Professional and novice audio describers: quality assessments and audio interactions

Sawako Nakajima*, Graduate School of Engineering Science, Akita University
Kazutaka Mitobe, Graduate School of Engineering Science, Akita University**

ABSTRACT

Empowering novice describers can reduce costs and expand access to high-quality audio descriptions (ADs). This study explored differences between novice and professional practices by analysing their ADs for a 3:42-minute scene from a Japanese fictional film. A film producer rated both the overall quality and volume quality of ADs. The perceived AD volume quality reflects the comprehensive volume experience within ADs beyond loudness. The assessment revealed that ADs created by ten novices using speech synthesis reached approximately 60% of both the overall quality and volume quality of published ADs with human voice. Kernel density estimation showed significantly lower mean loudness in published ADs than in novice ADs. Additionally, a significant negative correlation existed between perceived AD volume quality and mean film loudness during AD presentation across all AD sets. However, published ADs had longer durations compared to novice ADs. Contrasting cueing strategies were observed. Published ADs relied on film sounds, whereas novice ADs leaned on visual cues. Consequently, we developed a professional technique: carefully curating the film information to be heard and balancing AD placement to ensure the audio experience of both ADs and film sound without abrupt AD loudness increases. This sonic approach empowers novices to craft impactful ADs.

KEYWORDS

Audio description, partially blind, blind, film, speech synthesis, novice, loudness, multimodal cohesion, audiovisual translation, accessibility.

1. Introduction

Audio description (AD) is a form of narration that translates visual information in displayed events, such as character actions, expressions, and scene changes, into audio information. Providing for a diverse audience, including people who are blind, partially blind, and sighted, ADs enhance access and enrich their experience by expanding understanding and engagement with films, broadcasts, sports, live theatre, and museum tours. Owing to the growing consumption of online content, the need for

* ORCID 0000-0002-5898-7720; e-mail: nakajima@ie.akita-u.ac.jp

** ORCID 0000-0001-8446-9567; e-mail: mitobe@gipc.akita-u.ac.jp

https://doi.org/10.26034/cm.jostrans.2024.5980
© The Authors

AD has increased. The production process of AD for video content involves numerous steps, including script creation, narration recording, and AD adjustments for video synchronisation. This process necessitates collaboration among a diverse team comprising individuals who are blind or partially blind, audio describers, narrators, sound engineers, directors, and producers. Consequently, production costs are understandably high in both time and money (Lakritz and Salway, 2006; Plaza, 2017; Sade et al., 2012). In Japan, progress towards improving the limited availability of film AD has been slow.

The use of text-to-speech (Fernández-Torné, 2016; Fernández-Torné and Matamala, 2015; Kobayashi et al., 2010; Omori et al., 2015; Szarkowska, 2011; Walczak and Fryer, 2018) and automated AD generation technologies (Bodi et al., 2021; Campos et al., 2020; Campos et al., 2023; Han et al., 2023; Hanzlicek et al., 2008; Kurihara, et al. 2019; Oncescu et al., 2020; Wang et al., 2021) has been proposed to address these issues. However, synthetic and automated AD generation technologies have not yet matched human-generated AD in terms of quality (Fernández-Torné, 2016; Kobayashi et al*.,* 2010; Walczak & Fryer, 2018; Walczak & Iturregui-Gallardo, 2022). Human production remains the predominant method of AD film production and is extensively relied upon in Japan.

Previous studies have revealed the importance of the relationship between film sounds and AD (Fryer, 2010; Lopez et al., 2021, 2022; Remael, 2012; Remael et al., 2015; Reviers, 2018; Vercauteren, 2022; Vercauteren & Reviers, 2022). Remael (2012) noted that in the production of ADs, it is important to understand the role of film sound and the way it is created, as well as the expected changes and effects in perception when visual information is unavailable. In addition, it is important to understand how audiences who are blind or partially blind perceive film sounds. Lopez et al. (2021, 2022) demonstrated the potential to improve ADs under the 'Enhanced AD' project. By creatively employing sound design and binaural elements, they reduced verbal descriptions. Moreover, adding first-person narratives can provide information and enjoyment for audiences who are blind or partially blind in a way similar to what is currently known about ADs. Fryer (2010) examined the film sound–AD connection from the viewpoint of audio drama, advocating for the importance of coherence and constructing a sound effect classification. Analysing audio-described scenes through the lens of 'multimodal cohesion,' Reviers (2018) developed the multimodal transcription model using Fryer's categories. Through this model, Reviers highlighted the essential contribution of sounds in creating rich and vivid ADs and further argued that skilled integration of sound and descriptions enhances cohesion, transforming descriptions from implicit to explicit. Furthermore, Vercauteren and Reviers (2022) proposed a framework for analysing sound in ADs, drawing on Thom's (1999) 16 narratological functions of sound. Their analytical model can be used to apply the perspective of audionarratology to actual AD production, such as the decision-making process of describers. As documented by Sueroj and Lopez (2023), despite the potential of ADs to retain the intended functions of sounds within source materials, Thai AD guidelines give insufficient weight to their significant impact. Similar shortcomings are observable in Japan. Though existing guidelines do not always reflect the crucial role of film sound in AD, this understanding is increasingly recognised. In these ways, the dynamic evolution of AD demands constant innovation and discussion, fostering ever-greater creativity in describers. In addition, effective AD is a challenging process that demands a multifaceted perspective. Describers must empathise with audiences who are blind or partially blind, meticulously select information that aligns with the film's intent and narrative flow and express it vividly in language that complements the soundtrack. A previous study discussing AD as a video grounding benchmark (Soldan et al., 2022) suggested that manually producing high-quality AD and continuing to pioneer the expression of AD would not only improve the quality of automated AD and contribute to its adoption but also improve the quality of video-language research. Considering the further spill-over effect, it is important to develop an AD production support technology capable of delivering high-quality AD at reduced production costs.

Previous studies have proposed employing the services of novice audio describers to reduce AD production costs (Branje and Fels, 2012; Natalie et al., 2023). Natalie et al. (2023) investigated the quality of novice ADs for short videos, revealing challenges in areas such as learning and sufficiency. These findings suggest that novice describers have difficulties understanding the intended meaning of the videos and expressing it in words that are neither excessive nor inadequate. However, these prior investigations did not examine the multifaceted interplay between visual and auditory elements within the context of AD, nor did they propose a methodology for analysing the skill gaps associated with elements that pose difficulty for novice describers, which this study aimed to address.

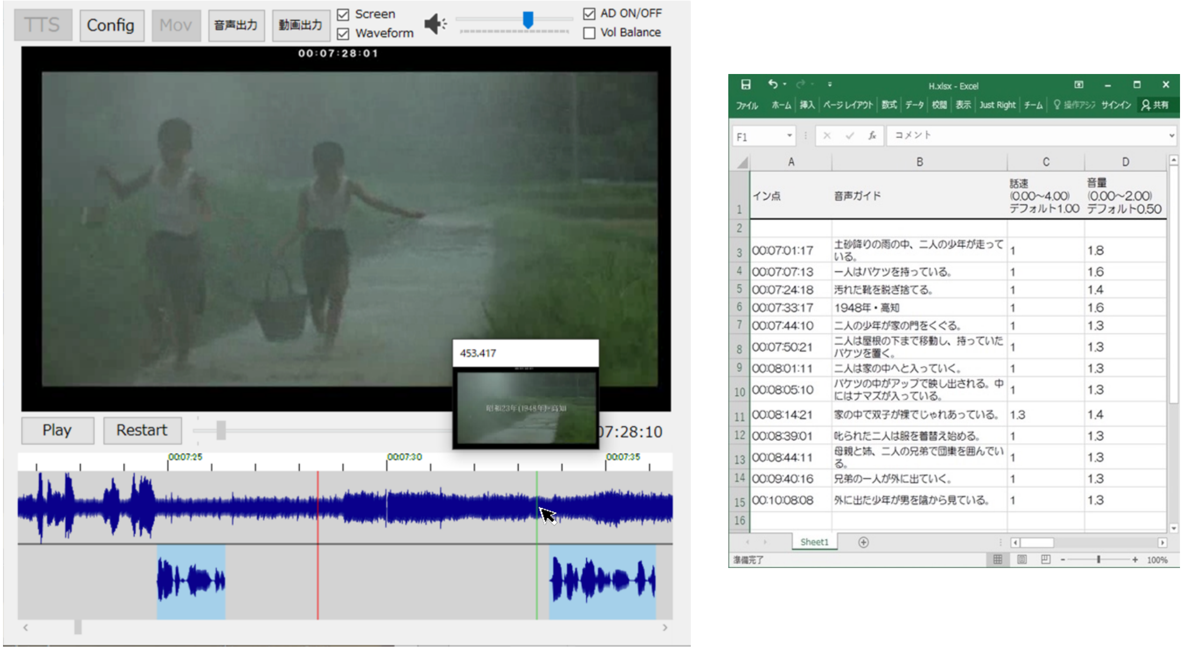
In this study, our focus is on the role of novice AD production in reducing production costs. We aim to identify and quantify the issues that novice describers face when working on long films and therefore to support them to produce high-quality ADs. First, the ADs of a scene in a Japanese fictional film of approximately two hours were produced by ten novice audio describers. Then, the film producer compared the quality of these ADs with that of professionally published versions. Moreover, the volume of the AD and film sound, as well as the factors triggering the timing of the AD presentation, were analysed to gain insights into the issues related to creating novice

ADs from a film soundscape and AD integration perspective. Finally, we discussed how the lack of consideration by novice describers of the sonic connection and flow between the AD and film sound leads to a decline in AD quality. The manuscript is organised as follows: Section 2 describes the methods used in the study. Section 3 discusses the results, and Section 4 presents the inferences drawn from analysing the results. Finally, Section 5 concludes the paper.

2. Methods

2.1 AD production by novice describers

Ten men participants (mean age of 22.3 ± 0.6 years), having no experience in AD production, took part in the experiment. Akita University approved this study based on Article 12(1) of the Code of Ethics for Human Subject Research. Informed consent was obtained from all participants before the start of the experiment. Figure 1, left, shows the main window of the AD production software, which is an improved version of the software developed in our previous study (Nakajima & Mitobe, 2022). The upper part of the main window displayed the film, and the lower part (below the film area) displayed the audio waveforms of the film and the ADs. Hovering the mouse over the audio waveform of the film while holding down the Ctrl key displayed a thumbnail. Right-clicking on the waveform area of the AD displayed a context menu, and clicking 'Create' displayed a window for creating a new AD. Then, the content of the AD, the speech rate (0.00 to 4.00), and the volume (0.00 to 2.00) used to synthesise the AD were entered, and the new AD was inserted at the position of the cursor. The inserted AD was highlighted in a light-blue box, as shown in Figure 1, and its position could be adjusted by dragging and dropping the box parallel to the time axis. The created AD could be edited or deleted using the 'Edit' and 'Delete' menus on the context menu, which were displayed when the light-blue box area of the relevant AD was right-clicked. Figure 1, right, shows the Excel sheet containing the details of the AD script, where the content, speech rate (0.00 to 4.00), volume (0.00 to 2.00), and insertion time of the AD created in the main window (Figure 1, left) were immediately reflected.



	A	B	C	D
1	イン点	音声ガイド	話速 (0.00～4.00) デフォルト1.00	音量 (0.00～2.00) デフォルト0.50
2				
3	00:07:01:17	土砂降りの雨の中、二人の少年が走っている。	1	1.8
4	00:07:07:13	一人はバケツを持っている。	1	1.6
5	00:07:24:18	汚れた靴を脱ぎ捨てる。	1	1.4
6	00:07:33:17	1948年・高知	1	1.6
7	00:07:44:10	二人の少年が家の門をくぐる。	1	1.3
8	00:07:50:21	二人は屋根の下まで移動し、持っていたバケツを置く。	1	1.3
9	00:08:01:11	二人は家の中へと入っていく。	1	1.3
10	00:08:05:10	バケツの中がアップで映し出される。中にはナマズが入っている。	1	1.3
11	00:08:14:21	家の中で双子が裸でじゃれあっている。	1.3	1.4
12	00:08:39:01	叱られた二人は服を着替え始める。	1	1.3
13	00:08:44:11	母親と姉、二人の兄弟で団欒を囲んでいる。	1	1.3
14	00:09:40:16	兄弟の一人が外に出ていく。	1	1.3
15	00:10:08:08	外に出た少年が男を陰から見ている。	1	1.3
16				

Figure 1. Use of AD production software for speech synthesis development. Main window with waveform and thumbnail (left); Excel sheet with the details of the AD script, synchronised with the AD content being edited in the main window (right).

The volume balance between the film sound and the overall ADs was adjusted from 0 to 100. In the experiment, the initial settings were 30 for the film sound and 10 for the AD sound, and the participants were instructed not to change the film sound until the end of the experiment. They could adjust the volume balance with the film sound by adjusting the numerical value of the AD. The personal computer volume was freely adjustable using the software. The AD could be outputted in the WAV format as an individual audio file according to the timing of each insertion. The software also included a function that mixed the film and AD and outputted it as a film file according to the volume balance set between the film sound and AD. The scene used in the AD production experiment was a 3 min 42 s video clip from the 6 min 58 s to 10 min 40 s mark of the Japanese film *Village of Dreams* (Yoichi Higashi, SIGLO Ltd./Palabra Inc., 112 minutes, Japanese, 1996). The AITalk® 4.1 SDK (AI Inc., 2017) software and 'nanako_22' engine voice (pitch = 1.00, range = 1.00) were used to create a synthesised AD. The AD production procedure is detailed below:

(1) Participants wore stereo headsets (Sanwa Supply, MM-HP117BK) connected to a laptop personal computer (MacBook Pro, 13 inch, 2020, Intel Core i5, 16 GB RAM, macOS 11.6.4) and operated the software.

(2) They watched a 30-min introductory AD production video for beginners made by a professional audio describer.

(3) The experimenter then explained the use of the AD production software for 30 min.
(4) After this, the participants created ADs using the AD production software for 1 h.
(5) On the next day, the participants spent 30 min adjusting the insertion position, speaking rate, and volume of the ADs using the AD production software.
(6) Participants were asked to mention the information in the scene to which they matched the 'in' and 'out' points of the AD, to gather details about their AD insertion positions.

Although participants were not asked to watch the entire film before producing the AD, the full-length film script, character roles, and names were provided for reference during the AD production experiment. The professional audio describer (with over 20 years of production experience) who gave the lecture in the introductory AD production video (step (2)) created the published AD of the film used in the experiment. The ADs produced by the participants were referred to in this study as novice ADs.

2.2 Quality assessment of ADs by the film producer

The producer of the film used in the experiment performed qualitative assessments of the produced ADs. In Japanese film AD production, final decisions traditionally lie with the producer and director after receiving input from individuals who are blind or partially blind. For the film under evaluation, although two producers were involved, one producer served as our evaluator. Having spearheaded the film's production and intimately collaborated with the director, the evaluator possessed the most comprehensive understanding of both the film itself and the film's published AD production process. With 40 years of film production experience and 18 years dedicated to AD and subtitling, the evaluator's unique insights, born from decades of film production and deep expertise in AD, guaranteed the precision of the evaluations. Twelve film clips were evaluated based on the produced novice ADs, the published ADs recorded by the human narrator (published human AD), and the synthesised ADs based on the published script (published synthesised AD). Although the published synthesised ADs were not exactly the published ones, they have been denoted as 'published synthesised AD' in this paper to simplify the comparison. The sample rate was 48,000 Hz, the audio format was stereo, the bit depth was FP32 (32-bit single-precision floating point), and the bitrate was 128 kb/s. AITalk® 4.1 SDK (AI Inc., 2017) and nanako_22 (pitch = 1.00, range = 1.00), which were used for the speech synthesis of the novice ADs, were also used for the published synthesised ADs. The content of the published synthesised ADs was the same as that of the published human ADs, and the volume of the ADs was adjusted using the Audacity 2.3.3 plugin dpMeter 5 to ensure that the sentences had the same loudness (LUFS) as the sentences in the published human ADs. Furthermore, the speech rate of the ADs was adjusted to match the presentation time of the published human ADs. No adjustments were made to the pauses (moments of silence) within the sentences of the ADs generated by the engine. The details of the evaluation procedure are described below.

(1) The film clip was watched with no AD.
(2) The film clip with the published human ADs was watched and evaluated.
(3) The film clip with the published synthesised ADs was watched and evaluated.
(4) The film clip with the novice ADs (in order from novices A to J) was watched and evaluated.
(5) The evaluated values were checked and changed if necessary.

A quantitative evaluation of AD was conducted, assessing five key aspects on a scale of 0 to 100 points: script content, insertion position, speech rate, volume, and overall rating.

3. Results

3.1 Quality assessed by the film producer

Table 1 lists the script content, insertion position, speech rate, volume, and overall AD rating assessment scores. The results for the novice ADs are presented as mean values. The published human ADs were used as a benchmark for comparison, and the percentages were calculated based on these scores. The novice ADs achieved, in order, scores of 60.0%, 54.4%, 46.0%, 63.9%, and 61.6%, whereas the published synthesised ADs performed better, achieving scores of 100.0%, 88.9%, 90.0%, 88.9%, and 94.7%.

	Script content	Insertion position	Speech rate	Volume	Overall
Novice ADs	48.0 (11.4)	49.0 (12.9)	46.0 (12.6)	57.5 (15.9)	58.5 (18.6)
Published synthesised ADs	80	80	90	80	90
Published human ADs	80	90	100	90	95

Score on a scale of 0–100 (standard deviation)

Table 1. Quality assessment by the film producer for the novice ADs, published synthesised ADs, and published human ADs.

3.2 Loudness of the AD and film sound

Given the crucial role of the auditory landscape in creating a seamless audiovisual experience, particularly for audiences who are blind or partially blind, this section analyses the volume balance between film sounds and ADs as originally adjusted by novice describers, elaborating potential factors contributing to their lower scores.

First, the film and AD of the novice ADs were mixed using the volume balance values set by the software volume balance adjustment function when the novice ADs were created. Then, the audio of the novice ADs, the published synthesised ADs, and the published human ADs were converted to a mono, 16-bit, and 768 kb/s format as a WAV file. For each WAV file, the momentary loudness (LUFS) was calculated every 400 ms at a rate of 10 Hz using FFmpeg version 5.0.1. Additionally, any silence was eliminated by setting a threshold value using Python 3.9, and each pair of AD and film sounds during the AD presentation was extracted. The novice ADs had a mean total duration of 49.73 s (SD 12.88). The published synthesised and human ADs were 63.40 and 66.80 s long, respectively. We investigated the relationship between this AD duration and the perceived volume quality of the ADs. In this study, 'AD volume quality' was operationalised as the comprehensive auditory experience, specifically the volume of ADs, rated by the film producer using a 0 to 100 scale (details in Table 1). This rating encompassed various aspects of volume, including clarity, loudness appropriateness, pleasantness, and other subjective qualities, without providing specific criteria beforehand. The film producer was simply instructed to 'Please rate the volume quality of the Ads'. The analysis included all twelve AD sets (ten novices, one published synthesised, and one published human) and showed a non-significant correlation coefficient of 0.222. The loudness distributions were obtained via kernel density estimation using the Gaussian_kde class of the Python SciPy library for each AD and film sound. Figure 2, left, shows the loudness distribution obtained by kernel density estimation for a single AD that was cut, and Figure 2, right, shows a pair of density distributions for the loudness of the AD in Figure 2, left, on the horizontal axis and the density distribution of the film loudness on the vertical axis. The coordinate with the largest distribution of AD and film loudness, indicated by the darkest colour, was calculated for all pairs of AD and film sounds.

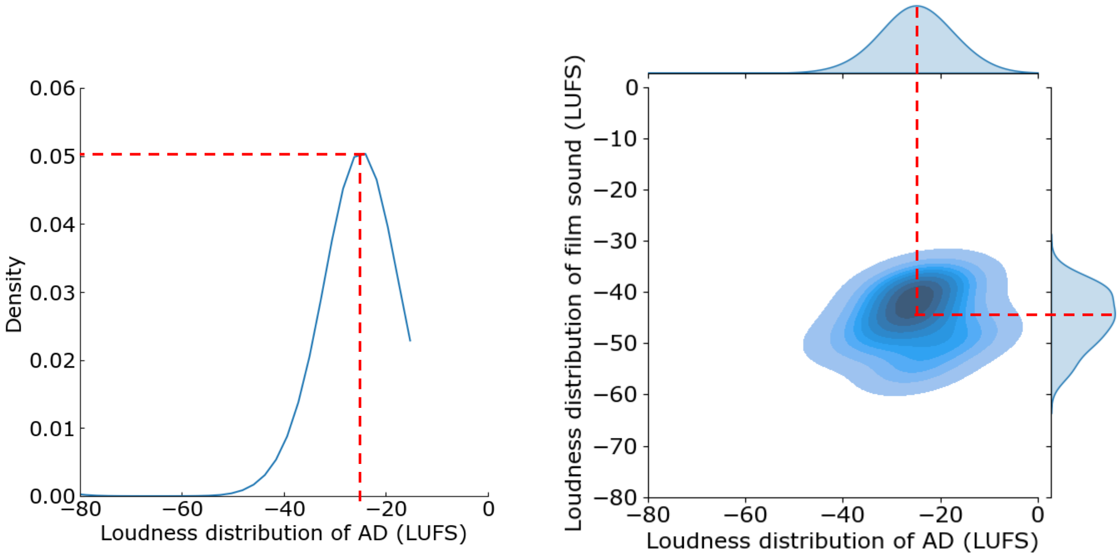


Figure 2. Maximum loudness coordinates of the AD and film sound values calculated by kernel density estimation. Loudness distribution of a single AD (left); loudness coordinates for a pair of the AD and film sound (right).

The shape of the density distribution and the magnitude of the maximum density were different for each AD and film sound. Therefore, the maximum density, considering the magnitude, was obtained by multiplying the maximum loudness obtained by the kernel density estimation from the depth of its density. Figure 3 shows the maximum loudness density coordinates obtained after density correction. There were 197, 28, and 30 samples for the novice ADs, published synthesised ADs, and published human ADs, respectively. Figure 3 shows that there was no difference in the distribution of the density-weighted maximum loudness of the film sound (shown on the vertical axis) between the novice ADs and the published synthesised ADs or between the novice ADs and the published human ADs. By contrast, the distribution of the density-weighted maximum loudness of the AD (shown on the horizontal axis) was shifted to a higher range in the novice ADs than in the published ADs.

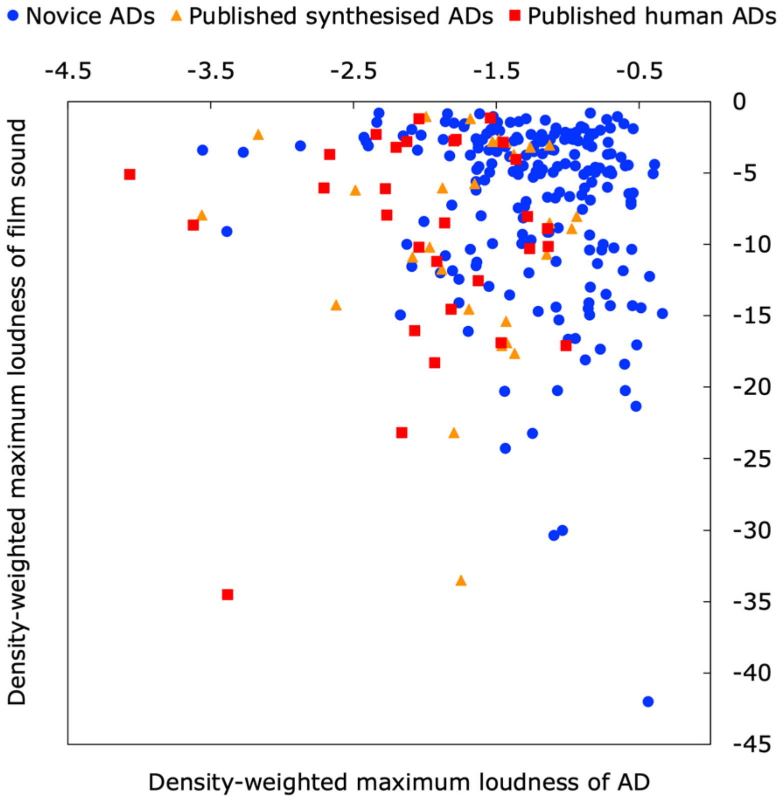


Figure 3. Density-weighted maximum loudness coordinates of the AD and film sound values calculated by kernel density estimation.

Figure 4 plots the mean density-weighted maximum loudness values of the ADs and film sound during the AD presentation for the novice ADs, published synthesised ADs, and published human ADs. Statistical analysis was performed using the Mann–Whitney U test and the Holm–Bonferroni method. The results showed no significant difference in the film sound. However, regarding the ADs, the values of the novice ADs were significantly higher than those of both the published human and synthesised ADs at the 0.1% level. However, there was no significant difference in these values between the published human ADs and published synthesised ADs. The Mann–Whitney U test provided an effect size r of 0.28 between the novice ADs and the published synthesised ADs, and 0.38 between the novice ADs and the published human ADs.

Figure 5 shows the correlation between the density-weighted maximum loudness of the film sound during the AD presentation and the AD volume quality score assessed by the film producer for all ADs, including the novice ADs and published ADs. The correlation coefficient of −0.636 revealed a significant negative correlation at the 0.5% level, with the slope and intercept of the line of best fit being −0.069 and −3.19, respectively.

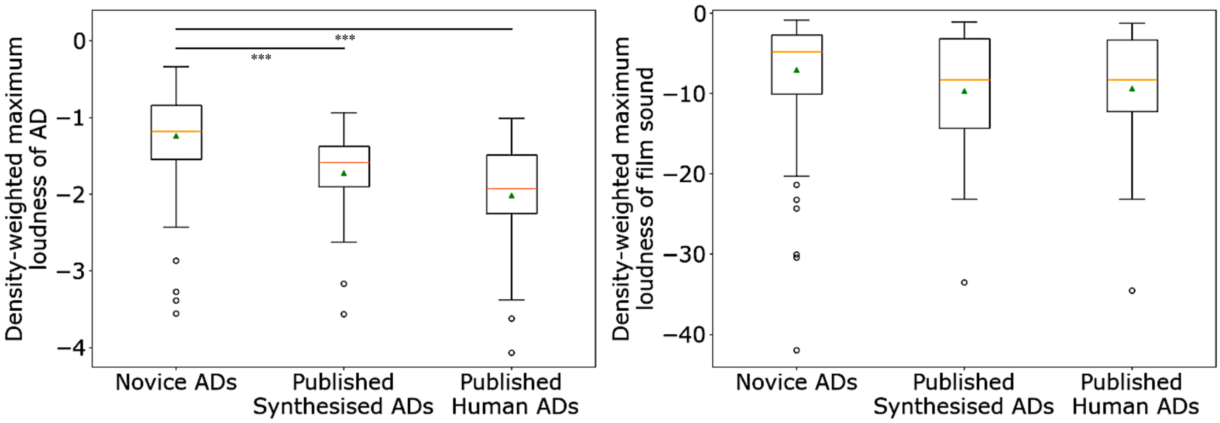


Figure 4. Mean density-weighted maximum loudness. AD (left); film sound (right).

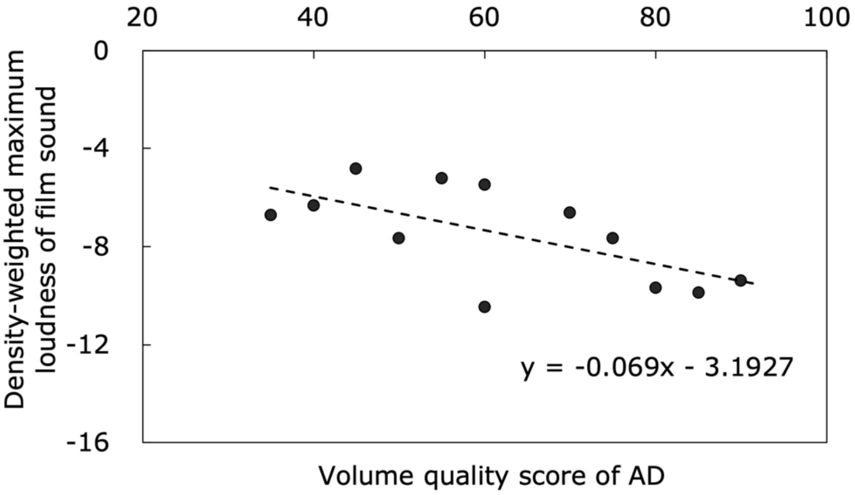


Figure 5. Correlation between the density-weighted maximum loudness intensity of the film sound during the presentation of AD and the AD volume quality score for all ADs, including the novice ADs and published ADs.

3.3 Ratio of the ADs triggered by audio information to those triggered by visual information

Beyond volume balance, this section delves deeper into the factors behind the quality disparity between novice and published ADs, specifically by examining the intentional choices and considerations guiding novice describers' AD placement.

The mean number of ADs produced by the novice describers was 19.2 (±3.9), and that of published ADs was 24. Figure 6 shows the types of information in the film content that trigger the insertion of the novice and published ADs. At the end of the novice AD production experiment, participants were asked to mention the kind of information in the scene that they matched the 'in' and 'out' points of the AD to (triggers). The responses were classified into six categories. In the figure, the proportion of ADs triggered by auditory information ('sound trigger') in terms of character lines and environmental sounds is shown in red, and the proportion of ADs triggered by visual information ('picture trigger') in terms of character actions, message displays (telop appearances), scene changes, and cut changes is shown in blue. Additionally, the figure shows that 55.7% and 44.3% of the novice ADs used visual and auditory information, respectively, as the AD insertion point. On the contrary, 25.0% and 75.0% of the published ADs used visual and auditory information, respectively, as the AD insertion point.

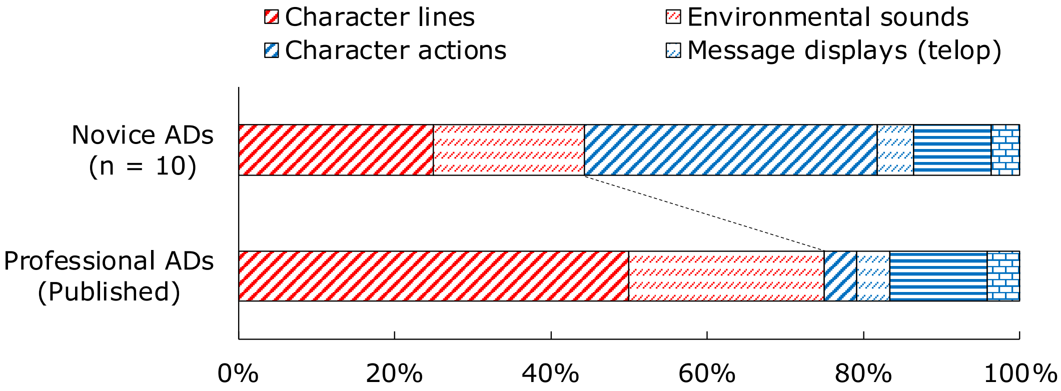


Figure 6. Types of information triggers in film content that determine the insertion position of the ADs created by novice and professional describers.

4. Discussion

In this study, to identify the specific issues faced by novice describers when producing AD and suggest suitable measures to mitigate them from a multimodal cohesion perspective, the ratings of the film producer were produced based on five codes: script content, insertion position, speech rate, volume, and overall rating of AD, with a maximum score of 100 points for each. Table 1 shows that the quality, volume, and overall rating of the novice ADs were 63.9, and 61.6% of those of the published human ADs. The mean volume score for the novice ADs was higher than that for the other codes for script content, insertion position, and speech rate; however, the individual differences were large.

Because the novice ADs were synthesised, whereas the published ADs were human generated, there were concerns about the influence of voice quality. Therefore, we also evaluated published synthesised ADs, in which only the voices of the recorded ADs were changed to the synthesised voices used for the novice ADs. The published synthesised ADs were slightly inferior to (up to 10 points lower than) the published human ADs under all the codes except content, with volume and overall scores of 88.9% and 94.7%, respectively. As described in the Methods section, the duration, speech rate, and volume of the synthesised ADs resembled those of the published human ADs. However, because the manner of speaking within a sentence depended on the speech synthesis engine, the pauses, speech rate, and volume within a sentence differed from those of the human voice. These differences were potentially responsible for the lower scores of the synthesised ADs, suggesting that only 11.1% of the quality loss in terms of volume was attributable to the synthesis engine used for the novice ADs and that the remaining 25.0% (36.1% decrease from the published human ADs to the novice ADs minus 11.1% decrease from the published human ADs to the published synthesised ADs), which accounted for approximately 70% of the quality loss, was because of the production skills of the novice describers.

Next, a quantitative analysis was performed to examine the momentary loudness (LUFS) of the ADs and film sounds, aiming to determine the reason for the decrease in the volume quality of novice ADs. The maximum loudness of the AD and film sound during the presentation of the AD was calculated using kernel density estimation, and the loudness with the maximum density was multiplied by the density to obtain the density-weighted maximum loudness. The analysis results showed that the loudness of the published human ADs and published synthesised ADs was significantly lower than that of the novice ADs. These results indicated that the published ADs, which had considerably higher volume quality than the novice ADs, maintained audibility without any increase in the volume of the ADs. In contrast, for film sounds during AD presentation, the mean value of the density-weighted maximum loudness calculated for each describer, including the published ADs, had a significant negative correlation with the volume score; the higher this score, the better the insertion of the AD when the film volume was low. Notably, the results showed no correlation between the AD presentation time and the volume score. Additionally, the published ADs were considerably longer compared to the average novice AD. Therefore, it can be concluded that the volume of the ADs and film during the AD presentation was set lower in the published ADs, but this cannot be attributed to the shorter AD presentation time.

We explored techniques acquired by professional developers to create ADs that would be easily audible, even at reduced AD volumes. The sound and the manner in which an AD is matched in a film strongly affect the ease of hearing an AD and understanding the corresponding content. If AD insertion is performed without considering the relationship between the AD and the film sound and flow, and only the timing of the displayed images to be explained by the AD is considered, the overlap with the sound of the film will probably increase. To determine the extent to which the novice describers were able to account for the relationship between the AD and film sound, we asked each describer to mention the information in the film that aligned with the insertion position of each AD. Figure 6 shows that most of the published ADs were triggered by auditory information such as character lines and environmental sounds, whereas a quarter (25%) of them were triggered by visual information such as character actions, message displays (telop appearances), and scene and cut changes. In contrast, more than half (55%) of the novice ADs were triggered by these visual cues. This discrepancy was supported by a comment from the professional describer who was the instructor in the novice AD production experiment, in which the novice describers often described the trigger for AD insertion as visual information rather than auditory information. The results of this study revealed that novice describers tended to be distracted by the visual information explained by the ADs and neglected to consider the sonic connection and flow between the film and the AD, which is important for people who are blind or partially blind to understand the content of a film easily and in depth.

This study revealed that novice describers often struggle to create a strong sonic connection between the film and AD. We analysed responses to the film information in which the insertion position (in-point, out-point) of the AD was aligned with the speech waveforms of the AD and film. An alternative perspective states that the visual information explained by the AD should be presented simultaneously with the AD. However, for instance, when adjusting the insertion position of an AD that explains the actions of certain characters, it is often necessary to connect the AD to the beginning or end of the lines spoken by the characters. This is not necessarily done when the actions occur to avoid drowning out the lines of the actions, but rather to enhance understanding of the subject and situation described by the AD, as well as the flow of the actions. To ensure that the audience hears the environmental sound or the sound of the movements of the characters, the AD may be presented slightly faster or later than the timing of these movements. In this study, all the novice describers intended to place the ADs such that they did not overlap with the character lines. However, there were differences between the novice describers in their level of attention to environmental and character movement sounds. For example, in the film clip of the experiment, there was a scene where twin boys were having dinner with their family when a neighbour visited them. The neighbour was angry because the twins had left a bucket of fish in the shed. Then, one of the twins urged the other to get up and throw the bucket of fish away by whispering, 'Go, go!' (いけー、いけー). In this scene, novice H, who had the highest AD volume quality score (85 points) among the novice describers, inserted the AD 'One of the brothers is going outside' (兄弟の一人が外に出ていく) after the line 'Go, go!' by the other, accounting for the auditory information of the whisper. By contrast, novice C, who had the lowest AD volume quality score (35 points), inserted the AD 'Yukihiko pushes Seizo’s shoulder' after being triggered by the visual information of the character touching his twin’s shoulder. As a result, the AD of novice C drowned out more than half of the whispered 'Go, go!' line. After the twins performed this action, their mother sighed. Novice A, who had a low AD volume quality score (45 points), inserted the AD 'Yukihiko is tapped on the shoulder by Seizo and goes to throw the fish away' to match the visual information of Yukihiko standing up. Consequently, the AD of this novice overlapped with the sigh uttered by the mother, resulting in it being only slightly audible in the form of noise.

Furthermore, even when the AD was inserted as a trigger for auditory information, a difference existed between the novice describers in terms of their attention to ensuring that the AD was related to the film sound. Before the aforementioned scene, there was one in which the twins ran into their houses during a storm, carrying fishing rods and buckets. There were some sheep in the yard with a roof, and their bleats of 'Mae, Mae' could be heard without overlapping with the conversation of the twins. For this scene, novice H placed the AD 'They go into the house' when the twins were running to avoid any overlap between the sound of their running footsteps and the subsequent bleating of the goats. This allowed the AD to be heard clearly, without any overlap with the bleating. By contrast, novice C inserted the AD 'Running to the main house' a little after their footsteps. However, as he did not consider the bleating, the AD overlapped with all the first bleats. Novice I, whose AD volume quality score (55 points) was slightly below the mean, inserted the AD 'Seizo and Yukihiko run into the house' after their dialogue. As in the case of novice C, the bleating overlapped with the AD by more than half and could not be heard clearly. Overall, these ADs used auditory information from the film as the insertion point. However, the impression of the film potentially changes depending on the attention of the describer towards the effect of the AD on the film sound.

These results suggest that ADs with a high volume quality considered the detailed environmental sounds that were important for the direction of the film, and that the describers were able to devise ways to ensure that the audience listened to these sounds carefully. A more multifaceted analysis is still needed to evaluate all the quality factors of the volume. Currently, the analysis of the loudness of the AD, the loudness of the film during the AD presentation, and the information that triggers the insertion of the AD provide valuable clues for extracting and improving novice ADs.

5. Conclusion

In this study, ten novice describers with no experience in AD production created an AD for a 3 min 42 s scene of an approximately two-hour Japanese film using AD production software. The results of the quality evaluation by the film producer showed that the novice ADs achieved an overall quality of 61.6% when the quality of the published AD was set to 100. The AD volume quality was 63.9%. After considering the reductions in voice quality owing to the use of synthesised voices, it was inferred that the remaining 25.0% of the 36.1% decrease was due to the production skills of the novice describers. For the novice and published AD, kernel density estimation was used to estimate the maximum loudness (LUFS) of the AD and film sound during the AD presentation, and the density-weighted maximum loudness was obtained by weighting the depth. The loudness values of the published ADs were significantly lower than those of the novice ADs. In contrast, the correlation analysis showed that the density-weighted maximum loudness of the film sound during the AD presentation was significantly lower when the volume quality of the AD was higher. By contrast, the total length of the AD duration of the published ADs was larger than the mean of the novice ADs. From these analyses, we inferred that professionally produced ADs were appropriately inserted at lower film volume levels to facilitate hearing, even with reduced AD volume, without reducing the amount of information. Responses to the question regarding the insertion position of the AD revealed that novice ADs were less triggered by auditory information than by visual information, whereas published ADs showed the opposite pattern. Furthermore, a detailed analysis of the individual AD scripts and the waveforms of the AD and the film sound revealed that novice ADs with high volume quality could be positioned in such a way that low-level environmental sounds could be heard. Based on these results, we hypothesised that novice describers tend to disregard the sonic connection and flow between the AD and film sound more than professional describers, leading to a quantitative change in the loudness of the AD and the film, thereby resulting in a decrease in the sound quality of the AD.

However, our study has several limitations. Although the film producer’s insights as an evaluator were valuable because of the intimate understanding of the film and its AD, future evaluations should prioritise the inclusion of the perspectives of individuals who are blind or partially blind and film directors to gain a more nuanced understanding of the sound–AD connection and its effectiveness for the target audience. Additionally, stricter controls are crucial for future evaluations with broader participant groups, encompassing blind and partially blind individuals, film directors, and other producers, to minimise potential order bias. Finally, acknowledging the gender bias introduced by our all-man novice describer sample, future experiments should require balanced participant groups in terms of age and gender to ensure that valid and generalisable findings are applicable to a wider range of film AD productions.

In future work, it will be necessary to confirm statistically whether the results of this study still hold when ADs are produced for different scenes and genres of films. In addition to the loudness analysis of the sound of the AD and the film, we aim to develop a method for evaluating the degree of sonic connection between the sound of the film and that of the AD, which can support the design of concrete measures to improve auditory landscape integration in AD production. The effective audio design of AD is one of the factors that influences overall quality, and its improvement can lead to the dissemination and expansion of AD expression. Furthermore, adopting a descriptive process that considers the sonic storytelling approach in AD enables people who are blind or partially blind to develop an enhanced understanding of the situations. These initiatives are anticipated to significantly enrich the lives of the blind and partially blind community as a whole.

Acknowledgements

This study was supported by Grant-in-Aid for Scientific Research (C) JP17K01553 and 22K12330. We thank Ryohei Kuriyama for his help with software development and data analysis.

References

AI Inc. (2017). AITalk SDK. https://www.ai-j.jp/products/sdk/

Bodi, A., Fazli, P., Ihorn, S., Siu, Yue-Ting, Scott, A. T., Narins, L., Kant, Y., Das, A., & YoonI, I. (2021). Automated video description for blind and low vision users. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Article 230, 1-7. ACM.

Branje, C. J., & Fels, D. I. (2012). LiveDescribe: Can amateur describers create high-quality audio description? *Journal of Visual Impairment & Blindness*, *106*(3), 154-165.

Campos, V. P., Araújo, T. M. U., Souza Filho, G. L., & Gonçalves, L. M. G. (2020). CineAD: a system for automated audio description script generation for the visually impaired. *Universal Access in the Information Society*, *19*, 99-111.

Campos, V. P., Gonçalves, L. M. G., Ribeiro, W. L., Araújo, T. M. U., Do Rego, T. G., Figueiredo, P. H. V., Vieira, S. F. S., Costa, T. F. S., Moraes, C. C., Cruz, A. C. S., Araújo, F. A., & Souza Filho, G. L.

(2023). Machine generation of audio description for blind and visually impaired people. *ACM Transactions on Accessible Computing*, *16*(2), 1936-7228.

Fernández-Torné, A. (2016). *Audio description and technologies: Study on the semi-automatisation of the translation and voicing of audio descriptions* [Unpublished doctoral dissertation]. Universitat Autnoma de Barcelona, Barcelona, Spain.

Fernández-Torné, A., & Matamala, A. (2015). Text-to-speech vs. human voiced audio descriptions: A reception study in films dubbed into Catalan. *The Journal of Specialised Translation*, *24*, 61-88.

Fryer, L. (2010). Audio description as audio drama. A practitioner's point of view. *Perspectives*, *18*(3), 205-213.

Han, T., Bain, M., Nagrani, A., Varol, G., Xie, W., & Zisserman, A. (2023). AutoAD: Movie description in context. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18930-18940. IEEE.

Hanzlicek, Z., Matousek, J., and Tihelka, D. (2008). Towards automatic audio track generation for Czech TV broadcasting: Initial experiments with subtitles-to-speech synthesis. In B. Yuan, M. Ruan, & X. Tang (Eds.) *Proceedings of the 2008 9th International Conference on Signal Processing*, 2721-2724. IEEE.

Kobayashi, M., O'Connell, T., Gould, B., Takagi, H., & Asakawa, C. (2010). Are synthesized video descriptions acceptable? *ASSETS '10: Proceedings of the 12th International ACM SIGACCESS Conference on Computers and Accessibility*, 163-170. ACM.

Kurihara, K. Imai, A., Seiyama, N., Shimizu, T., Sato, S., Yamada, I., Kumano, T., Tako, R., Miyazaki, T., Ichiki, M., Takagi, T., & Sumiyoshi, H. (2019). Automatic generation of audio descriptions for sports programs. *SMPTE Motion Imaging Journal*, *128*(1), 41-47.

Lakritz, J., & Salway, A. (2006). *The semi-automatic generation of audio description from screenplays.* (Dept. of Computing Technical Report CS-06-05). University of Surrey.

Lopez, M., Kearney, G., & Hofstadter, K. (2021). Enhancing audio description: Inclusive cinematic experiences through sound design. *Journal of Audiovisual Translation*, *4*(1), 157-182.

Lopez, M., Kearney, G., & Hofstädter, K. (2022). Seeing films through sound: Sound design, spatial audio, and accessibility for visually impaired audiences. *British Journal of Visual Impairment*, *40*(2), 117-144.

Nakajima, S., & Mitobe, K. (2022). Novel software for producing audio description based on speech synthesis enables cost reduction without sacrificing quality. *Universal Access in the Information Society*, *21*, 405-418.

Natalie, R., Tseng, J., Kacorri, H., & Hara, K. (2023). Supporting novices author audio descriptions via automatic feedback. In A. Schmidt, K. Väänänen, T. Goyal, P. O. Kristensson, A. Peters, S. Mueller, J. R. Williamson, & M. L. Wilson (Eds.) *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Article 77, 1-18. ACM.

Omori, K., Nakagawa, R., Yasumura, M., & Watanabe, T. (2015). Comparative evaluation of the movie with audio description narrated with text-to-speech. *IEICE Technical Report*, *114*(512), 17-22.

Oncescu, A.-M., Henriques, J. F., Liu, Y., Zisserman, A., & Albania, S. (2020). QUERYD: A video dataset with high-quality text and audio narrations. *Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2265-2269. IEEE.

Plaza, M. (2017). Cost-effectiveness of audio description production process: comparative analysis of outsourcing and 'in-house' methods. *International Journal of Production Research*, *55*(12), 3480-3496.

Remael, A. (2012). For the use of sound. Film sound analysis for audio-description: some key issues. *MonTi: Monografías de Traducción e Interpretación*, *4*, 255-276.

Remael, A., Reviers, N., & Vercauteren, G. (Eds.) (2015). *Pictures painted in words: ADLAB Audio Description guidelines.* EUT Edizioni Università di Trieste.

Reviers, N. (2018). Tracking multimodal cohesion in audio description: Examples from a Dutch audio description corpus. *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, *14*, 22-35.

Sade, J., Naz, K., & Plaza, M. (2012). Enhancing audio description: A value added approach. In K. Miesenberger, A. Karshmer, P. Penaz, & W. Zagler (Eds.), *Computers Helping People with Special Needs. Lecture Notes in Computer Science,* (Vol. 7382; pp. 270-277). Springer.

Soldan, M., Pardo, A., Alcázar, J. L., Heilbron, F. C., Zhao, C., Giancola, S., & Ghanem, B. (2022). MAD: A scalable dataset for language grounding in videos from movie audio descriptions. *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5016-5025. IEEE.

Sueroj, K., & Lopez, M. (2023). The challenges of existing Thai audio description guidelines. *Proceedings of Advanced Research Seminar on Audio Description (ARSAD 2023).* TransMedia Catalonia Research Group.

Szarkowska, A. (2011). Text-to-speech audio description: Towards wider availability of AD. *The Journal of Specialised Translation*, *15*, 142-162.

Thom, R. (1999). *Designing a movie for sound.* FilmSound.org. Learning space dedicated to the Art and Analyses of Film Sound Design.

Vercauteren, G. (2022). Narratology and/in audio description. In E. Perego, & C. Taylor (Eds.), *The Routledge Handbook of Audio Description*. Routledge.

Vercauteren, G., & Reviers, N. (2022). Audio describing sound – What sounds are described and how? Results from a Flemish case study. *Journal of Audiovisual Translation, 5*(2), 114-133.

Walczak, A., & Fryer, L. (2018). Vocal delivery of audio description by genre: Measuring users' presence. *Perspectives*, *26*(1), 69-83.

Walczak, A., & Iturregui-Gallardo, G. (2022). Artificial voices. In E. Perego and C. Taylor (Eds.), *The Routledge Handbook of Audio Description*. Routledge.

Wang, Y., Liang, W., Huang, H., Zhang, Y., Li, D., & Yu, L.-F. (2021). Toward automatic audio description generation for accessible videos. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Article 277, 1-12. ACM.

Data availability statement

Data pertaining to the film, such as the original image and sound, cannot be shared owing to restrictions placed on the film content; however, data supporting the results of this study, such as the novice audio describers' scripts and questionnaire results, are available on reasonable request to the corresponding author.