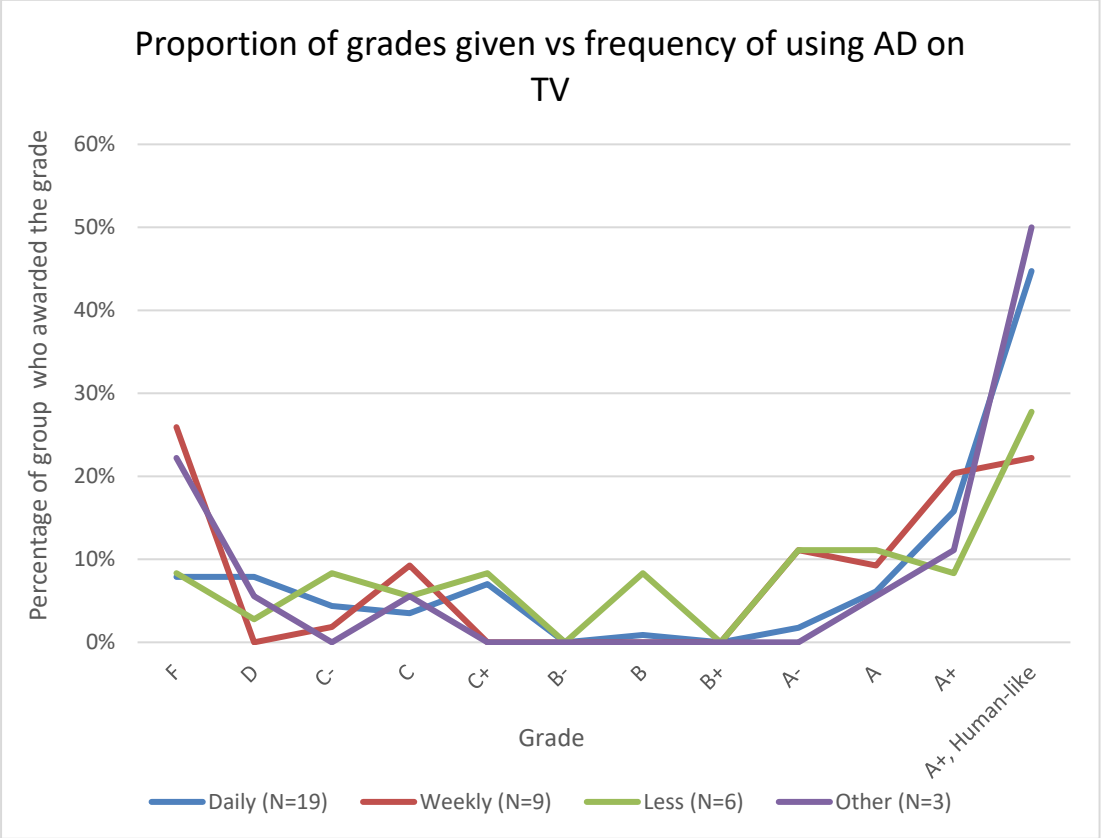

Fig 20. Proportion of grades given vs frequency of using AD on TV

The daily AD watchers spread a low number of grades across F to B- and then ramped up the grades sharply towards A+ human like to give the second highest proportion of the highest grade just short of viewers who selected 'other'.

The weekly AD watchers gave the most F grades and spread the higher grades across A- to A+ human-like.

Viewers who selected ‘less often’ spread their grades fairly evenly across all levels with an uptick at A+ human-like.

The viewers who selected ‘other’ gave a comparatively high number of F grades and then ramped up the grades sharply towards A+ human like to give the highest proportion of the highest grade by a small margin.

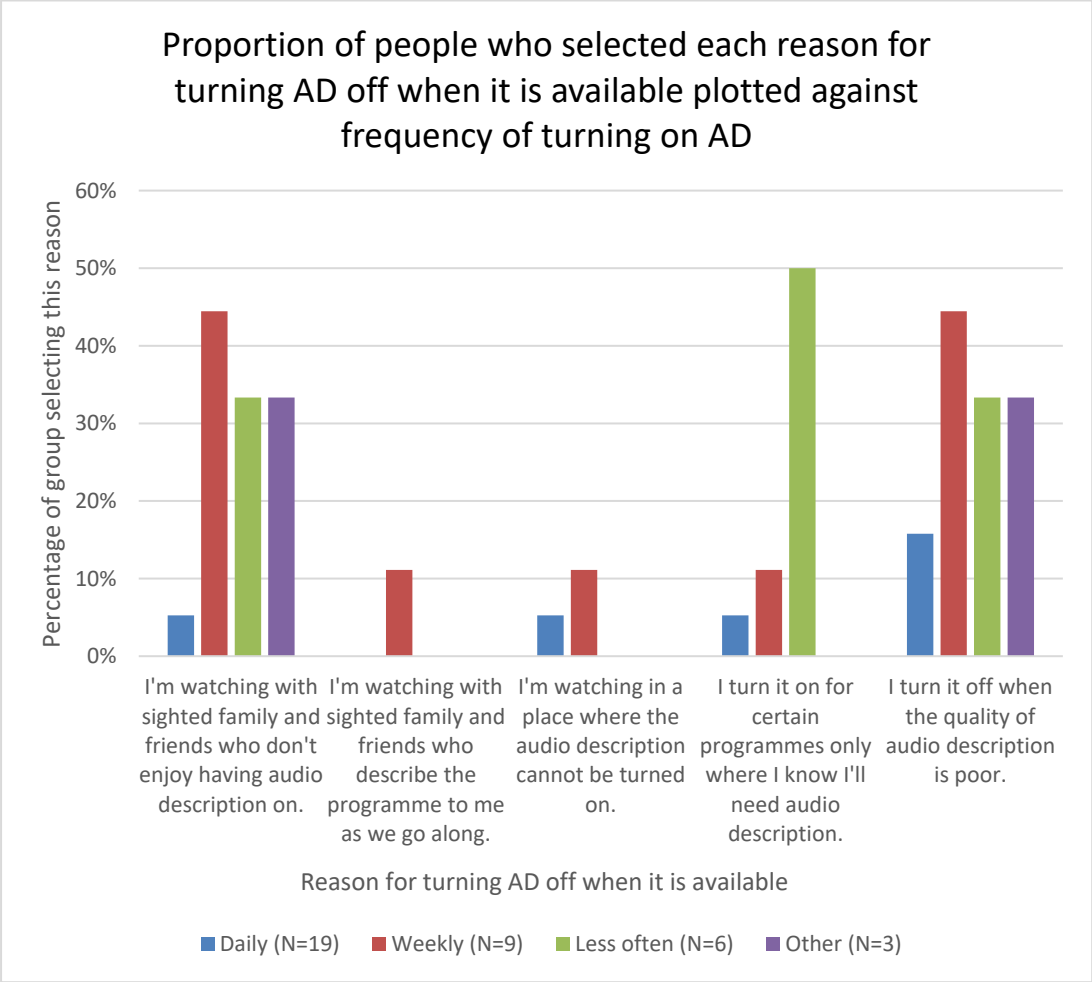

Fig 21. Proportion of people who selected each reason for turning AD off when it is available plotted against frequency of turning on AD

1.16.6 Reasons for turning off AD on TV

Reasons for disabling AD varied by frequency of usage:

- Daily AD Users: A small proportion (all under 20%) selected reasons for turning AD off, with none citing “I’m watching with sighted family and friends who describe the program to me.”

- Weekly AD Users: More likely to choose various reasons, except for “I turn it on for certain programmes only where I know I'll need AD” or “Other.”
- Occasional Users: Primarily turned off AD when they knew it would be unnecessary for specific programs. The top reasons included poor quality (one-third) and watching with people who didn’t enjoy AD, with a few selecting “Other.”
- Other’ Group: Two-thirds identified as audio describers and chose “I turn it off when the quality of audio description is poor,” with one-third citing “watching with sighted family and friends who don’t enjoy AD.”

1.17 Discussion

1.17.1 Information vs. entertainment

Participants viewed synthetic voices as more suitable for informational content, such as books, academic articles, and journals, particularly when used for education or work. AD was the next most acceptable category, with users recognising that AD could span both informative and entertainment contexts. Leisure reading, however, received less interest for synthetic voices, as many participants preferred the emotional engagement of human narration for entertainment.

Feedback aligned with this view, as scoring revealed that clips with factual or documentary content, such as the documentary and factual reality clips, received the highest ratings. Participants valued synthetic AD most in genres where the primary function is informational, while entertainment segments—requiring more nuance—were rated lower, indicating that emotional resonance may be lacking in synthetic AD for entertainment-oriented content.

One participant noted, “Synthetic AD has its place especially for factual content, however caution needs to be made on more emotive content where human emotion and expression are to be portrayed,” reflecting the general consensus that synthetic voices are best suited for content that is not emotionally complex.

1.17.2 Matching AD tone to content

Several participants emphasised the importance of tone and emotion in AD, suggesting that AD should feel integrated with the content’s tone rather than added as an external narrative. Synthetic AD, they noted, often sounds neutral or detached, which

can hinder immersion, especially for emotive genres. For example, comments highlighted how specific genres, such as theatre or drama, require “human emotion that fits the tone of the programme.” This expectation outlines the need for AD to align with the content’s energy and mood to create a cohesive experience.

Participants expressed concern that synthetic voices could lose the nuance of human reaction, and they worried about future AI potentially being unable to provide the “human views and decisions” needed to enhance leisure-oriented AD experiences. As one participant remarked, “I would prefer human for books for leisure and AD as hearing the humanity in a voice is wonderful,” highlighting the value users place on a human touch, particularly for genres that benefit from emotional resonance and responsiveness.

1.17.3 Sound mixing

Sound levels emerged as an unprompted yet significant concern in this study, highlighting participants' desire for balance between AD and the program’s soundtrack. This unprompted feedback suggests that sound quality and mixing are integral to effective AD, as background sounds can interfere with clarity. Participants specifically mentioned challenges with audio ducking—where the volume of the program audio is reduced to allow AD to be heard—indicating that achieving a balanced mix is crucial.

As one participant put it, “Attention does need to be paid to all of the different ‘sound levels,’ as background noise interferes with the hearing of everyone,” suggesting that poor sound quality or overpowering background audio detracts from the AD experience, particularly with synthetic voices, where clarity is paramount.

2 Report on the qualitative study on the use of synthetic speech for audio description

2.1 Introduction

This section of the report presents findings from a qualitative study examining synthetic voices in audio description (AD).

A focus group was organised to assess the quality of AD delivered with synthetic voices, exploring how these voices impact the

viewer experience. Participants provided insights on AD's clarity, naturalness, emotional resonance, and suitability across various content types. This session also offered a platform to discuss the broader implications of using synthetic voices for AD, touching on potential benefits and limitations. In order to preserve objectivity participants were not told that the voices were synthetic until after they have commented on each of the clips. Some participants guessed that some of the clips used synthetic voices however, leading to speculation about 'which' of the voices were computer generated.

2.1.1 Participants

Seven people with sight loss were recruited through RNIB's community networks (see Table 1). Participants shared a broad range of experiences with AD with a variety of interests, including watching TV, reading, and playing sports. The session was hosted by RNIB's Media, Culture, and Immersive Technologies Team.

Participant ID	Visual impairment status
01	Blind
02	Blind
03	Blind
04	Blind
05	Blind
06	Blind
07	Blind

Table 1: Focus group participants

This focus group highlighted experiences and perspectives from people with various degrees of sight loss, emphasising a common interest in accessible media and AD.

2.1.2 Stimuli

All participants completed a survey to share their responses to six sample clips featuring synthetic AD. The sample clips represented a diverse range of genres, including entertainment, drama, sports, documentary, and factual content:

1. Strictly Come Dancing (entertainment)
2. I'm a Celebrity (gameshow)
3. Rugby match (sports)
4. Riches (drama)
5. Living Next Door to Putin (documentary)

6. DIY SOS (factual reality)

These clips provided a basis for exploring how synthetic voices perform in delivering AD across different genres and allowed participants to express their preferences and concerns.

2.2 Key themes from the discussion

Themes were identified from transcripts after the session. These were then agreed within the research team.

2.2.1 Authenticity and emotional engagement

Participants noted that the voices often fell short in conveying the authentic, context-sensitive emotional engagement expected in AD, particularly in genres like sports and drama that benefit from dynamic delivery. This issue was highlighted in the rugby clip, where the synthetic AD voice sounded “bored” and “disconnected,” diminishing the intensity of the experience. As one participant noted:

“I thought the rugby was particularly unnatural. It just seemed a bit odd. I think it might be worth removing all the standard commentary and having an actual commentator doing it to make it more involved. You want someone who's passionate about what they're watching. It will feel like you're more involved. It almost sounded a little bit like an AI audio describer.”

For these participants, emotional alignment in AD was crucial for engagement. The group felt that human narrators are better able to adjust their tone and style to suit each moment, which keeps the AD lively and immersive. By contrast, synthetic voices’ lack of emotion often led to a flat and unengaging experience.

2.2.2 Humanising elements in audio description

The preference for humanised AD emerged strongly as participants appreciated when narrators expressed genuine reactions in response to on-screen events. These unscripted responses, like laughter or excitement, were viewed as adding warmth and relatability to AD—qualities that synthetic voices struggle to replicate. As one participant recalled:

“My favourite ever time in an audio-described theatre performance was when something unexpected happened, and the audio describer laughed. That, to me, really humanises things, which an AI voice can’t do. But that’s beside the point. I think being as human as possible and allowing for things like that is brilliant.”

Another participant cited examples from well-loved movies:

“There’s this describer who does The Avengers and a lot of the Marvel films, and he gets so involved, he gets so excited. I really enjoy listening to him.”

These experiences underscore that synthetic voices, lacking in spontaneity and emotional range, struggle to replicate the natural reactions that deepen viewers’ connection to content.

2.2.3 Clarity and pronunciation

Synthetic voices received mixed reviews regarding clarity. Generally, participants found the synthetic AD intelligible, but occasional inconsistencies—such as a monotonous tone in Strictly Come Dancing or overlapping AD in sports segments—detracted from the overall quality. One participant summarised:

“I’d say 98–99% of the time, I thought I understood absolutely everything that was being said. There were a few occasions... where there was a little bit of overlap between the AD and the commentators.”

This feedback points to a need for synthetic voices to maintain clear and accurate pronunciation while balancing clarity with contextual engagement.

2.2.4 Timing and accessibility challenges

Timing and synchronisation emerged as significant challenges with synthetic voices, particularly in fast-paced content like sports. Participants noted that synthetic AD sometimes failed to match the program’s rhythm, resulting in a disjointed viewing experience. For example:

“The rugby description was too slow, and then, as we said earlier, the Strictly clip started at a steady rhythm but then sped up suddenly. That made it feel a bit jarring.”

Although the instance in Strictly could be attributed to the production quality of the described tracks in synthetic, there were other irregularities that highlight the need for synthetic voices to better adapt to the natural pacing of different genres.

2.2.5 Accent and localisation

Aligning the accent of the AD voice with those heard in the program emerged as a significant factor in enhancing viewer immersion. Participants discussed the importance of contextually appropriate accents, with broad consensus that an AD voice matching the program's setting contributes greatly to the authenticity of the experience. This alignment was seen as enhancing the connection to the content, creating an immersive experience that participants felt was particularly effective in certain genres like drama.

In this study’s clip from Riches, the main character speaks English with a South African accent, and it was attempted to match the accent of the AD voice to the character’s accent. One participant noted the impact of hearing an accent familiar to their own background, which positively influenced their engagement with the clip:

“I grew up in South Africa, and the audio describer reminded me very much of the accents in the area where I grew up. So, yeah, it was interesting, and it is good to have similar accents to what’s going on in the programme, or people from similar backgrounds. I didn’t enjoy all of it, but I enjoyed most of it.”

An accent tailored to the accent of the main character in the clip increased the perceived naturalness of AD, and participants suggested that synthetic voices could benefit from replicating relevant accents to further boost immersion.

In addition to accents, participants also touched on the localisation of the AD script to align with the content’s cultural origin. Although not the main focus of this study, one participant noted the dissonance in watching the British show Slow Horses with an

American AD voice, suggesting that matching not only the accent but also colloquialisms to the program's origin improves immersion:

"At the moment, my partner and I have been watching Slow Horses on Apple TV, and the audio description is great, but it's an American accent, and it's a thoroughly British programme. So, you kind of feel like they missed a trick. It's kind of a tongue-in-cheek British programme, so they could have had more fun with the audio description with a British audio describer, I think. I quite appreciate it when they try to match the accent and colloquialisms to the programme itself. I think that works quite well."

The insights suggest that future AD efforts, whether synthetic or human, could benefit significantly from considering both accent alignment and script localisation to provide a more immersive and contextually accurate experience for viewers.

2.2.6 Continuity and familiarity of voices

Consistency in AD narration was widely appreciated, with participants highlighting that familiar voices across episodes or series contribute to viewer comfort and enjoyment. Drawing parallels to audiobooks, participants likened the role of an AD narrator to that of a favourite reader whose consistent voice adds a layer of continuity. Participants shared:

"I'm thinking about audiobooks which are sort of similar. We may have favourite readers, and when we can't get that reader, and someone else takes over, the book becomes less enjoyable."

"It's almost like having a favourite actor or actress, isn't it? You hear a describer's voice you recognise and think, "Right, they're going to be good at this."

The importance of continuity supports the idea that familiar voices can make AD more engaging, emphasising that synthetic voices could provide value if they offer consistent tonal quality across content.

3 Discussion and future considerations

This study highlights the potential for synthetic voices in audio description (AD) while highlighting the current limitations in terms of their suitability, especially for emotionally engaging content. Findings from both the qualitative and quantitative components of this study reveal key considerations for future implementation, emphasising areas where synthetic voices must improve to meet user expectations in AD.

3.1 Key themes

3.1.1 Emotional adaptability and content suitability

While synthetic AD shows promise in information-driven content, such as factual documentaries and non-fictional material, it currently lacks the emotional resonance required for entertainment and narrative-driven genres. Participants expressed a strong preference for human voices in contexts where AD enhances engagement, such as drama or entertainment, where a narrator's emotional tone and sensitivity to content is critical. Synthetic voices were best received in informational contexts, such as factual and non-fictional entertainment content, as these areas require less emotional engagement. As noted by one respondent, "Synthetic AD has its place, especially for factual content; however, caution needs to be made on more emotive content where human emotion and expression are to be portrayed." This distinction suggests a potential role for synthetic voices in specific genres but highlights that their current limitations are a significant barrier to their use across all content types.

3.1.2 Importance of tone and integration with content

The ability of AD to resonate with the content's tone was a central theme in participants' feedback. Many found synthetic voices too neutral or detached, which created a sense of disconnect from the material. This issue was particularly visible in clips from emotionally driven genres like drama and sports. Participants expressed that effective AD should harmonise with the emotional tone of the program, making it feel like a component of the viewing experience. One respondent remarked, "The audio needs to have human emotion that fits the tone of the programme. It needs to naturally blend in with it and not feel like a separate add-on as you're watching it," highlighting the need for synthetic voices to

improve in capturing and delivering emotional tones to enhance immersion.

3.1.3 Sound quality and mixing challenges

Unprompted feedback on sound mixing reveals that sound quality is integral to AD's clarity and effectiveness. Participants highlighted issues with audio ducking and the need for balanced sound levels to ensure AD remains clear and unobstructed by background audio. This feedback suggests that for synthetic AD to be viable, broadcasters must prioritise audio mixing practices, ensuring that AD remains accessible and prominent within the program's soundscape. As one participant noted, "Attention does need to be paid to all of the different 'sound levels,' as background noise interferes with the hearing of everyone." This finding suggests that poor sound quality and inconsistent mixing could exacerbate the limitations of synthetic voices, emphasising the importance of minimum sound quality standards in any synthetic AD implementation.

3.2 Recommendations for minimum quality standards for synthetic speech to be used for audio description

To bridge the gap between current synthetic AD capabilities and user expectations, establishing minimum quality standards for synthetic voices in AD is essential. Key recommendations include:

- **Emotional and tonal adaptability**: Synthetic AD must reflect the tone of various genres, whether informational, fictional, or humorous, to engage audiences effectively. Synthetic voices used in AD for emotional content (sports, drama, entertainment etc) should be capable of delivering a range of emotional tones and modulations to match the mood of the content.
- **Cultural sensitivity and localisation**: Matching accents and cultural nuances to the content improves audience immersion and helps make AD feel more contextually relevant. AD that aligns with the program's setting and audience can significantly enhance the viewing experience.
- **Consistent quality across episodes**: For serialised content, maintaining continuity in voice and tone across episodes is essential. Abrupt tonal changes can break immersion, and thus, consistent voice settings or models should be applied for multi-episode content.

- **Sound quality and mixing standards**: Sound levels for audio description are important regardless of the source of the voice. Any quality standard for AD needs to encompass the mixing levels.
- **Quality of synthetic voices:** Participants were concerned that synthetic voices would lower the quality of AD. Any voices used must match the 'quality' of a human voice and a Mean Opinion Score of the type used in this research could be used to develop a standardised test for the quality of a synthetic voice.

3.3 Conclusion

As synthetic speech technology evolves, its role in AD may expand, but minimum quality standards must be implemented to ensure the quality of AD does not decline. The findings in this report suggest that human narration remains preferable for most AD users, especially in emotive or narrative-driven content. AD users value humanising aspects of AD tracks and also prefer that the AD track matches the main content in terms of emotions, language style and accents.

If synthetic voices are to be adopted for AD, broadcasters and AD providers must actively involve user feedback to refine synthetic voice quality and maintain high standards that cater to the needs of AD users. This approach will be crucial to making synthetic AD a viable option that complements, rather than detracts from, the viewing experience.

---

[1] Audience expectations of access services across TV and Video-on-Demand Services. RIDC Ofcom 2024

[2] https://www.bbc.co.uk/rd/blog/2018-12-online-news-journalism-review

[3] https://www.bbc.co.uk/rdnewslabs/projects/script

[4] https://www.bbc.co.uk/rdnewslabs/projects/alto

[5] https://marhamilresearch4.blob.core.windows.net/gutenberg-public/Website/index.html

[6] Audience expectations of access services across TV and Video-on-Demand Services. RIDC Ofcom 2024

[7] Lewis, J. R. (2018). Investigating MOS-X ratings of synthetic and human voices. Voice Interaction Design, 2(1), 22