Journal of Audiovisual Translation

The Influence of Narrative Specificity and Voice Quality When Listening to Audio Descriptions: A Comparison of the Sighted and the Blind

Viveka Lyberg-Åhlander ✉
Faculty of Arts, Psychology and Theology/Speech Language Pathology, Åbo Akademi University, Turku, Finland

Department of Clinical Sciences Lund, Logopedics, Phoniatrics and Audiology, Lund University, Sweden

Jana Holsanova ✉

Department of Cognitive Science, Lund University, Sweden

Roger Johansson ✉
Department of Psychology, Lund University, Sweden

Abstract

Audio description (AD) serves as a vital means to make visual media accessible to non-sighted and visually impaired audiences. This study systematically investigates the impact of narrative specificity and voice quality on imageability and comprehension in both sighted and non-sighted populations. Twenty non-sighted participants, including congenitally blind individuals and those who lost their sight early in life, were compared with a group of 20 sighted participants, matched for verbal working memory capabilities. Participants listened to 50 short event descriptions, describing spatiotemporal relations with varying levels of narrative specificity, presented in both typical and dysphonic voices. After each event description, participants rated their ability to imagine the content, overall comprehension, listening effort, and listening enjoyment. Results indicate that high narrative specificity enhanced imageability in non-sighted individuals, especially for scenarios involving changes in motion, and, to some extent, for visuospatial relations, irrespective of sightedness. Additionally,

Citation: Lyberg-Åhlander, V., Holsanova, J., & Johansson, R. The Influence of Narrative Specificity and Voice Quality When Listening to Audio Descriptions:: A Comparison of the Sighted and the Blind. *Journal of Audiovisual Translation*, *7*(2), 1–25. https://doi.org/10.47476/jat.v7i2.2024.314

Editor(s): N. Reviers, J. Neves & G. Vercauteren.

Received: February 5, 2024

Accepted: August 23, 2024

Published: December 19, 2024

✉ viveka.lybergahlander@abo.fi https://orcid.org/0000-0002-7178-1604
✉jana.holsanova@lucs.lu.se https://orcid.org/0000-0002-7510-6526
✉ roger.johansson@psy.lu.se https://orcid.org/0000-0003-3434-2538

dysphonic voices increased listening effort and reduced enjoyment for non-sighted participants only. These findings underscore the importance of considering voice quality and narrative specificity in AD for non-sighted users and have implications for both professional audio describers and the development of automated AD systems.

Key words: audio description, voice quality, narrative specificity, spatiotemporal language, mental imagery.

Introduction

Audio description (AD) serves as a vital means to make visual media accessible to non-sighted and visually impaired audiences. The primary goal of AD is to enhance the accessibility of visual information and to offer visually impaired audiences a more comprehensive and detailed understanding, experience, and enjoyment of various forms of audiovisual media. To achieve this objective, the audio describer verbalises pertinent details from the visual content, including characters, objects, and the surrounding environments, as well as relevant narrative interactions among those elements. In doing so, the audio describer bridges the informational gaps, with the intent of eliciting "mental imagery" and enriching the process of meaning-making for the visually impaired audience (Holsanova, 2016; 2022; Vandaele, 2012). The effectiveness of this communication hinges critically upon what information is described, how the information is described, and how the message is aurally expressed (Holsanova, 2016; 2022; Holsanova et al., 2023; Johansson et al., 2023). A fundamental question in the realm of successful AD pertains to how such factors influence its effectiveness.

In the present study, we scrutinise two components within this intricate dynamic. Firstly, we examine the impact of narrative specificity pertaining to spatiotemporal circumstances involving characters and objects within spoken event descriptions on the processes of mental imagery and meaning construction for sighted and non-sighted individuals. Secondly, we delve into an exploration of the role played by the narrator's voice quality in shaping the holistic auditory experience.

Within the domain of audio description (AD), it has been acknowledged that spatiotemporal elements play a pivotal role (Remael & Vercauteren, 2015; Vandaele, 2012; Vercauteren, 2021; Vercauteren & Remael, 2015). However, as far as our knowledge extends, prior investigations have not explored the impact of narrative specificity or voice quality on the perceived effectiveness of AD communication by recipients.

1. Mental Imagery, Verbal Narratives and Sightedness

The capacity for mental visualisation has played a pivotal role in the evolution of human cognitive and communicative abilities, serving as a vital resource for internally simulating experiences in the absence of direct sensory input (e.g., Kosslyn et al., 2006; Pearson, 2019). These internal mental images are regularly conjured in various everyday scenarios, such as when individuals mentally recreate past events, plan for future occurrences, solve problems, or engage with captivating narratives. Substantial neurocognitive research has established that comparable cognitive processes are activated when individuals internally simulate an event and when they directly perceive the same event (e.g., Johansson et al., 2006; Kosslyn et al., 2006; Pearson, 2019). Thus, to grasp a specific situation conveyed within a spoken narration, listeners engage in mental imagery processes corresponding to a "situation model" (Van Dijk & Kintsch, 1983; Zwaan & Radvansky, 1998) of the described state-of-affairs (Bergen et al., 2007; Johansson et al., 2018; Stanfield & Zwaan, 2001; Zwaan

et al., 2002). The capacity to create such situation models has been empirically demonstrated to wield substantial influence over ongoing verbal comprehension (Gambrell & Jawitz, 1993; Garnham, 1981; McKoon & Ratcliff, 1992), enabling receivers of verbal narratives to vicariously immerse themselves in the depicted content (Zwaan, 2004; Zwaan & Madden, 2009). In prominent theories within cognitive linguistics, it is even claimed that the capacity to mentally create such internal models, provides the fundamental building blocks for being able to understand and talk about spatial relations in language (Lakoff & Johnson, 1980).

Research indicates that individuals with blindness can utilise mental imagery similarly to sighted individuals, engaging in cognitive processes like mental scanning or rotating imagined objects and producing spatially coherent drawings, despite never having experienced sight (Amedi et al., 2008; Afonso et al., 2010; Kerr, 1983; Röder & Rösler, 1998). However, the mechanisms underlying these abilities differ, with blind individuals relying more on haptic and motor imagery, which are cognitively more demanding and sequentially structured, affecting the processing of spatiotemporal information (Cattaneo & Vecchi, 2011; Noordzij et al., 2007; Postma et al., 2006).

In navigation, blind individuals use sequential information processing, identifying and connecting positions of landmarks relative to their bodies (e.g., left, right, in front of, behind), contrasting with sighted individuals' use of visual cues to form "mental maps" for orientation, independently of their bodily orientation (Postma et al., 2006). Phenomenologically, blind individuals report experiencing mental imagery in a schematic, non-visual manner, highlighting significant experiential differences from sighted individuals, who often perceive mental images as visually similar to actual percepts (Cattaneo & Vecchi, 2011).

In summary, while blind individuals employ mental imagery to construct situation models of described content, significant differences in cognitive processing and phenomenological experience exist between them and their sighted counterparts. These distinctions underscore the adaptability of the human cognitive system and the impact of sensory experiences on mental imagery. The study of mental imagery across sighted and non-sighted populations leverages a range of methodologies, including behavioural, physiological, and brain imaging techniques, with self-reports remaining a primary method for understanding individual experiences (Kosslyn et al., 2006; Pearson, 2019; Cattaneo & Vecchi, 2011).

1.1. Spatiotemporal Relationships and Narrative Specificity

In the context of AD, it is imperative to recognise the commonalities and distinctions in the processing of spatiotemporal information and mental imagery between sighted and blind individuals. Specifically, when defining and articulating spatiotemporal attributes among characters and objects within an event, it is known that the details of such narrative elements are fundamental for listeners to construct precise mental models of spatiotemporal circumstances (van Dijk & Kintsch, 1983), with direct consequences for narrative comprehension and memory retention (Zwaan & Radvansky,

1998). For instance, Wilken and Kruger (2016) have demonstrated that well-crafted AD, which pays attention to the spatial arrangement and presentation of visual elements in relation to each other, can significantly improve psychological immersion for audiences with visual impairment.

In the present study, we examine how different degrees of narrative specificity regarding spatiotemporal circumstances influence sighted and blind listeners' capacity to mentally visualise the described state of affairs. While narrative specificity broadly refers to the degree of detail, explicitness, and vividness provided in narrative descriptions of characters, settings, and events in storytelling (cf. Ryan, 2004), we focus specifically on the spatiotemporal elements that enable AD receivers to construct precise mental models of these circumstances. In this study, the concept of narrative specificity will thus be used exclusively in this regard.

1.1.1. Describing Spatial Relationships

Spatial relations constitute a fundamental component of human cognition and language, serving as a means for humans to convey information regarding the location, arrangement, and interaction of objects and entities within their environment (Blomberg, 2014; Svorou, 1994; Talmy, 2000). Central to the expression and conceptualisation of spatial relations are spatial prepositions, such as "on," "in," "above," or "beside", which delineate spatial configurations among entities (e.g., Choi et al., 1999; Pederson et al., 1998). These expressions are further specified by the utilisation of various spatial frames of reference, where a spatial relation can be described in relation to the properties of objects themselves (e.g., "The book is on the table"), relative to the viewpoint of an observer (e.g., "The book is to the left of the computer"), or relative to the perspective of a character within a narrative (e.g., "The book is in front of her"). The selection of such frames of reference significantly influences how individuals mentally organise space and conceptualise spatial relationships presented in verbal narratives (Levinson, 2003).

Furthermore, deictic expressions, which anchor these spatial configurations to a speaker's or character's perspective, also play an important role in this process. Terms like "here," "there," "this," and "that" provide essential contextual grounding to situate described entities within a specific spatial framework. This deictic spatial reference frame, often centred on a focal point such as the speaker's location or a key object in the narrative, helps direct attention and dynamically establish spatial relationships as the narrative unfolds. By integrating deixis with focal points, spatial relations, and spatial reference frames, verbal narratives can create a vivid and coherent spatial map, enhancing comprehension and engagement with the described events and environments (e.g., Hanks, 1992).

Thus, how spatial relations, frames of reference, and deixis, are expressed in AD profoundly affects how receivers conceptualise their mental models of described spatial circumstances, such as how a character is directed or positioned in relation to another character.

1.1.2. Describing Changes in Motion

Within cognitive linguistics, the term "path" pertains to the trajectory or route that an object or entity follows as it moves through space (e.g., Talmy, 2000). This concept is integral to understanding how languages express motion events, particularly in languages following the satellite-framed pattern, where path information is primarily encoded within verbs (e.g., "She walked across the street") (cf., Blomberg, 2014). Conversely, the "manner" of motion refers to the specific manner in which an action or movement is executed, providing further specification regarding how such movements occur (e.g., Talmy, 2000).

According to the semantic model of events proposed by Warglien et al. (2012), sentences articulate construal of events, with each event being profiled in distinct ways. For example, if a verb denoting an action or situation signifies the "change vector" (e.g., "move," "walk," "climb," "break"), it conveys the path of motion, whereas if it signifies the "force vector" (e.g., "push," "hit," "run"), it conveys the manner of motion (Gärdenfors, 2014; Warglien et al., 2012).

Consequently, the choice of whether and how to express the manner of motion, in addition to the path of motion, in AD significantly impacts how receivers conceptualise their mental models of the movement of entities within a spatial context.

1.1.3. Narrative Specificity and Audio Description

Creating AD requires a delicate balance between presenting visual information neutrally and objectively and elucidating it in a manner that enhances narrative comprehension. This balance involves selecting strategies for describing narrative events with varying degrees of specificity in description, directly impacting how spatial relations and movements are perceived within the larger narrative framework (Reviers, 2015; Remael et al., 2015; Holsanova, 2022). For instance, the absence of vision can introduce significant ambiguity in the conceptualisation of spatiotemporal configurations. Descriptions that lack directional or positional specificity can hinder understanding, as in the example "a man enters a bus and takes a seat in front of the woman," which does not clarify whether the man is facing the woman or not. Similarly, describing someone as sitting "next to" another person without indicating the specific side (left or right) or position (by the window or the aisle) can leave listeners guessing about the spatial arrangement (Holsanova, 2022).

Moreover, the choice of deictic focal point and spatial reference frame—whether scene-based or character-based—can introduce ambiguity. Scene-based descriptions provide spatial relations from the focal point of an external viewer. In contrast, character-based descriptions align spatial configurations with a character's perspective. The latter typically offers more narrative specificity by reducing ambiguity in AD recipients' mental models of how the described spatial arrangements relate to a protagonist.

In terms of motion, the absence of vision complicates the understanding of the manner of motion (e.g., “walk,” “march,” “run,” “strut”) within verb phrases, where different verbs convey varying degrees of narrative specificity regarding an entity’s movement. The choice of verb can significantly affect the narrative’s meaning (Wargelin et al., 2012), highlighting the importance of carefully selecting terms that offer the right level of detail and clarity for the audience.

Incorporating a detailed approach in AD, focusing on clarity, specificity, and the careful selection of spatial and motion-related linguistic expressions, can significantly enhance the visually impaired audience’s ability to construct accurate mental models of described spatiotemporal narrative circumstances. Thus, by navigating the complexities of spatial and motion information with precision, AD creators can provide a richer, more accessible narrative understanding, fostering greater engagement and enjoyment of media for all listeners.

In the present study, we will systematically investigate how the level of narrative specificity in spatial relations and changes in motion descriptions within spoken event narratives (akin to those in AD) influence the recipients’ conceptualisation and imageability of the described state of affairs.

In our previous research on AD production (Holsanova, 2016, 2022; Holsanova et al., 2023), we observed extensive variability in how audio describers specify and anchor spatial relations relative to spatial reference frames and focal points, as well as in the degree to which they explicate the manner of motion. Therefore, in the spatial domain, we will focus on how spatial relations are described through different reference frames and focal points. In the motion domain, we will concentrate on how the manner of action or movement is described. Note that, from a language processing perspective, we focus on the situation model level (Van Dijk & Kintsch, 1983; Zwaan & Radvansky, 1998). Crucially, this level transcends both the surface structure level (words, phrases, and their syntactic units) and the text base level (propositions and basic ideas conveyed by the verbal material). In contrast, the situation model level involves constructing a mental representation of the described situation, relying on inferences drawn from associated memories and general world knowledge to comprehensively understand the specific state of affairs among goal-relevant story elements (Zwaan & Radvansky, 1998).

1.2. Aural Properties of Spoken Descriptions

While the effective conceptualisation of spatiotemporal circumstances through AD critically depends on how the relevant properties are communicated, the success of this “meeting of the minds” is also contingent on how the described information is conveyed audibly by the audio describer (cf., Walczak, 2017; Walczak & Fryer, 2018). Research has demonstrated that the quality of the speaker’s voice exerts a significant influence on the listening effort expended by listeners, as well as their attitudes and comprehension of the spoken message (e.g., Lyberg-Åhlander et al., 2015, Rogerson & Dodd, 2005; Rudner et al., 2018). Listening effort can be defined as the effort the listener needs to make in terms of allocating cognitive resources to understand the spoken message (for a review, see

McGarrigle, et al., 2014). For example, when listening to noise or specifically when listening to a dysphonic (hoarse) voice, such listening effort increases (Sahlén et al., 2018). This results in the listeners processing the message more slowly and often missing words with content-bearing significance (Lyberg-Åhlander et al., 2015). This phenomenon not only directly affects comprehension but also elevates cognitive load, consequently increasing the effort and motivation required to engage with a spoken narrative.

The precise aspects of a speaker's voice that hinder message reception remain a subject of ongoing investigation. Nonetheless, studies have indicated that it is the overall deviant spectral characteristics of a dysphonic voice that heighten the cognitive workload for listeners. Even a mildly dysphonic voice, exhibiting traits that impact the high-frequency portion of the speech spectrum (such as breathiness and hyperfunction), has been shown to influence listening effort (Rogerson & Dodd, 2005).

As described earlier, blind individuals already contend with substantial cognitive demands when engaging in mental imagery and processing visuospatial information. Therefore, an additional burden stemming from increased listening effort could have profound implications for their capacity to process information effectively and on the perceived quality of the listening experience. Moreover, as a substantial body of evidence indicates that individuals with visual impairments frequently exhibit an enhanced sensitivity in their auditory perception as a compensatory mechanism for their visual deficit (cf., Collignon et al., 2009; Röder & Rösler, 2004; Sabourin et al., 2022), one would anticipate that the narrative experiences of non-sighted individuals would exhibit greater sensitivity to alterations within the auditory domain.

In the assessment of listening effort, a universally accepted benchmark for measurement has not yet been established. Commonly employed approaches encompass self-reports, behavioural observations, physiological metrics, and response time analyses (cf., McGarrigle et al., 2014). In the current investigation, self-reports were employed to evaluate both listening effort and listening enjoyment.

1.4. Present Study

The present study aimed to investigate the influence of narrative specificity and voice quality on imageability and the listening experience in audio descriptions of spatiotemporal relations, targeting both sighted and non-sighted listeners. In light of the pronounced variability observed among individuals with blindness in the realm of mental imagery processing (cf., Cattaneo & Vecchi, 2011), the current study is centred on congenitally blind individuals and those who experienced early-life sight loss.

The study had two primary objectives. Firstly, it examined how descriptions with varying levels of narrative specificity affected the perceived capacity to mentally visualise the depicted spatiotemporal scenarios. Secondly, it explored the impact of voice quality in the verbal narration on the perceived quality of the listening experience.

1.4.1. Hypotheses and Expectations

We hypothesised that descriptions with high narrative specificity would enhance the imageability of the described spatiotemporal relations. We anticipated that this effect would be more pronounced among the non-sighted participants.

We hypothesised that verbal narrations characterised by dysphonic voice quality would result in increased listening effort and a less satisfactory listening experience. We expected these effects to be more prominent in the non-sighted group.

To achieve these objectives, we conducted a study wherein both sighted and non-sighted participants listened to verbal descriptions that focused on spatial relations or motion changes. Following each description, participants provided ratings regarding (1) their ability to mentally visualise the description's content, (2) their overall comprehension of the verbal description, (3) the perceived pleasantness of the listening experience, and (4) the perceived level of effort required to engage with the description. The level of narrative specificity pertaining to the described spatiotemporal situations was manipulated to be either high or low, and the narrator manipulated her voice quality to be of either typical or mimicked dysphonic quality.

2. Method

2.1. Participants

In total, a cohort of 57 adults (38 female) with a mean age of 33.2 years (*SD* = 14.9) participated in the study. Among these participants, 27 individuals met the criteria for non-sighted status, defined in accordance with the World Health Organization's standards for blindness or severe visual impairment. The remaining 30 participants were sighted individuals with no reported visual impairments, recruited from personal networks and from the student body at Lund University.

The non-sighted participants were recruited through various organisations and networks dedicated to serving the visually impaired community in Sweden, including the Swedish Association of the Visually Impaired (SRF), the Swedish Braille Authority (Punktskriftsnämnden), the Swedish Agency for Accessible Media (MTM), and Young People with Visual Impairment (US). To maintain the focus of this study on individuals who had experienced early-life blindness, seven non-sighted participants were excluded. The exclusions consisted of one participant with mild visual impairment, five participants who had become visually impaired or blind during adulthood, and one participant due to technical complications. Among the remaining 20 non-sighted participants, 13 were congenitally blind, and the remaining seven had lost their sight or experienced severe visual impairment prior to the age of seven, with a minimum duration of 27 years of living with the condition.

Within the sighted group, three participants were excluded: two were excluded due to technical issues, and one due to impaired hearing. To mitigate the potential confounds stemming from cognitive disparities for processing verbal material among the two groups, we then selected 20 sighted participants (matching the sample size of the non-sighted group) based on the best-fit in terms of their performance in a verbal working memory task: the Competing Language Processing Task (CLPT) developed by Gaulin and Campbell (1994). As a result, the analyses and findings of this study are based on a total of 40 participants (24 female), with a mean age of 33.7 years (*SD* = 15.4). This participant pool consisted of 20 non-sighted individuals (10 female) with a mean age of 43.9 years (*SD* = 15.2) and 20 sighted individuals (6 female) with a mean age of 23.5 years (*SD* = 6.1). Thus, while the sighted group consisted of younger individuals, both groups were comparable in terms of critical cognitive capabilities, as measured by the verbal working memory task.

All participants provided informed consent, either in written or oral form, and received compensation in the form of vouchers. Ethical approval was obtained in advance from the Swedish Ethical Review Authority under registration number 2019-03445. All research methods adhered to the ethical standards outlined in the Swedish Act concerning the Ethical Review of Research Involving Humans and the Code of Ethics of the World Medical Association, as articulated in the Declaration of Helsinki.

2.2. Material

2.2.1. Auditory Stimuli

The auditory stimuli were presented with PsychoPy (Peirce et al., 2019) and were shared over Zoom. Participants listened to the material on headphones in their own personal setup. The stimuli were recorded in a floating room in a recording studio (LARMstudio[1]) with a sound-damping carpet and backdrops of sound-damping textiles on 75% of the walls. A sound technician monitored the recordings from a separate control room. The speaker, a voice healthy 55-year-old female with a neutral accent, was sitting in front of a microphone (Sennheiser MK4). The sentences were recorded first with a typical voice and thereafter with a simulated dysphonic voice.

In total, the material comprised 50 Swedish verbal event descriptions recorded in two versions (100 descriptions in total), describing spatiotemporal relations, and four verbal statements corresponding to the evaluation phase (the latter narrated with a typical male voice by a voice healthy 45-year-old male speaker). In one version, the event descriptions were narrated with a typical female voice, and in the other version, with a simulated dysphonic (hoarse) female voice. Ten of the recordings with the simulated dysphonic voice were assessed by three independent Speech Language Pathologists (SLP) specialised in voice. The dysphonic voice was assessed with the Stockholm Voice Evaluation

[1] https://www.humlab.lu.se/sv/utrustning/larm-studion/

Approach (SVEA, Hammarberg, 2000) to be mildly hyperfunctional and with moderate vocal fry. The vocal fundamental frequency (F0) and pace of speech were kept constant in both voice versions.

2.2.2. Event Descriptions

The event descriptions encompassed two distinct categories: (1) descriptions emphasising the spatial relationship between two central entities, and (2) descriptions concentrating on changes in motion, encompassing either a single central entity or two entities. All event descriptions adhered to a consistent narrative structure, commencing with the introduction of a contextual backdrop, followed by the delineation of a spatiotemporal state of affairs within that context, characterised by either high or low levels of narrative specificity for one or two entities.

For event descriptions focusing on spatial relations, there were two types: (i) spatial relations between an individual and an object (person-object) and (ii) spatial relations between two individuals (person-person) (see Fig. 1).

For person-object descriptions, high narrative specificity was characterised by descriptions wherein the spatial configuration between an object and an individual was explicated from the individual's egocentric reference frame, employing the prepositions "in front of" and "behind." Consequently, the event description adopted the focal point of the characters within the event, elucidating whether an object was in their field of view (in front of) or not (behind). In contrast, low narrative specificity was denoted by descriptions wherein the spatial configuration between an object and an individual was articulated from an allocentric reference frame, employing prepositions such as "to the right of" and "to the left of." Here, the event description embraced a scene-based perspective through an external viewer's focal point, wherein the spatial configuration between an object and an individual was delineated in relation to the scene's reference frame, without considering the character's viewpoint or orientation. Thus, in contrast to the high narrative specificity version, it is not possible to determine whether the object is in the character's field of view or not. In Table 1, this is illustrated by the spatial relationship between Anna and a backpack. In the version with low narrative specificity ("Anna stands to the right of the backpack"), the description uses a perspective-independent allocentric reference frame, which does not reference Anna's viewpoint and lacks details about orientation or potential interaction between them. In the version with high narrative specificity ("the backpack is behind Anna"), the description employs a perspective-based egocentric reference frame from Anna's viewpoint, with the backpack positioned relative to her orientation (behind), suggesting that she is unaware of it.

For person-person descriptions, high narrative specificity was characterised by the depiction of the spatial configuration between the two individuals, taking into account both individuals' focal points and employing prepositions such as "opposite." In such instances, the event description encompassed pertinent relational attributes of both characters, including their relative orientations, making it possible to, for instance, determine whether they are facing each other or not. In contrast,

low narrative specificity was characterised by descriptions wherein the spatial configuration between the two individuals solely considered the perspective of one character, employing prepositions such as “beside” and “in front of.” Consequently, in low narrative specificity descriptions, pertinent relational aspects between the characters, such as their relative orientations to each other, remained unspecified, making it impossible to, for instance, determine whether they were facing each other or not. Note that in contrast to the person-object descriptions, the prepositions “behind” and “in front of”, will here give rise to lower narrative specificity, highlighting that it is not the linguistic content of the prepositions themselves that are offering the specification, but the whole situation model as construed by the spatial relations, the reference frame, and the focal point in relation to the narrative context (cf. Zwaan & Radvansky, 1998; Blomberg, 2014). In Table 1, this is illustrated by the spatial relationship between Lisa and Maja. In the version with low narrative specificity, the description is from Lisa’s viewpoint, specifying their relative positions without indicating whether they are facing each other or back-to-back. In the version with high narrative specificity, the description includes both Lisa’s and Maja’s viewpoints, specifying that they are facing each other, suggesting direct interaction between them.

Concerning event descriptions emphasising motion changes, there were three types: (i) motion changes involving a person (person), (ii) motion changes pertaining to an object subjected to the actions of a person (person-object), or (iii) motion changes involving a person engaging with another person (person-person) (Fig. 1). In all three subtypes, low narrative specificity entailed event descriptions wherein the path of motion was expounded through relatively neutral verb phrases. Conversely, high narrative specificity was characterised by event descriptions that offered specifications regarding the manner of motion, encompassing factors such as variations in force vectors associated with the manner of motion (e.g., increased/decreased force vectors). See Figure 1 for categorical schematics and Table 1 for specific examples of event descriptions with high and low narrative specificity, expressed through the specification of the manner of motion.

Note that the alteration of narrative specificity for spatial relations and motion changes is not linguistically equivalent at the word or propositional level. However, the present study does not aim for such equivalence; instead, it focuses on specificities at the level of the situation model. This level transcends individual words and linguistic units, emphasising the construction of mental models of specific spatiotemporal states of affairs based on inferences drawn from associated memories and general world knowledge (Zwaan & Radvansky, 1998).

Figure 1

Schematics of the Two Categories of Event Descriptions – Spatial Relations and Motion Changes – and the Different Types of Event Descriptions Within Those Categories

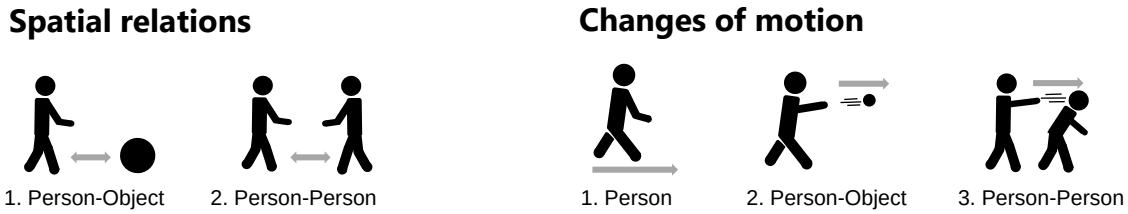


Table 1

Examples of High and Low Narrative Specificity Over the Different Event Descriptions

	High Narrative Specificity	Low Narrative Specificity
Spatial relations		
Person-Object	At the train station. Anna is on a platform. On the platform lies a black backpack. The backpack is **behind** Anna.	At the train station. Anna is on a platform. On the platform lies a black backpack. Anna stands **to the right of** the backpack.
Person-Person	On the train. Lisa is in a train compartment. Lisa's sister Maja is also there. Lisa sits **opposite** Maja.	On the train. Lisa is in a train compartment. Lisa's sister Maja is also there. Lisa sits **in front of** Maja.
Motion changes		
Person	In School. It's Monday morning. Frank **rushes through** the classroom door.	In School. It's Monday morning. Frank **enters** the classroom door.
Person-Object	In the grocery store. It's Friday night. Fredrik **throws** his goods onto the checkout conveyor belt.	In the grocery store. It's Friday night. Fredrik **puts** his goods onto the checkout conveyor belt.
Person-Person	In the park. It's Wednesday morning. Annika is with her baby daughter on the playground. Annika **pushes** her into a baby swing.	In the park. It's Wednesday morning. Annika is with her baby daughter on the playground. Annika **puts** her into a baby swing.

2.3. Design and Procedure

Due to the data collection period in 2020, which coincided with the COVID-19 pandemic, the study was administered remotely through the Zoom platform using online video calls. The experiment was

executed on the experimenter's computer and shared with participants via the audio-sharing feature. To minimise auditory disturbances during the primary experiment, the experimenter's camera remained deactivated, and the microphone was muted. Prior to the experiment's commencement, participants were apprised of the procedures and provided their informed consent verbally, which was audio-recorded and securely stored.

The experimental protocol consisted of three distinct phases. Firstly, the CLPT working memory task was administered. Subsequently, a practice session, emulating the overarching procedure of the main experiment, was conducted to acquaint participants with their tasks. Finally, the main experiment was executed.

During the main experiment, participants were presented with brief event descriptions, delineating either spatial relations or changes in motion, characterised by varying degrees of narrative specificity. Participants were unaware of the manipulation of narrative specificity and remained uninformed regarding the different categories and types of event descriptions. In total, each participant encountered 50 event descriptions, encompassing 20 event descriptions centring on spatial relations (10 Person-Object, 10 Person-Person) and 30 event descriptions focusing on changes in motion (10 Person, 10 Person-Object, 10 Person-Person).

Half of the event descriptions exhibited high narrative specificity, while the remaining half featured low narrative specificity, with an equal distribution of high and low narrative specificity across the distinct types and categories of event descriptions. The presentation order of event descriptions and narrative specificity was randomised within each experimental session.

Furthermore, half of the descriptions were presented using a simulated dysphonic voice, while the other half utilised a typical voice. The allocation of dysphonic and typical voices was evenly spread across event description categories and types, as well as narrative specificity levels, with a randomised sequencing in each experimental session.

Subsequent to each event description, participants provided ratings on a 1–6 scale, evaluating (1) their ability to mentally visualise the description's content, (2) their overall comprehension of the verbal description, (3) the perceived pleasantness of the listening experience, and (4) the perceived level of effort required to engage with the description. Participants were prompted to vocally respond to these four aspects following an auditory voice cue, which included the terms "imageability," "comprehension," "enjoyment," and "effort." Participants were instructed to respond promptly and accurately, and their responses were recorded by the experimenter pressing the corresponding key corresponding to the participant's answer.

2.4. Analytical Approach

We analysed the data in three steps. Firstly, across the sighted and non-sighted groups, we examined how narrative specificity influenced the imageability of spatial relations and motion changes.

Secondly, across the sighted and non-sighted groups, we examined how voice quality influenced the listening experience. Finally, we explored potential interactions between voice quality and narrative specificity within this context.

Statistical analyses were carried out utilising Generalised Linear Mixed-Effect Models (Gallucci, 2019) with the aid of jamovi version 1.6.23 (The jamovi project, 2019). These models, incorporating data from all data points, offer enhanced statistical power compared to conventional analyses of variance. Participants and Scenarios were included as random effects (intercepts). The model fit assessment involved contrasting the deviance of the proposed model with an unconditional null model, encompassing solely the intercept and random factors. A backwards selection approach was employed in constructing models with multiple independent variables, commencing with a maximal model comprising all variables and interactions. Likelihood-ratio tests were then used to compare models, progressively eliminating non-significant effects until no further model changes yielded a significant likelihood-ratio test ($p < .05$). Models were fitted employing restricted maximum likelihood (REML). Satterthwaite approximations were employed to evaluate the significance of individual predictors.

Responses with unreasonable short (< 0.2 s) or long (> 20 s) response times were excluded from the analyses (0.15%).

3. Results

3.1. Narrative Specificity and Imageability

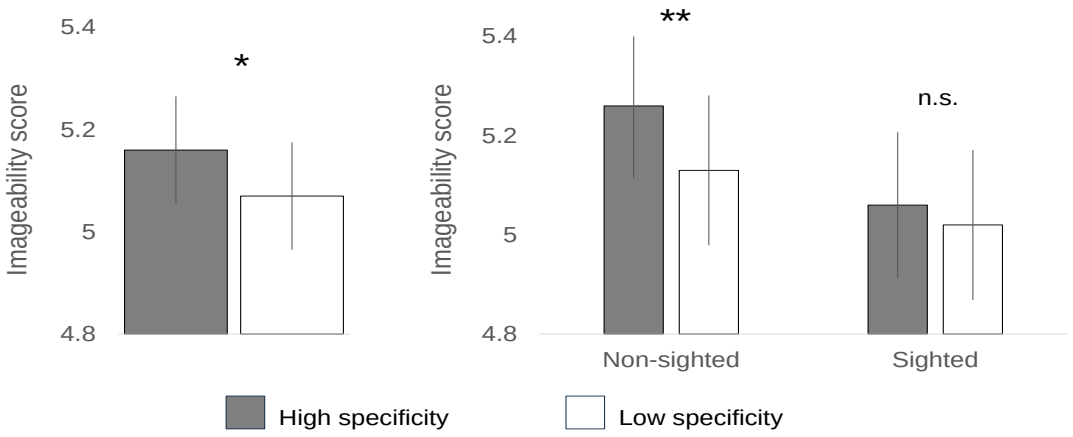
To examine imageability, we first created a null model (AIC = 4727, R^2 = 0.43), including only the dependent variable (imageability rating) and the intercept (participants and scenarios modelled as random effects). To test the effect of narrative specificity, the null model was then contrasted against models including fixed effects of narrative specificity (high, low), sightedness (sighted, non-sighted), and description category (spatial relations, motion changes). The best model fit, including all fixed effects and their interactions (AIC = 4710, R^2 = 0.44), revealed significant main effects of narrative specificity (F = 6.7, β = 0.09, SE = 0.04, t = 2.58, p = .01), description category (F = 5.7, β = 0.11, SE = 0.04, t = 2.38, p = .03), a two-way interaction effect between description category and sightedness (F = 10.9, β = 0.23, SE = 0.07, t = 3.30, $p < .001$), and a three-way interaction effect between narrative specificity, description category, and sightedness (F = 4.8, β = 0.30, SE = 0.14, t = 2.19, p = .03).

In alignment with our hypotheses, imageability ratings were higher for high narrative specificity (see Fig. 2 – left). Post-hoc analyses indicated that this effect was primarily driven by the non-sighted group (see Fig. 2 – right). Furthermore, it was observed that, in the sighted group, imageability ratings were overall higher for descriptions concerning motion changes than spatial relations.

To further disentangle the three-way interaction effect between narrative specificity, sightedness, and description category, separate analyses were conducted for descriptions concerning spatial relations and motion changes.

Figure 2

Average Imageability Ratings for High and Low Specificity Across All Event Descriptions Over All Participants (left) and When Grouped According to Sightedness (right)



Note: Error bars denote the standard errors of the mean. n.s., nonsignificant. *p <.05, **p<.01.

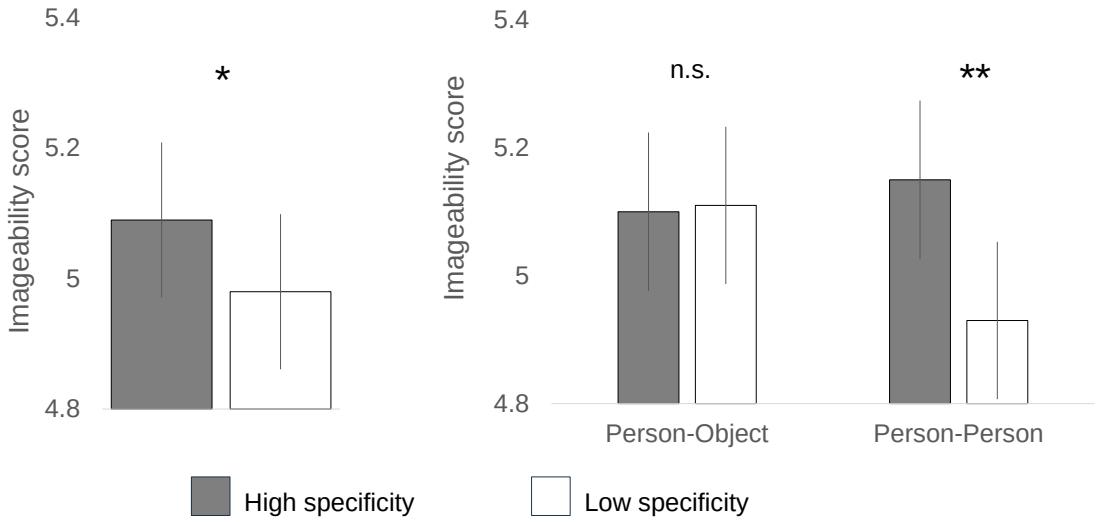
3.1.1. Spatial Relations

To examine the imageability of visuospatial relation, we first created a null model (AIC = 1923, R^2 = 0.47), including only the dependent variable (imageability rating) and the intercept (participants and scenarios modelled as random effects). To test the effect of narrative specificity, the null model was then contrasted against models including fixed effects of narrative specificity (high, low), sightedness (sighted, non-sighted), and spatial relation type (person, person-object, person-person). The best model fit, including fixed effects of narrative specificity and spatial relation type, and their interaction (AIC = 1919, R^2 = 0.48), revealed a significant main effect of narrative specificity (F = 4.4, β = 0.11, SE = 0.05, t = 2.10, p = .04), and a two-way interaction effect between narrative specificity and spatial relation type (F = 5.2, β = 0.24, SE = 0.10, t = 2.3, p = .02).

Consistent with our predictions, imageability was rated higher for high narrative specificity (see Fig. 3 - left). However, post-hoc analyses indicated that this effect was only significant for spatial relations involving person-person interactions (p = .002; see Fig. 3 – right). Contrary to expectations, no effect of sightedness was observed.

Figure 3

Average Imageability Ratings for High and Low Specificity Across All Event Descriptions Focusing on Spatial Relations (left) and When Separating the Descriptions into the Two Description Types: Person-Object, Person-Person (right)



Note. Error bars denote the standard errors of the mean. n.s., nonsignificant. *p <.05, **p <.01.

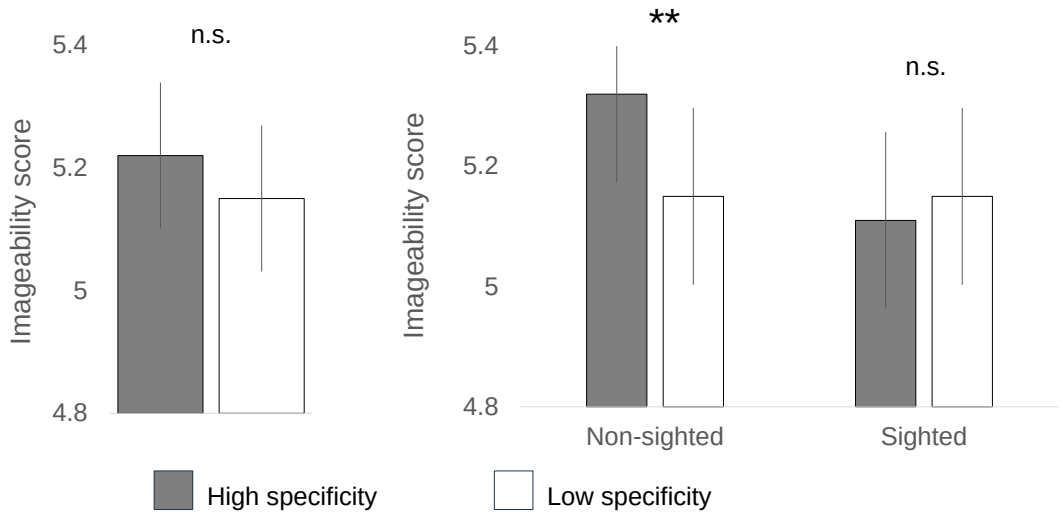
3.1.2. Motion Changes

To examine the imageability of motion changes, we first created a null model (AIC = 2851, R^2 = 0.40), including only the dependent variable (imageability rating) and the intercept (participants and scenarios modelled as random effects). To test the effect of narrative specificity, the null model was then contrasted against a model including fixed effects of narrative specificity (high, low), sightedness (sighted, non-sighted), and type of motion change (person, person-object, person-person). The best model fit, including fixed effects of narrative specificity and sightedness, and their interaction (AIC = 2848, R^2 = 0.42), revealed a significant two-way interaction between narrative specificity and sightedness (F = 5.8, β = 0.21, SE = 0.09, t = 2.4, p = .02).

Post-hoc analyses revealed that imageability was rated higher for high narrative specificity exclusively within the non-sighted group (p = .003; see Fig. 4). Hence, in accordance with our predictions and in contrast to the sighted group, imageability received higher ratings for high narrative specificity in the non-sighted group, regardless of the type of motion change.

Figure 4

Average Imageability Ratings for High and Low Specificity Across All Event Descriptions Focusing on Motion Changes Over All Participants (left) and When Grouped According to Sightedness (right)



Note. Error bars denote the standard errors of the mean. n.s., nonsignificant. ***p*<.01.

3.2. Narrative Specificity and Overall Comprehension

When we conducted similar analyses for overall comprehension, no significant main effects for narrative specificity or significant interactions between sightedness and narrative specificity emerged (*p-values* > .11). Consequently, the reported effects of narrative specificity appear to be specific to the ability to conjure mental images of the described scenarios and did not exert an influence on the overall comprehension of the verbal descriptions.

3.3. Voice Quality and Listening Experience

To investigate the influence of voice quality on the listening experience, separate models were developed to examine listening effort and listening enjoyment.

3.3.1. Listening Effort

To examine the influence of voice quality on listening effort, we first created a null model (AIC = 4779, R^2 = 0.53), including only the dependent variable (listening effort) and the intercept (participants and scenarios modelled as random effects). The null model was then contrasted against models including

fixed effects of voice quality (normal, dysphonic) and sightedness (sighted, non-sighted). The best model fit included both fixed effects and their interaction (AIC = 4761, R^2 = 0.54), and revealed a significant main effect of voice quality (F = 4.8, β = 0.08, SE = 0.03, t = 2.20, p = .03), and an interaction effect between voice quality and sightedness (F = 18.7, β = 0.30, SE = 0.07, t = 4.3, p < .001).

Post-hoc analyses revealed that listening effort was rated higher for descriptions narrated with the dysphonic voice only in the non-sighted group (see Fig. 5 – left). Thus, in accordance with predictions and in contrast to the sighted group, descriptions narrated with dysphonic voice quality elicited higher ratings for listening effort in the non-sighted group.

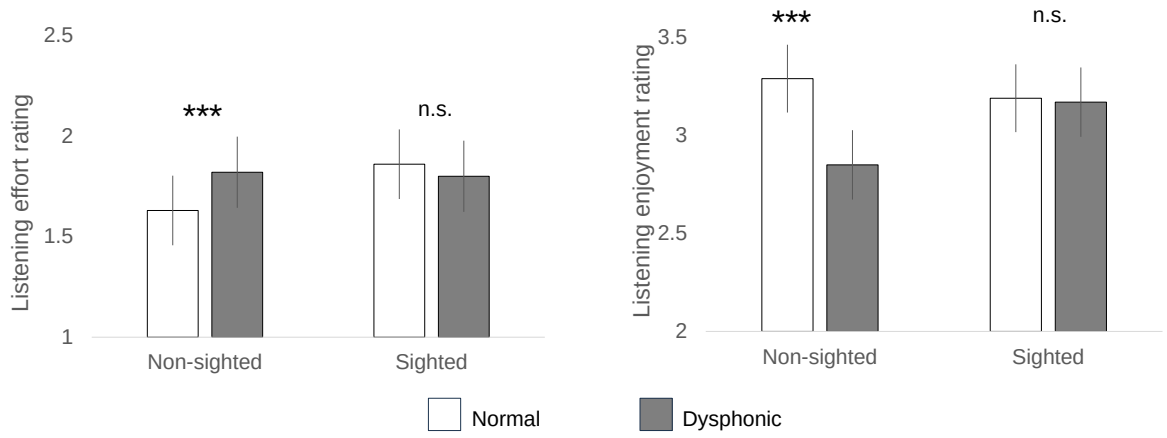
3.3.2. Listening Enjoyment

To examine the influence of voice quality on listening enjoyment, we first created a null model (AIC = 5276, R^2 = 0.55), including only the dependent variable (listening enjoyment) and the intercept (participants and scenarios modelled as random effects). The null model was then contrasted against models including fixed effects of voice quality (normal, dysphonic) and sightedness (sighted, non-sighted). The best model fit included both fixed effects and their interaction (AIC = 4761, R^2 = 0.54). It revealed a significant main effect of voice quality (F = 30.3, β = -0.21, SE = 0.04, t = 5.50, p < .001), and an interaction effect between voice quality and sightedness (F = 20.1, β = - 0.35, SE = 0.08, t = 4.5, p < .001).

Post-hoc analyses revealed that listening enjoyment was rated lower for descriptions narrated with the dysphonic voice only within the non-sighted group (see Fig. 5 – right). Thus, in accordance with predictions and in contrast to the sighted group, descriptions narrated with dysphonic voice quality elicited lower ratings for listening enjoyment in the non-sighted group.

Figure 5

Average Listening Effort (left) and Listening Enjoyment (right) Ratings for Sighted and Non-Sighted Participants When Grouped According to Voice Quality (normal, dysphonic)



Note. Error bars denote the standard errors of the mean. n.s., nonsignificant. *** p <.01.

3.4. Interactions Between Voice Quality and Narrative Specificity

Lastly, in order to investigate potential interactions between voice quality and narrative specificity, we introduced both voice quality and narrative specificity variables, along with their interaction, into all previously established models encompassing imageability, overall comprehension, listening effort, and listening enjoyment. Nonetheless, no statistically significant interactions between voice quality and narrative specificity were observed at any level (*p-values* > .12).

4. Discussion

The results of the present study provide valuable insights into the influence of narrative specificity and voice quality on the imageability and listening experience in audio descriptions of spatiotemporal relations, considering both sighted and non-sighted participants.

The study found a significant main effect of narrative specificity on imageability. High narrative specificity led to higher imageability ratings, aligning with the idea that detailed descriptions enhance the capacity to mentally visualise the described scenarios (Holsanova, 2016; 2022). Importantly, this effect was more pronounced among non-sighted participants, suggesting that detailed audio descriptions are particularly beneficial for individuals without sight. These findings highlight the importance of providing rich and detailed descriptions in audio descriptions to improve the possibility of properly imagining visual content for visually impaired audiences (Holsanova, 2016; 2022).

When examining spatial relations specifically, the effect of narrative specificity was significant, with higher imageability ratings for high narrative specificity. This effect was, however, only evident for spatial relations involving person-person scenarios. This finding aligns with previous research (Holsanova 2016, 2022; Wilken & Kruger, 2016) and suggests that detailed descriptions are particularly beneficial when conveying interpersonal interactions within spatial contexts. However, contrary to expectations, sightedness had no significant effect, indicating that both sighted and non-sighted participants benefited from high narrative specificity.

When considering the absence of a narrative specificity effect for person-object descriptions, it is essential to acknowledge that the manipulation of narrative specificity predominantly hinged on variances in spatial reference frames and focal points. Low narrative specificity was characterised by a perspective-independent allocentric reference frame without details about orientation or potential interaction between the character and the object, while high narrative specificity entailed the utilisation of a perspective-based egocentric reference frame, where the object is described relative to the viewpoint and orientation of the described character. The distinction in reference frames and deictic focal points may not have yielded a substantial impact on the degree of narrative specificity within the limited scenarios employed in this study, such as determining whether a backpack was situated within the described character's field of view or only "somewhere" within the same spatial context as the character. The significance of the bag concerning the character remains rather

unspecified in both scenarios, with only an implied relevance to the narrative. If the meaning were to be further elucidated, such as specifying the presence of a bomb in the bag, the narrative specificity of the spatial configuration in relation to the character's field of view would likely have greater implications for the imagined situation model. Future investigations should explore the influence of narrative meaning on spatial relationships in more depth.

A significant interaction between narrative specificity and sightedness was observed for motion changes. High narrative specificity led to higher imageability ratings, but this effect was only significant for the non-sighted group. Sighted participants, on the other hand, did not show a significant improvement in imageability with high narrative specificity. This suggests that detailed descriptions are more critical for enhancing imageability in non-sighted individuals when it comes to motion changes.

It is worth noting that the effect of narrative specificity was specific to imageability and did not significantly influence overall comprehension of the verbal description. This implies that the improved ability to imagine spatiotemporal properties within a narrative is not directly related to a general understanding of the verbal information itself (cf., Holsanova, 2022), highlighting the importance of considering how the specific situation model – which goes beyond words, grammar, and propositions (e.g., Bergen, et al., 2007; Zwaan & Radvansky, 1998; Johansson et al., 2018) – is communicated in AD (Holsanova, 2016; 2022). However, since the present study does not include an objective measure of comprehension of the described spatiotemporal state of affairs, we cannot make any claims about the accuracy of participants' mental models in this regard. This limitation warrants further studies specifically designed to address these aspects.

Our study also investigated the impact of voice quality on the listening experience. We found that dysphonic (hoarse) voice quality increased listening effort and decreased listening enjoyment. Crucially, this effect was significant for non-sighted participants but not for sighted participants. These findings indicate that voice quality plays a central role in the listening experience, especially for visually impaired individuals who rely heavily on auditory information. Previous research has shown the adverse impacts of a dysphonic voice on both listening effort and listening enjoyment (Lyberg-Åhlander et al., 2015; Rogerson & Dodd, 2005), along with findings that an irregular voice (dysphonic) places increased demands on the listener's working memory capacity (Imhof et al., 2014), it is conceivable that more cognitive resources were needed for the non-sighted listeners when listening to the dysphonic voice. Given that individuals without sight already contend with substantial cognitive demands when engaging in mental imagery and processing visuospatial information (Cattaneo & Vecchi, 2011), the additional burden of increased listening effort is likely to have profound implications for their capacity to process spatiotemporal information effectively. The dysphonic voice was simulated by the speaker. Although the expert panel judged the voice as dysphonic, it may have been perceived by the listeners as unnatural and could, as such, have influenced the results. However, since Rogerson and Dodd (2005) showed that even a mildly dysphonic voice affects the listener's perception, this risk can be seen as minor and cannot explain the difference between the sighted and non-sighted group.

Surprisingly, no significant interactions were found between voice quality and narrative specificity in any of the measured dimensions, suggesting that the influence of voice quality on the listening experience and imageability is relatively consistent regardless of the level of narrative specificity. Further research may be needed to explore potential nuanced interactions between these factors.

Collectively, these findings have practical implications for content creators and audio describers seeking to improve the accessibility and user experience of visual media for individuals with visual impairments. Detailed audio descriptions are crucial, particularly for conveying spatial relations and motion changes, as they significantly enhance the mental visualisation of content for both sighted and non-sighted individuals. Additionally, maintaining good voice quality in audio descriptions is essential to minimise listening effort and maximise listening enjoyment, especially for the non-sighted audience.

In conclusion, this study provides valuable insights into how narrative specificity and voice quality influence the imageability and listening experience of audio descriptions for spatiotemporal relations, shedding light on the complex interplay between these factors and their distinct effects on sighted and non-sighted individuals. These findings underscore the importance of tailoring audio descriptions to meet the diverse needs of audiences with varying visual abilities.

References

Afonso, A., Blum, A., Katz, B. F., Tarroux, P., Borst, G., & Denis, M. (2010). Structural properties of spatial representations in blind people: Scanning images constructed from haptic exploration or from locomotion in a 3-D audio virtual environment. *Memory & Cognition*, *38*(5), 591–604.

Amedi, A., Merabet, L. B., Camprodon, J., Bermpohl, F., Fox, S., Ronen, I., Kim, D. S., & Pascual-Leone, A. (2008). Neural and behavioral correlates of drawing in an early blind painter: A case study. *Brain Research*, 1242, 252–262.

Bergen, B. K., Lindsay, S., Matlock, T., & Narayanan, S. (2007). Spatial and linguistic aspects of visual imagery in sentence comprehension. *Cognitive Science, 31*(5), 733–764.

Blomberg, J. (2014). *Motion in language and experience: Actual and non-actual motion in Swedish, French and Thai* [Doctoral dissertation, Lund University]. https://portal.research.lu.se/files/3888051/4433942.pdf

Cattaneo, Z., & Vecchi, T. (2011). *Blind vision: The neuroscience of visual impairment*. MIT Press.

Choi, S., McDonough, L., Bowerman, M., & Mandler, J. M. (1999). Early sensitivity to language-specific spatial categories in English and Korean. *Cognitive Development, 14*(2), 241–268.

Collignon, O., Voss, P., Lassonde, M., & Lepore, F. (2009). Cross-modal plasticity for the spatial processing of sounds in visually deprived subjects. *Experimental Brain Research, 192*(3), 343–358.

Gallucci, M. (2019). GAMLj: General analyses for linear models. [jamovi module]. https://gamlj.github.io/

Gambrell, L. B., & Jawitz, P. B. (1993). Mental imagery, text illustrations, and children's story comprehension and recall. *Reading Research Quarterly, 28*(3), 264–276.

Gärdenfors, P. (2014). *The geometry of meaning: Semantics based on conceptual spaces*. MIT press.

Garnham, A. (1981). Mental models as representations of text. *Memory & Cognition, 9*(6), 560–565.

Gaulin, C. A., & Campbell, T. F. (1994). Procedure for assessing verbal working memory in normal school-age children: Some preliminary data. *Perceptual and Motor Skills, 79*(1), 55–64.

Hammarberg, B. (2000). Voice research and clinical needs. *Folia phoniatrica et logopaedica*, *52*(1-3), 93-102.

Hanks, W. F. (1992). The indexical ground of deictic reference. In A. Duranti & C. Goodwin (Eds.), *Rethinking context: Language as an interactive phenomenon* (pp. 43–76). Cambridge University Press.

Holsanova, J. (2016). Cognitive approach to audio description. In: A. Matamala & P. Orero (Eds.), *Researching audio description: New approaches.* Palgrave Macmillan.

Holsanova, J. (2022). The cognitive perspective on audio description: Production and reception processes. In C. Taylor & E. Perego (Eds.), *The Routledge handbook of audio description* (pp. 57–77). Taylor & Francis.

Holsanova, J., Blomberg, J., Blomberg, F., Gärdenfors, P., & Johansson, R. (2023). Event segmentation in the audio description of films: A case study. *Journal of Audiovisual Translation, 6*(1), 649–692.

Imhof, M., Välikoski, T. R., Laukkanen, A. M., & Orlob, K. (2014). Cognition and interpersonal communication: The effect of voice quality on information processing and person perception. *Studies in Communication Sciences, 14*(1), 37–44.

Johansson, R., Holsanova, J., & Holmqvist, K. (2006). Pictures and spoken descriptions elicit similar eye movements during mental imagery, both in light and in complete darkness. *Cognitive Science*, 30, 1053–1079.

Johansson, R., Oren, F., & Holmqvist, K. (2018). Gaze patterns reveal how situation models and text representations contribute to episodic text memory. *Cognition,* 175, 53–68.

Johansson, R., Rastegar, T., Lyberg-Åhlander, & Holsanova, J. (2023). Event boundary perception among the visually impaired in audio described films. *PsyArXiv*. https://osf.io/preprints/psyarxiv/7bmdx

Kerr, N. H. (1983). The role of vision in "visual imagery" experiments: Evidence from the congenitally blind. *Journal of Experimental Psychology: General*, 112, 265–277.

Kosslyn, S. M., Thompson, W. L., & Ganis, G. (2006). *The case for mental imagery*. Oxford University Press.

Lakoff, G., & Johnson, M. (1980). The metaphorical structure of the human conceptual system. *Cognitive Science, 4*(2), 195–208.

Levinson, S. C. (2003). *Space in language and cognition: Explorations in cognitive diversity* (Vol. 5). Cambridge University Press.

Lyberg-Åhlander, V., Brännström, K. J., & Sahlén, B. S. (2015). On the interaction of speakers' voice quality, ambient noise and task complexity with children's listening comprehension and cognition. *Frontiers in Psychology*, 6, 871.

McGarrigle, R., Munro, K. J., Dawes, P., Stewart, A. J., Moore, D. R., Barry, J. G., & Amitay, S. (2014). Listening effort and fatigue: What exactly are we measuring? A British Society of Audiology Cognition in Hearing Special Interest Group "white paper". *International Journal of Audiology, 53*(7), 433–445.

McKoon, G., & Ratcliff, R. (1992). Inference during reading. *Psychological Review, 99*(3), 440.

Noordzij, M. L., Zuidhoek, S., & Postma, A. (2007). The influence of visual experience on visual and spatial imagery. *Perception, 36,* 101–112.

Pearson, J. (2019). The human imagination: The cognitive neuroscience of visual mental imagery. *Nature Reviews Neuroscience, 20*(10), 624–634.

Pederson, E., Danziger, E., Wilkins, D., Levinson, S., Kita, S., & Senft, G. (1998). Semantic typology and spatial conceptualization. *Language, 74*(3), 557–589.

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E. J., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, 51, 195–203.

Postma, A., Zuidhoek, S., Kappers, A. M., & Noordzij, M. L. (2006). Haptic spatial orientation processing and working memory. *Cognitive Processing, 7*(1), 181–181.

Remael, A., Reviers, N., & Vercauteren, G. (Eds.). (2015). *Pictures painted in words: ADLAB audio description guidelines*. EUT Edizioni Università DiTrieste.

Remael, A., & Vercauteren, G. (2015). Spatio-temporal settings. In: *Pictures painted in words: ADLAB audio description guidelines*. EUT Edizioni Università DiTrieste, Section 2.1.2, 24–27.

Reviers, N. (2015). The language of audio description in Dutch: Results of a corpus study. In A. Jankowska & A. Szarkowska (Eds.), *New points of view on audiovisual translation and accessibility* (pp. 167–189). Peter Lang.

Rogerson, J., & Dodd, B. (2005). Is there an effect of dysphonic teachers' voices on children's processing of spoken language? *Journal of Voice, 19*(1), 47–60.

Rudner, M., Lyberg-Åhlander, V., Brannstrom, J., Nirme, J., Pichora-Fuller, M., & Sahlen, B. (2018). Listening comprehension and Listening Effort in the primary School classroom. *Frontiers in Psychology*, 9, 1193

Ryan, M.-L. (2004). *Narrative across media: The languages of storytelling*. University of Nebraska Press.

Röder, B., & Rösler, F. (1998). Visual input does not facilitate the scanning of spatial images. *Journal of Mental Imagery*, 22, 165–181.

Röder, B., & Rösler, F. (2004). Compensatory plasticity as a consequence of sensory loss. In G. Calvert, C. Spence, & B. E. Spence (Eds.), *The handbook of multisensory processes* (pp. 719–747), MIT Press.

Sabourin, C. J., Merrikhi, Y., & Lomber, S. G. (2022). Do blind people hear better? *Trends in Cognitive Sciences, 26*(11), 999–1012.

Sahlén, B., Haake, M., von Lochow, H., Holm, L., Kastberg, T., Brännström, K. J., & Lyberg-Åhlander, V. (2018) Is children's listening effort in background noise influenced by the speaker's voice quality? *Logopedics Phoniatrics Vocology, 43*(2), 47–55.

Stanfield, R. A., & Zwaan, R. A. (2001). The effect of implied orientation derived from verbal context on picture recognition. *Psychological Science, 12*(2), 153–156.

Svorou, S. (1994). *The grammar of space*. John Benjamins.

Talmy, L. (2000). *Toward a cognitive semantics* (Vol. 2). MIT Press.

The jamovi project. (2019). Jamovi. (Version 1.6.23) [Computer software]. https://www.jamovi.org

Vandaele, J. (2012). What meets the eye. Cognitive narratology for audio description. *Perspectives, 20*(1), 87–102.

Van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. Academic Press.

Vercauteren, G. (2021). Insights from mental model theory and cognitive narratology as a tool for content selection in audio description. *Journal of Audiovisual Translation, 4*(3), 6–24.

Vercauteren, G., & Remael, A. (2015). Spatio-temporal settings. In A. Maszerowska, A. Matamala, & P. Orero (Eds.), *Audio description. New perspectives illustrated* (pp. 61–80). John Benjamins.

Walczak, A. (2017). Immersion in audio description: The impact of style and vocal delivery on users' experience [Unpublished doctoral dissertation]. Universitat Autònoma de Barcelona.

Walczak, A., &Fryer, L. (2018). Vocal delivery of audio description by genre: measuring users' presence. *Perspectives, 26*, 69–83. https://doi.org/10.1080/0907676X.2017.1298634

Warglien, M., Gärdenfors, P., & Westera, M. (2012). Event structure, conceptual spaces and the semantics of verbs. *Theoretical linguistics, 38*(3–4), 159–193.

Wilken, N., & Kruger, J. L. (2016). Putting the audience in the picture: Mise-en-shot and psychological immersion in audio described film. *Across Languages and Cultures*, *17*(2), 251–270.

Zwaan, R. A. (2004). The immersed experiencer: Toward an embodied theory of language comprehension. In B. H. Ross (Ed.), *The psychology of learning and motivation* (Vol. 44, pp. 35–62). Academic Press.

Zwaan, R. A., & Madden, C. J. (2009). Embodied sentence comprehension. In D. Pecher & R. A. Zwaan (Eds.), *Grounding cognition: The role of perception and action in memory, language, and thinking* (pp. 224–245). Cambridge University Press.

Zwaan, R. A. & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin, 123*(2), 162–185.

Zwaan, R. A., Stanfield, R. A., & Yaxley, R. H. (2002). Language comprehenders mentally represent the shapes of objects. *Psychological Science, 13*(2), 168–171.