A stormy sea. A three-masted ship with half-reefed sails can be seen behind high waves with foaming crests.

Jana Holsanova, Lund University
Anita Hildén, Leknyttan Utbildning AB
Maria Salmson and Veronica Kesen Tundell, Tundell och Salmson AB
The study has been carried out with the help of funding from the Swedish Post and Telecom Authority (PTS) within the project "Audio description and audio subtitling on TV". The project is one of the winning contributions to the PTS innovation competition on the theme "Usable media services".

The report authors would like to thank:
DyslexiförbundetFMLS
Synskadades Riksförbund, SRF
Eva Hedberg, project manager for the "Audio description and audio subtitling on TV" project.
All of the participants.
Many thanks to you all!

Would you like the report as an audio or digital version?
Visit Tundell Salmson Lär: www.tslar.se.
Here you can find the report
- to order on CD in DAISY format
- as a downloadable PDF

Tundell Salmson Lär
Tel.: +46 (0) 8 720 20 80
E-post: info@tslar.se
www.tslar.se

AUDIO DESCRIPTION and AUDIO SUBTITLES
-- a study of user preferences

With guidelines for audiovisual media

© 2015 Jana Holsanova, Lund University
© 2015 Anita Hildén, Leknyttan Utbildning AB
© 2015 Tundell och Salmson AB
Editor: Maria Salmson
Design: Jenny Salmson
Printing: Print 24

AUDIO DESCRIPTION
and AUDIO SUBTITLES

– A STUDY OF USER PREFERENCES

With guidelines for audiovisual media

Jana Holsanova, Lund University
Anita Hildén, Leknyttan Utbildning AB
Maria Salmson and Veronica Kesen Tundell,
Tundell och Salmson AB
"I've seen films where they describe the environment before the film starts and also say a little bit about how people are dressed and the relationships between them. It makes things much easier when the film starts."
1 INTRODUCTION

1.1 Background

This study is part of a larger project financed by the Swedish Post and Telecom Authority. The aim of the entire project has been to find a functional technical solution for streaming soundtracks with audio description and audio subtitles for live broadcast TV to smartphones. It must be possible for these soundtracks to be heard simultaneously, in a synchronised fashion with the programmes which are being visually described and/or have audio subtitles. It must be possible for the users to choose to play either one soundtrack – in other words, audio subtitles or audio description – or both simultaneously. The soundtracks will be listened to via an app and the user will be able to control the sound volume directly through their telephone.

The project involves Tundell och Salmson AB, Sveriges Television SVT, Synskadades Riksförfund SRF, DyslexiförbundetFMLS, Syntexa (Daniel Ehrt) and Media-Vision (Per Hållander). The project manager is Eva Hedberg.

The study has been developed as a collaboration between Veronica Kesen Tundell and Maria Salmson, Tundell och Salmson AB; Jana Holsanova, Lund University and Anita Hildén, Leknyttan Utbildning AB.

Jana Holsanova has developed the study design, implemented the study and analysed Part One of the study regarding audio description with the help of Maria Salmson. Anita Hildén has assisted with the analysis of Part Two on audio subtitles.

1.2 Concepts

**Audio description:** A description in words of what is happening in an image\(^1\).

**Audio subtitles:** Reading aloud of text that occurs in moving pictures\(^2\).

**Text in moving pictures:** All sorts of text in moving pictures which can need to be made accessible, for example subtitles, tables, scrolling text, captions or text included as part of the action, for example when somebody writes a message on a paper serviette.

**Subtitles:** The text used to translate or clarify dialogue, normally placed in the bottom half of the viewing window. (There is also text for users who do not have access to sound, for example people with hearing impairment. This text can be in the same language as a dialogue or a translation. This type of text can also contain descriptions of noises other than speech, for example "a door slams".) It is only necessary to read out the subtitles where the original language cannot be understood by the viewer.

**Caption:** Text in the image (often in the upper part of the display window) which describes who is speaking, where the image is taken from et cetera.

**Additional information:** Spoken information about things in a film/programme which are not related to the events depicted, but which can be required to obtain a deeper understanding and clearer image of what is going on. This can be a description of the environment, clothes or of the person’s age, profession or relationships.

**Speech synthesis:** Text read by a synthetic voice, as opposed to read by a human voice.

**Participant, informant:** The study’s test subjects.

---

\(^1\) Image in the broad sense: Photographs, film, theatre, objects at an exhibition et cetera.

\(^2\) Moving images in the sense of video, film, TV programmes et cetera.
User: All those who use audio description or audio subtitles.

Score: The value on a scale of 1 to 7 (where 7 always represents the best case) which the participant considers best corresponds to their experience.

1.3 Objective and purpose

The objective of this study is to draw up recommendations/guidelines for what can function as good audio description from the perspective of a number of users. Even though audio description has traditionally been produced for blind people and those with visual impairment, recent experience shows that many more people can benefit from the service. Audio description can contribute to an increased understanding for those without visual impairments but who for some reason have difficulty following on-screen action or what is taking place in an image.

We also want to investigate the needs and preferences of audio subtitle users. The target group for this service is large and consists of people including children, the elderly, people with visual impairment, reading difficulties/dyslexia, aphasia and other cognitive impairment. Audio subtitles can also be useful for second language learners.

The aim has also been to give TV companies and other actors guidance about the users’ priorities in terms of what should be made accessible through audio description and/or audio subtitles, how this can best be done and to provide information to be used when balancing costs against needs.

1.4 Research

Audio description

Audio description plays a major role in ensuring that blind and visually impaired people are included in society. It is important to carry out research into this area in order to provide further training to visual describers and to develop audio description.

The task of the visual describer is to offer people with visual impairment a richer and more detailed experience. In order to achieve this, the visual describer selects relevant information from the visual scene and expresses it in speech, using vivid and graphic descriptions and activating internal images in the visually impaired (Forceville & Holsanova, in progress). The visual describer describes events, environments and people, including their appearance, clothes and facial expressions. It is also important to "time" the description so as not to disrupt the dialogue etc. To conjure up internal images for visually impaired/blind people in the audience, the visual describer uses linguistic formulations which are vivid and graphic and which stimulate the imagination (Holsanova, in press).

Audio description takes place either live during a film or theatre performance or is based on a script prepared in advance, read out and recorded as an extra soundtrack, for example on DVDs or for playback via an app. Audio description is an extremely complex phenomenon. The following questions emerge: What is important in order to understand the action? What do visually impaired people want to hear and not hear? How should the description be formulated so that the recipient understands and can empathise with the narrative? (Blomberg, Holsanova & Gårdenfors, in progress). The visual describer struggles constantly with the question of what should be described, how it should be described and when it should be described (Holsanova, in progress).

Audio description is a young research field. Initially, it was primarily practitioners and pioneers within audio description and related interest groups (above all in Germany and the USA) around the turn of the millennium who began to reflect about their activities and carry out surveys, interviews and focus group investigations relating to audio description.
Research interest has increased since then (primarily in the UK and Spain) and there are currently four theses, four anthologies and around 70 international publications relating to different aspects of audio description. Research is ongoing within the framework of a variety of disciplines using different methods to study audio description (Holsanova 2015).

Until only a few years ago there was no research on audio description in Sweden or Scandinavia. The first Swedish research initiative began five years ago in the form of workshops in audio description organised by Jana Holsanova at Lund University and Cecilia Wadensjö at Stockholm University, in 2010, 2012 and 2014. The vision behind the three workshops was to create networks between visual describers, users and researchers (Holsanova & Wadensjö, in press). The workshops became a meeting place for researchers from cognitive science, interpretation and translation science, professional visual describers, users, educators within audio description and representatives from government authorities and organisations that work with disabilities, Braille etc.

To date, only a small part of the international research has been devoted to the recipient perspective. In Sweden, we are extremely interested in how users perceive audio description and therefore carry out research into the cognitive aspects of audio description from a recipient perspective (Holsanova, forthc.). Research into aspects such as mental images is described in the book “Syntolkning teori och praktik” (Audio description: theory and practice) (Holsanova, Andrén & Wadensjö, eds., in press) which will be published shortly. The book presents contributions from the above named workshops and builds on knowledge from researchers, visual describers, educators, users and interest organisations. The aim of this broad collaboration is to establish audio description as a research subject in Sweden.

Research has an important task: to create an overall picture of the processes that underlie verbal descriptions of visual scenes. This applies both to the production and reception of audio description; in other words, how the visual describer formulates the verbal description of environments, characters and their gestures, facial expressions, glances, body movements and actions, and how people with visual impairment assimilate this information.

**Audio subtitles**

Text is common in moving pictures. This can mean subtitles translating text from a foreign language, but also tables/diagrams, captions describing who is in the image, the environment the person is in or text with important societal information, for example about fire safety. In addition, there are also often pieces of text involved in the action of the film; for example, something written on a serviette or an email conversation on a computer screen. Text of any sort is an obstacle to those who can’t or are unable to read. Text which is important for comprehension or experience must be made available by being audio subtitled.

Audio subtitles can be created with speech synthesis or a human voice. In Sweden, since the early 2000s SVT has offered audio subtitles with speech synthesis. The service has required the use of two digital boxes and many people have perceived it as being difficult to use. From 2016, SVT will offer audio subtitles in a more user-friendly manner.

In 2015, the Swedish Film Institute has launched a new infrastructure for audio subtitles and audio description. This involves audio description and audio subtitles provided with human voices. The sound files are downloaded in advance to an app on a smartphone and played in sync with the film soundtrack. The sound files are developed to be used with the service in cinemas, but the idea is that these should also be used with films on DVD and TV.
There has been some research into audio subtitling with a view to make foreign films and TV available for people who can’t or are unable to read the text. Audio subtitles are often sufficient for those who only have reading difficulties, but for those with visual impairment a combination of audio description and audio subtitles can be necessary as the audio subtitles alone are insufficient (Braun and Orero 2010). Research has been carried out which shows that audio description, or a combination of audio description and audio subtitles, also provides useful support for people with cognitive disabilities (Ramael 2012).

In a study of audio subtitles with speech synthesis on Danish TV, users were able to “like” the service in relation to various programme types; news, documentaries and drama. The participants in the study included people with visual impairment, reading difficulties, the elderly and second language learners. The participants noted three general problems with the Danish service; the synchronisation and the level and quality of the sound. For drama in particular, the users experienced the greatest problems. For example, they found it difficult to determine who had said what, who was present, and what was going on. People with visual impairments experienced the greatest difficulties (Kvistholm Thrane 2013).

For people with dyslexia/reading and writing difficulties, reading comprehension was improved by bimodal reading where the reader is able to see the text at the same time as it is read out, compared to merely hearing and not seeing the text. This is explained by the fact that bimodal reading reinforces the associations between sound and writing. It is also assumed to release cognitive resources for higher-level comprehension processes (Montali & Lewandowski, 1996).

**Guidelines and standardisation**

Standards and guidelines for audio description and audio subtitles have been drawn up or are in progress within the international standardisation organisations ITU (International Telecommunication Union) and ISO (International Organization for Standardization).

Guidelines for audio description have been drawn up and are ready for publication within the ISO’s work on user interfaces (ISO/IEC Joint Technical Committee 1, SG35/WG6) : ISO/IEC TS 20071-21. The same group is now working on guidelines for audio subtitles. The ITU’s focus group for audiovisual media (FG-AVA) has published a series of technical reports about accessible media, including on the subjects of visual interpretation and audio subtitles. These can be downloaded free of charge on the FG-AVA page on ITU.int. (Part 5: Final report of activities: “Audio/Video description and spoken captions” and Part 13: Audio characteristics for audio descriptions and/or spoken subtitles) which includes technical requirements. Part 13 includes a statement that audio description and audio subtitles should be accessible as separate files without extra charge.

The European collaborative project Hybrid Broadcast Broadband For ALL (HBB4ALL) is working to achieve accessible TV for all users.
"Sometimes you want information, sometimes too much information can kill the atmosphere in some way ... The most important thing is that you find out what's important for the action or what moves the action forwards ... "
2 Method

2.1 Audio description

We have developed a new method for evaluating how the users perceive different versions of audio description.

First, we carried out telephone interviews with the users based on short examples. The participants stated their preferences, which in turn generated a number of important variables for the tests in the next phase. On the basis of these variables, we developed alternative audio description versions in three genres: drama, science and documentaries/facts. These alternative versions were enhanced, for example with descriptions of environment, gestures and facial expressions, feelings etc.

As a next step, we invited two groups of participants to a meeting. The meeting consisted of two parts. In the first part, the participants assessed different aspects of the audio description versions using a score (on a scale of 1 to 7). In the second part, they commented during a focus interview on the quality and details of the audio description versions. Which type of mental performances do the visually impaired obtain from the different types of description? Which types of descriptions are most relevant for them? Are the descriptions understandable? Is anything missing? Do the users feel that they are participating?

The aim of the study has been to examine how verbal descriptions are perceived by the participants, what type of descriptions they prefer and why.

We also tested audio subtitles in combination with audio description and the need for a summary when viewing a new episode of a series.

2.2 Audio subtitles

The method for testing audio subtitles/text and images is the same as that for audio description. Two groups of users assessed audio subtitles, both with speech synthesis and human voices and were then able to choose which of the versions they preferred. We also tested the reading of text and images, in this case captions and audio captions. In addition, the participants were able to test how well the audio subtitles worked when a Swedish voice was subtitled for audibility reasons (poor telephone connection).

Finally, a discussion increased our understanding of their reactions and choices.
"It's extremely important that the audio describer adapts themselves to the type of film or programme they're describing."
3 Results

3.1 Part One: Visual description

3.1.1 Introduction

The visual describer struggles constantly with the question of what should be described, how it should be described and when. So how visual describers should work so that visual description contributes to understanding and empathy? In the context of visual description, direct feedback from the users is often not possible. In addition, the users have different backgrounds, interests and needs. For us, therefore, this study is extremely useful in helping us to find out how users experience visual description and what they prefer.

The main objective of this study is to obtain feedback from the users by means of assessment of concrete examples of visual description of different types.

3.1.2 Research questions

Our preliminary research showed that visually impaired people and other users:

• would like visual description of documentaries and news items
• have different views of how far visual description should go in terms of descriptions of feelings, gestures and facial expressions
• would like to have information about the film/programme before the screening
• would like to have a summary of the previous episode of a series
• avoid programmes in languages they do not understand

On the basis of the preliminary research, we have selected the following research questions.

Do users want visual description that describes as much as possible; in other words, even people’s feelings, facial expressions, gestures and other things (below called DETAILED) or would users prefer a more restricted version (below called MINIMAL)?

1. Is the film/programme’s genre of significance when the users choose between detailed and minimal visual description?
2. Do the users prefer a human voice or is it acceptable to have speech synthesis in films/programmes with subtitles?
3. How great is the need for additional information? Do the users want descriptions of environments and clothing/appearance? Do they want an introduction before a film or a summary before a new episode of a series?

3.1.3 Test films

To test the research questions, we have selected the following films:

Research questions 1) and 2)

Two versions of the first three films were shown; first a MINIMAL one in which the visual describer says what happens in the film without adding anything about environment, visual expression etc. And following this a DETAILED version in which the visual describer adds more descriptions of the environment, interpretations of feelings et cetera.
The documentary **FLIES** is a short film about the breakdance craze. This is a documentary with a rapid tempo and many complicated movements which is difficult to describe in only a few words. Unfortunately we could not test the film on both groups due to technical problems. However, we have still included the score from the group who saw the film and which clearly shows how the visual description contributed to the experience of the film (see Diagram 1).

The nature programme **SEALS** is a section from a film about a seal colony off the coast of the border between North and South Korea. This film has its own narrator who talks about the seal colony. However, the speaker doesn’t say much about the environment, how the seals move in groups beneath the water etc. Here the visual describer can add a lot of information.

The drama **HUGS** is a section from the Swedish film Hur många kramar finns det i världen? (How many hugs are there in the world?) This section comes from the middle of the film and is therefore taken out of context. For the minimal version, only the events are described. In the detailed version, descriptions of the environment, feelings, visual expressions and gestures are added.

**Research question 3)**

**BATHING MICKEY** is a short Danish film with a Danish narrator, Mickey, who speaks Danish. In the film there are subtitles which translate what Mickey says. The film has a visual description in Swedish.

Two versions were shown. In the first a synthesised voice reads the subtitles; in the second the subtitles are read by a human voice (a different voice to that of the visual describer).

**Research question 4)**

**SUMMARY.** A short summary intended to introduce the **HUGS** section was played.

### 3.1.4 The test

During the test, the participants were divided into two groups of five and six participants. Each group sat at a table with a name tag in front of each participant. Each participant had a number of test forms, a sheet of paper for each film version on which four rows of seven raised points represented the scores 1 to 7. On the wall opposite the table the films were shown on a computer monitor.

First, both versions of the three first films were shown. After each showing of a film, the following questions were asked, and the participants answered using a scale of 1 to 7, where 7 is the highest.

- Was the visual description informative?
- Was the visual description comprehensible?
- Did the visual description contribute to empathy?
- Could you picture what was going on?

After showing the test films, the participants had a break followed by discussions of the different versions. During the discussions, the conversation was recorded so that it could be transcribed later.

During the discussions, the test film **BATHING MICKEY** with subtitles, together with a summary intended for the section from **HUGS** were shown.
3.1.5 The participants

The eleven participants (five women and six men) in the study had different serious visual impairments, from completely blind (five people) to partial blindness/low visual acuity (two people).

Their ages ranged from 25 to 70 years, with a predominance of upper middle age. Eight of the participants are employed, one unemployed and two are pensioners. During the test, the participants were divided into two groups, with five people in one group and six in the other.

All completely blind participants were in group 1 (with six participants), while group 2 (with five participants) contained more people with residual sight.

3.1.6 The participants' responses to test questions

Test 1 FLIES

The first film, FLIES, was shown to only one group of participants (see 3.1.3).

FLIES MINIMAL (FM)

Few high scores were given here. Even though some participants considered that the minimal audio description was completely understandable (7), the average was very low at between 2.2 and 3.0. Four of the participants gave consistently very low scores (average 1.3) and one gave medium-high (average 3.6).

FLIES DETAILED (FD)

The visual describer says so much that it is possible to follow the film but there is very little space to describe the movements etc. as the programme has its own narrator who speaks almost the whole time.

The detailed audio description contributes to a great deal of understanding, but it is still difficult for viewers to understand what is going on. Nobody gave the highest score to the question “Could you picture what was going on?”. But three participants gave a 5. One participant gave the lowest score to all questions in both versions. One gave consistently high scores (5s or 7s). Picturing what was going on produced the lowest average score, 3.8, for the detailed description, while information, understanding and empathy received higher scores.

FLIES was not included in the final discussion about which version the participants preferred and why, because only one group had been able to see the film.

Test 2 SEALS

SEALS MINIMAL (SM)

Two participants consistently gave the highest score. The lowest score was a 2 for the question about whether the audio description was informative. There were also two 3s for the question about whether the audio description contributed to empathy. Otherwise, only good scores were given.

SEALS DETAILED (SD)

The detailed variant improved the experience still further. No score was lower than 4. There was one 4 for the question about empathy and one for the question about whether the participants could picture what was going on, although not from the same person. The same person who gave only 7s for the minimal version did the same for the
detailed version. And two more people gave only 7s. 6s and 7s dominated strongly.

Picturing what was going on received the lowest average score (6.1).

Both the minimal and detailed audio description now gave much greater understanding and empathy than the two versions of FLIES.

Test 3 HUGS

HUGS MINIMAL (HM)

One person consistently gave a score of 7. The lowest score was 3s, above all for the question on whether the audio description was informative. Otherwise, 4 (= neither good nor bad) was a normal score here. The best average score was received by the question about whether the participants could picture what was going on (average 5.3) and the worst by the question on informativeness (average score 4.5). The minimal variant of HUGS received a slightly lower score than the minimal variant of SEALS.

HUGS DETAILED (HD)

The detailed version gave the highest average score for all variants. All scores but one were a 6 or 7.

Diagram 1 shows the average score the groups gave regarding informativeness, understanding, empathy and the possibility of understanding what was going on in both versions of the three films (note that the result for FM and FD represents group 1 only).

If we compare the factual film SEALS with the drama KRAMAR, we can see that the largest point difference is between the minimal and detailed versions of KRAMAR. A drama therefore has much to gain from detailed audio description – and this applies for all of our test questions. Informativeness, understanding, empathy and the possibility of understanding what was going on increase considerably. For the factual film, the difference is significantly smaller.
We obtain a clearer image of the difference between the minimal and detailed versions if all minimal or detailed versions are combined as in **Diagram 2**.

![Diagram 2](image)

**Diagram 2.** The average for all minimal and detailed versions, per genre.

In **Diagram 3**, we compare the average score the groups gave to the different test films regarding informativeness, understanding, empathy and the possibility of understanding what was going on.

![Diagram 3](image)

**Diagram 3.** Average score for the different assessment questions, per test film.

In this diagram too it is clear that all detailed versions receive higher scores than the minimal ones in all test questions, but also that the minimal versions of **SEALS** and **HUGS** were given higher scores than the detailed version of **FLIES**, except as regards informativeness where the minimal version of **HUGS** received a lower average score than the detailed version of **FLIES**.

One interesting question is whether the assessments of the test questions differ in any crucial manner between the two groups, since group 1 contains more blind test people and group 2 more with residual sight.

In **Diagram 4**, the differences in average score are compared between the groups regarding **SEALS MINIMAL**.
Diagram 5 provides a corresponding image for **SEALS DETAILED**

Diagram 4. Average score for both groups for the test film **SEALS MINIMAL**

Diagram 5. Average score for both groups for the test film **SEALS DETAILED**

Regarding **SEALS MINIMAL**, group 1 has better understanding, empathy and perception of what is happening than group 2, while informativeness seems to have been lost for group 1 in the minimal variant. In the detailed version, informativeness is equally great for both groups.

Regarding understanding, the differences between the groups are very small and there is also very little difference between the two versions. Regarding empathy, it is improved for both groups in the detailed version. But group 1 still gives a slightly higher empathy score. Finally, regarding understanding what is going on, group 2 now gives a much higher score for the detailed version, while group 1 remains on the same score as for the minimal version.
Diagram 6 shows both groups' average scores for HUGS MINIMAL while diagram 7 shows the scores for HUGS DETAILED.

Diagram 6. Average score for both groups for the test film HUGS MINIMAL

Diagram 7. Average score for both groups for the test film HUGS DETAILED

Here we can detect the same pattern as for SEALS, at least in terms of understanding and empathy, where group 1 gives a higher score than group 2 in the minimal version. The detailed version receives a high score from both groups and there are very small differences between them.

Because the groups are so small, we cannot draw any conclusions from these results, but we can detect a tendency – namely that completely blind people would have more to gain from a MINIMAL audio description than users with residual sight – a trend that it could be useful to investigate further.
3.1.7 Comparison between versions. Summary of group discussion

When both versions of the film had been shown, a discussion was launched about what the users preferred and why. The conversation was recorded.

Comparison between versions of SEALS.

The detailed version was preferred by all participants, partly because the minimal version was considered to be almost unnecessary because it adds very little to what the narrator's voice says.

Several users felt that they had obtained what they wanted from the programme without audio description. However, if they had sat and focused on this specific programme, they would have preferred the second version as it gave the entire experience. It was clear that the second version was felt to be better, more informative and that it facilitated interpretation.

Several people also pointed out that male visual describers should be chosen for female speakers and vice versa. The test film contained female speaker voices and female visual describers.

When it comes to "poetic" interpretations in the detailed version, the majority of participants seemed to accept this, but some of them pointed out that it is not possible to know whether what the visual describer says is correct. But, as one comment explained, it must be possible in that situation to rely on the visual describer's interpretation. It must also be possible to rely on the fact that there is nothing to interpret when it is quiet.

Finally, a desire was expressed to have name tags et cetera – in other words audio captions – read out.

Comparison between versions of HUGS

All participants preferred the more detailed version. One person expressed the opinion that much greater concentration is required to follow the minimal version, and another that in the detailed version the atmosphere and relationships between people became clear.

On the other hand, descriptions of feelings and visual expressions can be used more sparingly where these are clear from the context. "For example, if someone is crying, you don’t need to say that, because we can hear it", said one participant. And another: "We have often had much more practice than sighted people in listening to voices and we have different opinions about what audio description should be like. Should it replace what can be seen, or should it describe what the images are conveying?"

In films with quick scene changes, the visual describer’s role can be to help by saying who is speaking and where they are. As the film progresses, the listener becomes more used to the voices and no longer needs that information. And of course different types of interpretation are required for detective stories and nature films. It is extremely important to vary the audio description according to what type of programme is being described.

The two final tests were presented during the group discussion.

Test 4 SUMMARY

A SUMMARY of HUGS was shown, of the type that would be used if this section of the film had been a new episode in a series. The summary gives information about the characters and the environment in which the film is set.
The question is whether the participants would like a summary or whether they can manage just as well without.

Here, opinion was split. Some people thought that it would be useful to know a little about the interiors and environment in advance, because this kind of information can’t be presented during the actual film. Some people thought that it is not up to the visual describer to summarise earlier episodes. A visually impaired person who has missed an episode has the same problem as a sighted person who has missed it. "But", responded another participant, “the sighted person gets this kind of information using their sight during the episode. And we miss out on that.”

Another person questioned whether summaries are required. You watch a series from episode 1 and don’t start halfway through. And if you have missed one of the episodes you look for it on the Internet. Some people protested and said that not everyone can cope with the technology.

But if there is going to be a summary, it must explain what has happened in the previous episode and not provide general information about the programme, which the users said the summary in question did.

But most participants agreed that they would like an introduction that describes environments, how people are dressed et cetera before they start to watch a film and that it would be useful to have this introduction in an app.

**Test 5 BATHING MICKEY**

Finally, the Danish film Bathing Mickey, which is subtitled in Swedish, was shown. The film is visually described and we have added audio subtitling for the subtitles.

In version 1, the subtitles are read out with speech synthesis and in version 2 with a human voice (a different voice to that of the visual describer).

The question is which version the users prefer.

Several people think that the speech synthesis version is clearer but they get distracted by incorrect pronunciation and emphasis. And the fact that they are listening more to how the voice sounds than to what it is saying. The majority of participants prefer the human voice. But one opinion is that the use of speech synthesis makes clear the difference between audio subtitling, the speech in the film and the audio description.

One participant says that if more effort could be put into the speech synthesis so that it didn’t pronounce and emphasise incorrectly, it would be almost the same as human speech. On the other hand, it takes time to work with it and this is impossible in situations such as news broadcasting. Then the participants would rather accept the unprocessed speech synthesis version than not have the text read out at all.
3.2 Part Two: Audio subtitles

3.2.1 Introduction

Many foreign films and TV programmes are now subtitled. However, this doesn't help people who are unable to read the text, but who instead require the text to be read out. This can be done with a human voice, but in many cases it is not possible to create audio subtitling. Then speech synthesis can be an alternative.

But there are also other texts in images, such as information signs and audio captions that state who is talking. There are also subtitles of text which is in Swedish; for example, unclear speech (for example in a telephone interview) or a dialect which can be difficult to understand.

The main objective of this study is to obtain feedback from the participants by means of assessment of concrete examples of audio subtitling with human or speech synthesised voices.

3.2.2 Research questions

Unlike the audio description section, we did not carry out any preliminary research for this section. However, we did have discussions with representatives of people with dyslexia/reading and writing difficulties and developed the following research issues as being useful to test:

1) Is speech synthesis sufficient for audio subtitling, or is it better to use a human voice?

2) Is there a difference between drama and factual programmes regarding the choice of reading voice?

3) Is it important that everything that can be read (in other words all the text visible in the image) is read?

3.2.3 Test films

To test our research questions, we used recordings of different types and with different degrees of text, including with text in captions, audio captions and subtitling of Swedish voices.

Bathing Mickey (BATHING MICKEY). Bathing Mickey is a short Danish film with a Danish narrator, Mickey, who speaks Danish. In the film there are subtitles which translate what Mickey says.

Report feature from Gaza. (REPORT) Features which are not in the Swedish language are subtitled. This also includes audio captions.

Report feature with telephone interview. (TELEPHONE INTERVIEW). As the sound quality is poor for the voice on the telephone, the entire interview is subtitled. Both the interviewer and the voice on the telephone are speaking Swedish.

Report feature with subtitled captions. (CAPTIONS.)
3.2.4 Test questions

During the test, the participants were divided into four groups with four to five participants in each group. Each group sat at a table with a name tag in front of each participant. Each participant had a number of test forms. In the left-hand margin of the paper were nine points and against the first six points, in horizontal rows, the figures 1 to 7 representing scores of 1 to 7 where 7 was the best. The three lowest points (7, 8 and 9) were used to mark which of the versions was best or alternatively if they were equal. On the table the films were shown on a computer monitor.

Round 1

To test research question 1, MICKEY was shown first with speech synthesis reading the subtitles and then with a human voice reading the subtitles.

The same research question was tested in the same way with the report feature from Gaza.

The test form asked the test participants to assess:

1) how well they heard the audio subtitling (7 = heard everything extremely well)
2) if the audio subtitling disrupted their experience of the feature (7 = did not disrupt at all) and
3) if it was disruptive that they could not simultaneously hear the original voices in the feature (in other words, the people whose voices were subtitled in the feature). (7 = not at all disruptive)

When the participants had seen the four features (two with speech synthesis and two with a human voice), they ticked one of points 7-9 (7 = the speech synthesis is better than the human voice, 8 = the human voice is better than the speech synthesis, 9 = it doesn’t make any difference).

This round concluded with a discussion which was recorded and later transcribed.

Research question 2 was tested with the same feature as in research question 1, and in the following discussion the participants were requested to state whether the type of programme makes a difference – for example a fictional film such as MICKEY compared to a news feature such as REPORT – in the choice between human voice and speech synthesis.

Research question 3 was tested in various ways in the following three rounds.

Round 2

In Report feature with telephone interview, (TELEPHONE INTERVIEW), the subtitle caption was read out by speech synthesis, which to a large extent then overlapped the original Swedish voices. The informants were requested to assess:

1) if they could clearly hear the voice
2) if the speech synthesis disrupted their experience of the feature, and
3) if, despite the speech synthesis, they could hear and understand what the original voices were saying.

This type of clash can occur if all features with subtitles are audio subtitled. In the following discussion, we wanted to know whether this was acceptable or whether the participants had suggestions for solutions.
Round 3
After this, a report feature with subtitled captions was shown (CAPTIONS). The captions were read out by a human voice.

First, the caption was read out immediately it was displayed on the screen, and then it often clashed with other speech. In the next version, the caption was read out as soon as possible without disrupting other speech. The participants were requested to assess which of the versions worked best and whether they wanted the captions read out at all.

Round 4
Even the report feature from Gaza contains text, in the form of audio captions. An audio caption is a graphic element inserted onto the image which, for example, gives information about who is speaking. This feature was now shown again, but with the information being read out by a human voice. Because the feature was subtitled, the complication arose that the audio for the subtitling sometimes clashed with the audio for the audio caption.

In addition to responses to the test form of the same type as in round 1, a discussion was held after rounds 3 and 4.

3.2.5 The participants
The filmed features were shown to four groups with 4 to 5 people in each group (18 informants in total) and with different degrees of dyslexia/reading and writing difficulties or other reading problems such as concentration difficulties or autism. Immediately after each feature, the informants filled in a form with questions specific to the feature.

3.2.6 Comparison between versions. Summary of group discussion.

Drama
The first film sequence, Bathing Mickey (hereafter MICKEY) is a short film in which the main character talks Danish. At the same time that the Swedish text is being shown at the bottom of the film, it is read out using speech synthesis in Swedish.

After having seen this film sequence, the test participants evaluated whether they could clearly hear what the speech synthesis was saying. They graded their responses from 1 (did not hear anything) to 7 (heard extremely well). All responses showed that they all heard the text well and four of the participants even felt that they heard it extremely well.

The next film sequence was once again MICKEY, with the difference that the text was read by a human voice. The majority of participants considered that they clearly heard what the voice said, but for half of the group the score fell by one point. Only one of the test participants considered that it was easier to hear the human voice and three test participants felt that they could hear the text equally well.

The next question was about whether the speech synthesis disrupted their experience of watching the film. Here, responses varied from two participants who felt that it was extremely disrupting to two who felt that it not disrupt the experience at all. Here the variation was individual and did not covariate with the score of how clearly they heard the voice. The same question was asked when they had seen the same film sequence with a human voice. The responses were individual and a pattern cannot be seen among the participants. However, they responded that the experience of watching the film was disrupted slightly less.
It was to a large extent the same situation when it came to scoring whether they were disturbed by the fact that they could not hear the original Danish voice completely. The responses to this question were equally spread from 1 to 7 just like the previous question. However, when they watched the film sequence with the human voice they responded that this was less disturbing.

Next, the test participants evaluated both variants of speech synthesis and human voice. One person felt that the speech synthesis was better than the human voice and also responded that the speech synthesis was easier to hear. The other participants responded that the human voice was better despite the fact that half of them had responded that they heard the synthetic speech better.

In the interviews it emerged that human speech was better than speech synthesis for films. “Human speech is more pleasant, the speech is more differentiated and you can follow it in a different way.” Another person preferred the human voice “when you’re supposed to feel changes in feelings and so on, particularly in films.” Some people also pointed out that for children and young people, who were not as used to speech synthesis, it was important and undoubtedly also easier for them to listen to human speech.

Several test participants found the Danish voice in the background disruptive. They felt that both voices were equally strong and that it was necessary to tone down the background voices so that they could concentrate on the human voice or the speech synthesis. Some people said that it was difficult and said that with concentration difficulties it was hard to know which voice to listen to.

The suggestion from the test participants was that it would be useful to be able to regulate the volume of the voices in the app. Some other participants suggested having an earbud only in one ear and using the other ear to listen to the film. They also pointed out that it was not possible to completely turn off the sound from the film as then they would miss music and other important sound information.

It is more important to have human speech in films or TV series for relaxation purposes as speech synthesis means that the programme becomes a bit too artificial. “Not as relaxing, you sit there and make fun of the pronunciation and then you kind of lose the entertainment value.”

**News feature**

The test participants watched a news feature from Swedish television; a report from Gaza where the interviews were subtitled. The subtitles were read out using speech synthesis and a human voice.

The result showed that six of the test participants gave a score one point higher for their perception of clearly hearing the text with speech synthesis compared to with a human voice. For two people, the two were the same and two people felt that the human voice could be heard better.

When they then assessed which voice they felt was best, half of them responded that they felt the human voice was better, while four people felt that it didn’t make any difference. Only one person responded that the speech synthesis was better and this was not the person who felt that the speech synthesis was better in films.

To the question about whether the voice that read the text disrupted the news feature, half of the participants responded that it did not disrupt it at all, while the rest felt that it did. The same people felt that the news feature was disrupted as felt that the film sequence was disrupted.

When the same feature had a human voice reading the text, only one person felt that
their experience of watching the feature was somewhat disrupted and the remaining people responded that their experience was not disrupted.

To the third question about whether they felt it was a problem they could not hear what the original voices said, the majority responded that it was not a problem. The person who felt that it was a problem also felt that it was a problem not to be able to hear the interviewees’ voices.

In this news feature the people being interviewed spoke Arabic and of this one of the test participants said: "Then you couldn’t understand and then your ears seek out the Swedish voice reading out the subtitles."

Diagram 8 shows the participants’ average scores for how they perceived the news features with the subtitles read by speech synthesis and by a human voice.

Diagram 8. The average score for the experience with the audio subtitling per test question

**News feature with subtitles and Swedish speech in the background**

This test task consisted of a report feature with a telephone interview with a man speaking Swedish where the interview was captioned as a result of poor sound quality. The text was read out using speech synthesis.

Of the test participants, four felt that they could clearly hear what the speech synthesis said while the rest considered that it was difficult to hear and two people could not hear at all. For this news feature, the majority of participants considered that the speech synthesis disrupted the feature and only two people considered that they could hear everything the Swedish voice said.

During the interviews, it emerged that they found it difficult to ignore the Swedish speech in the telephone interview and only listen to the speech synthesis. Several people said that when the interview in Swedish was also subtitled, text and unclear speech can complement each other. "They subtitle it too. Then you can probably put it all together, even if the telephone is a bit crackly. If you have the subtitles to help it should be possible."

**News feature with subtitled captions**

The next news feature contained inserted captions. The captions were on the screen at the same time that people were talking. The text in these captions was read out. On the first occasion, the text was read in sync with the showing of the caption, and in the second it was adapted to the other speech.

Eight of the test participants felt that they heard the text well or very well regardless of whether the text was read out in sync or adapted. However, the two people who had difficulties hearing the text felt it was slightly easier to hear when it was adapted.
Despite the fact that they felt it was easy to hear the text, half of the participants felt that they could not hear the other voices and the remainder that they had difficulties hearing the voices. The adapted reading increased the possibilities of everyone hearing better what the other voices said.

In a comparison of reading out the caption so that it did not clash with other speech or not reading at the caption at all, half of the participants responded that they would like the caption read out so it did not clash with other speech. The remainder felt that they did not want any captions read out at all.

During the discussion, the majority felt that “it was awful, because you missed out on a lot of the actual news article”. They wanted the opportunity to choose to have these captions read out or not. "This was terribly messy. It got really difficult." "You had to listen to three voices at the same time."

**News feature with subtitles and audio captions.**

The final film sequence was a news feature from Gaza; the same one that they had seen earlier but now with reading of texts describing who was speaking and where they were. Despite this, the test participants’ perception of hearing the caption voice was considerably worse. Two people felt that they could hear the caption voice very well but simultaneously felt that they could not at all hear what the others said. Nor could they hear the subtitles clearly. The other participants felt that they found it difficult to hear both the caption voice and what the other speakers said.

"You missed out on a great deal of the actual article. I don’t need to know the name of a man in Gaza and the information is completely irrelevant compared to the news in itself." When they discussed the captions in general, several people said that there can be news features where captions could be important. Some people said that they have time to read because the caption is shown for a while. One person said that "if it’s important societal information, and you really can’t keep up, then it’s important for it to be read out."

One wish was for the presenter to contribute the information included in the captions. An alternative was for space to be created in the sound flow to insert reading of the text present in the captions.

**Diagram 9** shows the experience of reading information other than the subtitles.

![Diagram 9](image)

Diagram 9. The participants’ average scores for how clearly they could hear the captions being read at the same time compared to adapted to other speech and the reading of audio captions which clashed with the reading of the subtitles in the feature.
Summary of discussion

In summary, all participants were happy that the texts were being read out and felt that it was a human right for everyone to be able to acquaint themselves with information.

It is exciting to see that the majority felt the speech synthesis was slightly easier to hear but that they chose a human voice. It was most important to have a human voice for films and TV series. For news features they were more equal. Here, speech synthesis could be more acceptable.

During the test, the participants felt that speech synthesis was largely good but they suggested that it would be useful to be able to individually select the preferred voice via an app. They also said that they would want to be able to have a female voice for certain occasions and a male voice for others.

They wanted to be able to regulate the sound level for both reading and original sound.

The test participants have experience of different readers and feel that high quality is important in a reader. "If you choose a poor reader, one that you can’t follow, then I think that you ruin the film completely."

They also felt that it would be useful to be able to decide themselves if they want to listen to a human voice or speech synthesis.

The test tasks that led to the biggest discussion were those with audio captions, audio subtitles and original sound. They found it difficult to perceive all of the voices and messages. "The last one, where there were three voices on top of each other, that was an altercation and you didn’t know whether to listen to the Arabic or Swedish voices."

The test participants suggested that instead of subtitles or at the same time as subtitles, an introductory announcement should be provided. "Now you’re going to hear NN from blah blah newspaper, or whatever. They can say now we’re going to this feature with this person and say the name." "When it’s a foreign name, it can be useful to hear the right pronunciation, they can be difficult to read."

The participants also said that if captions are to be read, the newsreader must pause in their speech while the captions are read out so that you don’t end up with several voices at the same time.
"I want to hear as much as possible and there is nothing as disruptive as an audio describer, but it's the least bad option because otherwise you understand nothing. I enjoy watching opera and then it's always disruptive because you have to hear both the opera and the audio describer. But I want everything. "
4 Summary and guidelines

4.1 Answers to our research questions

4.1.1 Audio description

Do the participants want a audio description that describes as much as possible or do they prefer a more restricted version?

- The majority prefer the detailed version and also choose the versions that contain descriptions of expressions of emotion etc. even when, in the discussions, they say that it is necessary to be careful with too extensive interpretations.

Is the film/programme's genre of significance when the user chooses between DETAILED and MINIMAL audio description?

- From the discussion, it emerged that it is more important to have a detailed audio description in a fictional film/drama than in a factual film/documentary. See also Diagram 2 where the differences are clear.

Do the participants prefer a human voice or is it acceptable to have speech synthesis in films/programmes with subtitles?

- The majority feel that human voices are preferable. But they would rather accept speech synthesis version than not have the text read out at all.

How great is the need for additional information? Do participants want descriptions of environment, clothes and appearance? Do they want an introduction before a film or a summary before a new episode of a series?

- Many want to choose to have additional information, introductions or summaries via an app or link.

4.1.2 Audio subtitles

Is speech synthesis sufficient for audio subtitling, or is it better to use a human voice?

- It appears to be about as easy to hear a synthetic voice as a human one. But if they can choose, the participants prefer a human voice.

Is there a difference between drama and factual programmes regarding the choice of reading voice?

- During the discussion, it emerged that human voices are preferred in fictional films/dramas. The choice of voice is important and the participants wanted to be able to choose themselves between a male and female voice (also applies to speech synthesis).

Is it important that everything that can be read is read, even captions?

- From the test, it emerged that reading of captions is often disruptive, but that it is better that they are read out when they do not disrupt other speech in the feature, even if they are not read out in sync with the caption being displayed. Half of the participants felt that it was better not to read the captions at all.

- It is also problematic when Swedish voices are supplemented with subtitles. It is difficult to simultaneously listen to the Swedish voices and the audio subtitling, and most of the content is lost. In such situations, it is important that users are able to turn off the audio subtitling themselves.
4.2 Guidelines

Audio description

- Audio description should never disrupt the film dialogue.
- Use a female visual describer if there is a male speaker and vice versa.
- Make an audio description as detailed as possible without disrupting the film dialogue. Gestures and visual expressions are also of value.
- Don’t interpret feelings which are evident and can be read from how the voices sound; sighs, crying and so on.
- Describe environments, clothing etc. as far as possible.
- If possible, read the subtitles with a voice other than that of the visual describer.
- For drama films, provide an introduction that talks about things that are difficult to understand during the film: the main characters, how they look, how they live and so on.

Audio subtitles

- Important societal information that scrolls across the screen must be read out one or more times, ideally without other important information being lost.
- Choose a human voice for audio subtitles for drama or TV series.
- Speech synthesis is acceptable for factual programmes and news.
- Speech synthesis is better than no reading at all.
- Reading of captions (name, function, location etc.) or text which is part of the narrative should be avoided if it clashes with other important information. The importance of the text for comprehension and experience of the whole is crucial in deciding whether it should be read out or not. Those in charge of the audio subtitling must prioritise. Ideally, the user should be involved in the process.
- A newsreader can make features more accessible by inserting information conveyed in text into their presentation.

Prioritisation

Where prioritisation must be carried out between different types of programmes/genres, on the basis of this study we can state the following:

- Audio description is particularly important for drama or series programmes/films. Factual programmes with their own narrator can make the programme understandable without audio description.
- An introduction can in some cases go a long way, for example in making a factual programme understandable.
- Use a human voice for reading subtitles for drama/film and series.
- Speech synthesis can be used to read subtitles for news programmes, documentaries or similar programmes.
Conclusion

The study has provided answers to our research questions, but also given rise to new ones. Some areas which can be of interest for further study:

There are interesting differences between how completely blind and visually impaired people view the necessity and degree of audio description. Our groups were too small for us to be able to draw any definite conclusions. In comparing the study's main areas – audio description and audio subtitles – we also feel that the differences can be larger between blind and visually impaired people than between visually impaired people and those with reading difficulties.

Another interesting question is when in the flow a subtitle should be read out. For the visually impaired, it is important that they are able to hear the voices of those speaking, so that they know who they are, before the reading begins. For people with reading difficulties, the need appears to be the opposite. If the text is shown before the reading begins, they automatically look at the text and the user begins the laborious process of reading. They then risk missing a lot of both what is said and what is shown. If the reading instead begins immediately before the text is shown, the user can focus on listening and watching.

Is it possible to find a balance for this that both groups can accept?

This study has not examined live interpretation of films and TV programmes. This is an area that we feel should be studied in more depth and we hope that a future study can bring more clarity to how the users view this type of audio description.
REFERENCES

References, Audio description


References, Audio subtitles


Would you like the report as an audio or digital version?
Visit www.tslar.se.
Here you can find the report
- to order on CD in DAISY format and
- as a downloadable PDF

Tundell Salmson Lär
Tel.: +46 (0) 8 720 20 80
E-post: info@tslar.se
www.tslar.se