International Telecommunication Union

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

T.701.25

(03/2022)

SERIES T: TERMINALS FOR TELEMATIC SERVICES

User interfaces - Accessibility and human factors

Guidance on the audio presentation of text in videos, including captions, subtitles and other on-screen text

Recommendation ITU-T T.701.25

ITU-T

ITU-T T-SERIES RECOMMENDATIONS

TERMINALS FOR TELEMATIC SERVICES

Facsimile – Framework	T.0–T.19
Still-image compression – Test charts	T.20–T.29
Facsimile – Group 3 protocols	T.30–T.39
Colour representation	T.40–T.49
Character coding	T.50–T.59
Facsimile – Group 4 protocols	T.60–T.69
Telematic services – Framework	T.70–T.79
Still-image compression – JPEG-1, Bi-level and JBIG	T.80–T.89
Telematic services – ISDN Terminals and protocols	T.90–T.99
Videotext – Framework	T.100–T.109
Data protocols for multimedia conferencing	T.120–T.149
Telewriting	T.150–T.159
Multimedia and hypermedia framework	T.170–T.189
Cooperative document handling	T.190–T.199
Telematic services – Interworking	T.300–T.399
Open document architecture	T.400–T.429
Document transfer and manipulation	T.430–T.449
Document application profile	T.500–T.509
Communication application profile	T.510–T.559
Telematic services – Equipment characteristics	T.560–T.619
General multimedia application frameworks	T.620–T.649
User interfaces - Accessibility and human factors	**T.700–T.799**
Still-image compression – JPEG 2000	T.800–T.829
Still-image compression | JPEG XR	T.830–T.849
Still-image compression – JPEG-1 extensions	T.850–T.899

For further details, please refer to the list of ITU-T Recommendations.

Recommendation ITU-T T.701.25

Guidance on the audio presentation of text in videos, including captions, subtitles and other on-screen text

Summary

Audiovisual content (such as video) often contain text, which cannot be easily accessed by a wide section of the audience. While captions/subtitles provide text alternatives to audio elements in audiovisual content, other on-screen text may have various functions. It can be part of the story (as a message written on a piece of paper by one of the characters) or it can provide additional information (such as graphs, emergency alerts or superimposed titles). Complementarily, audio description provides a description of audiovisual content auditorily, including captions/subtitles and other on-screen text if present and are of particular benefit to persons who, for different reasons, cannot access on-screen text. However, some users may only require captions/subtitles and other on-screen text to be made accessible as audio because they already have access to other visual content such as the images.

Recommendation ITU-T T.701.25 provides guidance for audiovisual content producers, distributors and exhibitors on the audio presentation of captions/subtitles and other on-screen text. It acknowledges the relationship with existing access services such as audio description. While considering current implementations, as well as future possibilities suggested by research, and bearing in mind possible trade-offs between quantity and quality, this document positions itself for situations in which various access services coexist and users are given the choice to select those best suited to their needs.

Recommendation ITU-T T.701.25 is twin with the published ISO/IEC TS 20071-25:2017 "Information Technology – User interface component accessibility – Part 25: Guidance on the audio presentation of text in videos, including captions, subtitles and other on-screen text" developed by ISO/IEC JTC1 SC35.

History

Edition	Recommendation	Approval	Study Group	Unique ID[*]
1.0	ITU-T T.701.25	2022-03-29	16	11.1002/1000/14973

Keywords

Accessibility, audio presentation of text in videos, captions, on-screen text, subtitles.

[*] To access the Recommendation, type the URL http://handle.itu.int/ in the address field of your web browser, followed by the Recommendation's unique ID. For example, http://handle.itu.int/11.1002/1000/11830-en.

FOREWORD

The International Telecommunication Union (ITU) is the United Nations specialized agency in the field of telecommunications, information and communication technologies (ICTs). The ITU Telecommunication Standardization Sector (ITU-T) is a permanent organ of ITU. ITU-T is responsible for studying technical, operating and tariff questions and issuing Recommendations on them with a view to standardizing telecommunications on a worldwide basis.

The World Telecommunication Standardization Assembly (WTSA), which meets every four years, establishes the topics for study by the ITU-T study groups which, in turn, produce Recommendations on these topics.

The approval of ITU-T Recommendations is covered by the procedure laid down in WTSA Resolution 1.

In some areas of information technology which fall within ITU-T's purview, the necessary standards are prepared on a collaborative basis with ISO and IEC.

NOTE

In this Recommendation, the expression "Administration" is used for conciseness to indicate both a telecommunication administration and a recognized operating agency.

Compliance with this Recommendation is voluntary. However, the Recommendation may contain certain mandatory provisions (to ensure, e.g., interoperability or applicability) and compliance with the Recommendation is achieved when all of these mandatory provisions are met. The words "shall" or some other obligatory language such as "must" and the negative equivalents are used to express requirements. The use of such words does not suggest that compliance with the Recommendation is required of any party.

INTELLECTUAL PROPERTY RIGHTS

ITU draws attention to the possibility that the practice or implementation of this Recommendation may involve the use of a claimed Intellectual Property Right. ITU takes no position concerning the evidence, validity or applicability of claimed Intellectual Property Rights, whether asserted by ITU members or others outside of the Recommendation development process.

As of the date of approval of this Recommendation, ITU had not received notice of intellectual property, protected by patents/software copyrights, which may be required to implement this Recommendation. However, implementers are cautioned that this may not represent the latest information and are therefore strongly urged to consult the appropriate ITU-T databases available via the ITU-T website at http://www.itu.int/ITU-T/ipr/.

Table of Contents

Page

Introduction

Videos are omnipresent in our digital society and are used to inform, educate and entertain audiences. Videos often contain text, which cannot be easily accessed by a wide section of the audience. This text includes captions/subtitles and other on-screen text that is part of the visual content.

Captions/subtitles provide text alternatives to audio elements. Other on-screen text may have various functions. It can be part of the story, as a message written on a piece of paper by one of the characters. It can also provide additional information, such as graphs, emergency alerts or superimposed titles.

Persons who, for different reasons, cannot access on-screen text will benefit from an audio presentation. This oral rendering is often part of audio description (see [b-ITU-T T.701.21], an access service providing a description of audiovisual content auditorily, including captions/subtitles and other on-screen text if present. However, some users may only require captions/subtitles and other on-screen text to be made accessible because they already have access to other visual content such as the images.

This Recommendation provides guidance for video producers, distributors and exhibitors on the audio presentation of captions/subtitles and other on-screen text. It acknowledges the relationship with existing access services such as audio description. While considering current implementations, as well as future possibilities suggested by research, and bearing in mind possible trade-offs between quantity and quality, this document positions itself for situations in which various access services coexist and users are given the choice to select those best suited to their needs.

Recommendation ITU-T T.701.25

Guidance on the audio presentation of text in videos, including captions, subtitles and other on-screen text

1 Scope

This Recommendation provides recommendations on the audio presentation of captions/subtitles and other on-screen text for use in all types of videos regardless of the language and technology being used to transmit and present the recorded or live video.

This Recommendation applies to making captions/subtitles and other on-screen text accessible to users with various needs, including but not limited to people with learning and reading disabilities, people with cognitive disabilities, people who are blind or have low vision, older people, and non-native language speakers. It does not apply to captions/subtitles or other on-screen text whose content is already provided in the soundtrack in a language and a way users can access.

This Recommendation provides guidance on spoken captions/subtitles as a stand-alone access service, but it also provides guidance on how to integrate spoken captions/subtitles, other spoken on-screen text and audio description, if needed, in different types of videos.

NOTE 1 – Extensive guidance on audio description is provided in [b-ITU-T T.701.21].

This Recommendation does not consider the devices or transmission mechanisms used to deliver and play the content or the audio presentation of text in videos. These devices include, but are not limited to televisions, computers, wireless devices, projection equipment, digital versatile disk (DVD) and home cinema equipment, cinema equipment and other forms of user interface technology. Therefore, this Recommendation does not consider transcoding files for the various video and audio outputs.

NOTE 2 – Technical matters of transmission and distribution are covered by other international standards (e.g., MPEG standards and other technical international standards such as [b-IEC 62731]).

This Recommendation acknowledges the various needs and preferences of users, as well as the different approaches to the audio presentation of text in videos.

It applies to audio presentations intended to be heard simultaneously along with the original video.

2 References

The following ITU-T Recommendations and other references contain provisions which, through reference in this text, constitute provisions of this Recommendation. At the time of publication, the editions indicated were valid. All Recommendations and other references are subject to revision; users of this Recommendation are therefore encouraged to investigate the possibility of applying the most recent edition of the Recommendations and other references listed below. A list of the currently valid ITU-T Recommendations is regularly published. The reference to a document within this Recommendation does not give it, as a stand-alone document, the status of a Recommendation.

None.

3 Definitions

3.1 Terms defined elsewhere

None.

3.2 Terms defined in this Recommendation

This Recommendation defines the following terms:

3.2.1 General terms

3.2.1.1 programme: Complete unit of a recorded or live video.

3.2.1.2 programme category: Classification of programmes.

NOTE – Programme categories are not necessarily mutually exclusive.

EXAMPLE – Programme categories include documentary, news and information, and drama.

3.2.1.3 video: Combination of audio and visual content presented together in a synchronized manner via ICT.

NOTE – While the visual content is often presented using a screen, it might also be presented via other technologies, e.g., a projected hologram.

3.2.2 Audio-specific terms

3.2.2.1 audio description; descriptive audio: Audiovisual content described in an audio modality.

NOTE 1 – Audio description can also be used to describe locations, directions and objects.

NOTE 2 – Audio description can be used to describe sound not easily identified or coming from an unknown source or location.

3.2.2.2 narrator: Person(s) and/or technology which voices the alternative audio information.

NOTE 1 – Alternative audio information includes audio description and/or spoken captions/subtitles and/or spoken on-screen text.

NOTE 2 – In audio description, narrators are also referred to as describers or audio describers.

3.2.2.3 spoken captions/subtitles; audio captions/subtitles: Captions/subtitles that are voiced over the audiovisual content.

NOTE – In this Recommendation, the term "spoken captions/subtitles" will be used.

3.2.2.4 spoken on-screen text: Text, other than captions/subtitles, that is voiced over the audiovisual content.

3.2.3 Language of presentation terms

3.2.3.1 original language: Language in which audiovisual content is produced.

3.2.3.2 dubbing: Secondary audio version of a video produced in a language other than the original language of the video and timed to match the voicing of the original actors.

NOTE – The secondary audio version is lip-synchronized and replaces the original dialogue, which cannot be heard.

3.2.3.3 voice-over: Secondary audio version of a video produced in a language other than the original language of the video which overlaps with the voicing of the original actors.

NOTE – The secondary audio version is not lip-synchronized and does not replace the original dialogue, which can still be heard.

3.2.3.4 captions/subtitles: Transcription or translation of audio content visually presented together with the content.

NOTE 1 – Transcriptions or translations include speech and/or non-speech information.

NOTE 2 – Transcriptions or translations are often suitable for use as an alternative or a complement to the audio content.

3.2.4 Production terms

3.2.4.1 pre-mixed production: Process that involves delivering alternative audio information together with the audio stream of the video as one single audio track.

NOTE 1 – Although pre-mixed is used in the definition to refer to the audio mix, it can also refer to a production process in which the original audio (including both the soundtrack and the dialogues), the alternative audio information in the video and the original visual content are delivered together.

NOTE 2 – Pre-mixed productions are also referred to as broadcast-mixed productions.

3.2.4.2 receiver-mixed production: Process that involves delivering alternative audio information separately from the audio stream of the video and having them mixed in a device controlled by the user.

NOTE – Alternative audio information can be delivered to the user or be downloaded or streamed from the Internet as separate services or as services mixed in different combinations.

4 Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

DVD	Digital Versatile Disk
SDH	Subtitles for the Deaf and Hard-of-hearing

5 Conventions

None.

6 Framework and process considerations

6.1 General

6.1.1 Alternate names for the audio presentation of text in videos

Terms used in this Recommendation vary according to country, language, region, video content, and type of text in videos.

They include audio subtitles, spoken subtitles, spoken captions, audio captions, to refer to the audio presentation of captions/subtitles. For the purpose of this Recommendation, these terms are synonymous, as defined in clause 3.

They include audio text, to refer to the audio presentation of all text in videos.

NOTE – In some jurisdictions, there are precise usages defined for one or more of these terms. Individuals can consult their own country's regulations for the locally appropriate terminology.

6.1.2 Motivation for the audio presentation of text in videos

Videos are everywhere in our society, and they often contain text such as captions/subtitles and other on-screen text.

This is particularly the case in countries or regions where captions/subtitles are used to translate content into another language.

All these texts share two features: they are visual and at the same time they are verbal. Users who might not be able to fully access the content include:

– users who cannot access the visual elements: persons with sensorial disabilities such as blind and visually impaired audiences, and also persons who for other reasons (for instance, not being in front of the video display) cannot see the visuals; and

– users with difficulties to access the written verbal content.

NOTE 1 – Reading is a complex cognitive process and, very often, the coexistence of visual stimuli and the speed at which written text is presented makes it difficult for certain users to access written text: this includes persons with reading difficulties caused by a lack of reading ability, dyslexia or cognitive diversity, but also includes children, the elderly and people learning a new language.

Not being able to access text in videos has a direct impact on the understanding and enjoyment of videos. It also implies that certain users are excluded from educational, cultural and social contexts (e.g., when a movie is discussed by colleagues in informal contexts).

Facilitating access to text in videos improves the viewing experience in terms of comprehension and enjoyment, and guarantees access in critical emergency situations where information is provided via text.

Providing an audio presentation of text in videos enhances access to video content.

NOTE 2 – While nowadays the audio presentation of text generally contains spoken captions/subtitles only, users also request the audio presentation of other on-screen text.

NOTE 3 – While audio description usually includes an audio presentation of all relevant captions/subtitles and other on-screen text, some users only require written text to be made accessible since they already have access to the visual content.

6.2 Types of text in videos

6.2.1 Captions/subtitles

Video content can include:

a) text which provides a translation of the original language of the video, addressing users who do not understand the spoken words;

b) text which provides a translation of the original language of the video plus additional features (character identification, sound effects indications, etc.), addressing users who do not understand the language and cannot access the audio;

c) text which provides a transcription of the original language, addressing users who do not have access to the audio; the transcription can be a verbatim or an edited version of the spoken words, including additional features (character identification, sound effects indications, etc.);

d) text which provides a transcription of the original language, addressing users who for various reasons have difficulties understanding the spoken words (e.g., strong dialect, bad quality audio).

Depending on function, users, country, language, and region, this text can be referred to as captions, subtitles, subtitles for the deaf and hard-of-hearing (SDH), intralingual or same-language subtitles, and interlingual subtitles.

An audio presentation of captions/subtitles should be provided only when the original dialogue cannot be understood by the user.

6.2.2 Other on-screen text

Video content can include:

a) logos, which can display text often indicating the name of the company;

b) credits: opening credits present a selection of the main production and cast members, whilst end credits include a more extensive list with additional information;

c) superimposed titles, which are used to present the title of the video, often at the beginning, or to insert additional information such as the spatio-temporal settings (some videos also include intertitles);

d) scrolling texts included at the beginning of a video (e.g., to introduce an episode);

e) graphs, tables and other figures including text;

f) scrolling tickers (crawlers/slides), providing emergency alerts, breaking news, running messages from social media, and other information; and

g) popping-up text messages or captions.

Elements which are part of the story being told in the video can also contain text that is important to be made accessible.

EXAMPLE 1 – An object, a piece of paper or a computer screen which is part of a movie can include text.

An audio presentation may not be needed when the information provided by the on-screen text is already delivered auditorily.

EXAMPLE 2 – If a news presenter already delivers orally the information included in text presented on screen, there is no need to provide an audio presentation of such text.

Text in videos can use colour, type font, positioning, size and other effects for creative and artistic purposes, adding additional layers of meaning. It is important for these visual elements to be transmitted, if possible, to users who do not have access to the visuals.

6.3 Types of audio presentations of text in videos

6.3.1 Live and recorded spoken captions/subtitles and spoken on-screen text

Live spoken captions/subtitles and spoken on-screen text can be used for live and recorded programmes. They involve creating and delivering audio presentations of text in videos in real time along with the delivery of the original video.

Live productions are typically distributed initially in real time, but might be recorded for later redistribution. In this case, recorded spoken captions/subtitles and spoken on-screen text should be prioritized.

NOTE – Although being transmitted live, captions/subtitles can sometimes be semi-prepared when there is a script available. This is known as semi-live captioning/subtitling or as-live captioning/subtitling. The preparation method (live/semi-live) will have an impact on the quality of the caption/subtitles and, consequently, on its audio presentation.

Recorded spoken captions/subtitles and spoken on-screen text are used for recorded programmes and involve creating and recording the audio presentation of text in videos prior to delivery. They allow for careful planning and evaluation before delivery, and are therefore recommended where possible.

6.3.2 Pre-mixed and receiver-mixed productions

In pre-mixed productions, the audio presentation of text in videos is created and mixed with the original audio before its distribution.

In receiver-mixed productions, there are different possibilities:

a) The audio presentation of text in videos is created before distribution and mixed in the user device.

NOTE 1 – The audio presentation can also be retrieved from the Internet.

b) Text content is sent to the user device, where it is converted into an audio presentation by a text-to-speech system, allowing users to select the synthetic voice of their choice.

NOTE 2 – Text content can also be retrieved from the Internet.

NOTE 3 – Receiver-mixed productions allow users with different needs to access the same video simultaneously.

EXAMPLE – One user accesses the video with the audio presentation of text using headphones while another user watches it in the same room without the audio presentation of text.

6.4 Creating and delivering audio presentations of text in videos

6.4.1 Narrator preparation

The audio presentation of text in videos may be delivered by a human voice or by a synthetic voice or a combination of both.

The narrator, be it a person or a technology, should have the following:

a) good native language skills;

b) the ability to articulate.

When using synthetic voices, special care should be taken in their selection. Voice quality, naturalness and reading speed vary depending on the synthetic voices. Availability and quality of synthetic voices vary depending on the language.

NOTE 1 – User testing indicates that human voices are generally preferred in certain programme categories, especially in drama, but synthetic voices are also accepted and might be prioritized, especially when users are familiarized with them. Research indicates that, when speech synthesis is used, it might avoid confusing the human voice of the original dialogue with the synthetic voice.

NOTE 2 – Synthetic voices reading split subtitles (i.e., when one sentence is divided into two or more subtitles) might produce unnatural pauses that can affect the comprehension or enjoyment of users.

6.4.2 Volume

The audio presentation of text in video should be prepared in sound objects so that users can adjust the volume of the original audio and the volume of the audio presentation of text in video separately. This is only possible in receiver-mixed productions.

In pre-mixed productions, where one single audio track is received by the user, the volume of the audio presentation of text in videos should have a good mix with the volume of the original audio, including both the dialogues and the soundtrack.

6.4.3 Audio quality

Cut off, microphone noise, background noise, levels between the audio presentation of text in videos and the original soundtrack, and other audio factors should be mixed and filtered so as to guarantee comprehension and enjoyment by users.

6.4.4 Inclusion of users in creating the audio presentation of text in videos

The full range of users should be included in the process of creating audio of text in videos where possible.

NOTE – This can take the form of focus groups, employment of users as marketers or community consultants. Users might also be included in situations where stakeholder groups are actively involved in the creation and marketing processes.

6.4.5 Inclusion of users in evaluating the audio presentation of text in videos

If evaluation is part of the process of creating audio presentations of text in videos, representative users should be involved.

NOTE 1 – It is important to evaluate recorded spoken captions/subtitles and spoken on-screen text prior to delivery wherever possible.

NOTE 2 – It is important for evaluations to focus on the understandability.

Users' feedback should be systematically registered and taken into account to improve the process.

6.5 Synchronization

Spoken captions/subtitles should generally be delivered in synchronization with the captions/subtitles displayed on screen. They should also be synchronized with visual action such as body language, where possible, and take into account audio elements such as music and other sounds.

When spoken captions/subtitles are delivered in synchronization with the captions/subtitles on screen, the original dialogue might be heard on the background or not. When the original dialogue is not heard, the final effect is similar to dubbing. When the original dialogue is heard underneath, a voice-over effect is achieved.

NOTE – Research on user preferences is still needed.

Spoken on-screen text may be delivered in a synchronized or a non-synchronized manner.

EXAMPLE – Where time constraints prevent reading all the opening credits in a movie in a synchronous way, a selection is provided either before or after they appear on screen.

Credits should be read at a normal pace without rushing, and effort should be made to include as many as possible.

6.6 Text-files

Captions/subtitles can be produced in a way that they are always visible (open) or in a way that they can be turned on and off by the user (closed). Open captions/subtitles are generally burned in the video.

For the purpose of accessibility, captions/subtitles and other on-screen text should be included as separate text-files, as this is the best option for converting text into audio.

7 Guidance on the audio presentation of text in videos

7.1 General considerations

7.1.1 Different strategies

Different strategies apply when:

a) creating separate tracks for the spoken captions/subtitles, for the spoken on-screen text, and for the audio description, which can be used independently or combined to suit the needs of diverse users;

b) integrating the spoken captions/subtitles and spoken on-screen text in the audio description.

NOTE 1 – Although the first option caters for the needs of a wider section of the population, both scenarios are considered in this Recommendation when describing the strategies used.

NOTE 2 – The integration of spoken on-screen text in the spoken captions/subtitles track is easier to achieve when a human narrator is used.

7.1.2 User considerations

The following aspects are important to consider:

a) At the beginning of each programme, a verbal and a visual notification should be presented to make users aware that an audio presentation of text is available.

b) The availability of the audio presentation of text in a programme should be identified by a standardized logo, both on the screen and in any media where it is advertised.

   NOTE 1 – It is recognized that logos might vary from country to country. It is important that logos are as consistent across as many jurisdictions as possible.

c) Users should be easily able to access information (both in real time and in advance) that identifies when the audio presentation of text is available and for which programmes.

d) Users should be provided with an easy to use means of selecting and changing between presenting:

1) spoken captions/subtitles;

2) spoken on-screen text;

3) audio description;

4) any combination of the previous access services;

5) no access services.

NOTE 2 – Audio description traditionally includes a verbal rendition of text in videos. However, when mixing traditional audio descriptions with the audio presentation of text, overlapping and inconsistencies might occur. This is why an audio description that does not include the audio presentation of text is more suitable when interacting with spoken captions/subtitles and spoken on-screen text. This caters for the needs and preferences of a wider sector of the population.

7.1.3 Availability across technologies

The audio presentation of text in videos should be available regardless of the technology being used to transmit and/or present the original content.

NOTE 1 – Some technologies used for transmission include cable, satellite, Internet, DVD, stream, catch up.

NOTE 2 – Some technologies used for presentation include television, computer, smart phone, tablet, and cinema.

7.1.4 Consistency within a video and programme series

The style of the audio presentation of text should be consistent throughout a video and the programmes in a series.

7.2 Developing the audio presentation of text in videos

7.2.1 Clarity in the audio presentation of text in videos

The audio presentation of text in videos should be clearly voiced in a manner that can be understood by its intended users.

7.2.2 Reading/delivering the audio presentation of text in videos

Spoken captions/subtitles can be delivered in an acted or a non-acted way. Voices with different accents can also be used.

NOTE 1 – User preferences concerning emotional narrations and accents vary.

Different voices matching the gender and the age of the original voices should be used where possible.

NOTE 2 – More research is needed, especially in countries where voice-over is delivered by a single narrator.

7.3 Levels of importance

Too much information can interfere with a user's ability to understand the programme.

When various types of text coexist simultaneously in a video, it may not be possible to include all the text in the audio presentation.

EXAMPLE 1 – A video can present different types of text simultaneously: for instance, subtitles, a scrawling text and a graph.

Levels of importance should be determined by the persons responsible for developing the audio presentation of text in videos and will vary between different contexts. These levels are: essential information, significant information, helpful information, and unhelpful information. The levels are based on clause 4.4 of [b-ITU-T T.701.21], but "unhelpful information" in this Recommendation replaces "irrelevant information" in [b-ITU-T T.701.21].

Levels of importance should be used to select the text in videos to be included in the audio presentation.

EXAMPLE 2 – An emergency alert is considered essential information and is prioritized over other text.

NOTE – It is useful to include users in making this decision where possible.

7.4 Guidance on identifying the audio presentation of text in videos

The audio presentation of text in videos should be clearly identifiable from other audio content.

Users should also be able to distinguish the types of text in videos (captions/subtitles, logos, superimposed titles, credits, emergency alerts, etc.) being presented.

This can be achieved by using:

a) Voices: Different voices are generally used to distinguish between audio description and the audio presentation of text in videos.

   EXAMPLE – Two voices, male and female, are used. One is used to read the audio presentation of text in videos and one to read the audio description script.

b) Prosody: When only one voice is available, changing the prosody can help the user infer that a text is being read.

   NOTE 1 – Although the quality of synthetic voices is improving, prosody is still a commonly reported issue.

c) Descriptive words: A descriptive word indicating the type of text that will be read can also be used.

   EXAMPLE – The presentation of text is preceded by a phrase to identify the type of text such as: "a caption reads", "subtitle", "text message" or "emergency alert".

NOTE 2 – While the use of earcons (audio signals) as an identification mechanism has been proposed, there is insufficient research to standardize their use at the present time.

7.5 Guidance on identifying characters

Different voices may be used to distinguish between various characters in a video.

When only one voice is used, prosody helps distinguish between various characters in a video.

Characters can also be identified by including their name before reading the caption/subtitle.

EXAMPLE 1 – "Mary: I like going to the cinema".

NOTE 1 – When captions/subtitles are integrated in the audio description, they can be rendered as direct speech or indirect speech, indicating in both cases the character who voices the text.

EXAMPLE 2 – "It's raining, says Mary" or "Mary says that it's raining".

NOTE 2 – When spoken captions/subtitles and audio description are integrated, the audio description can help the user infer who is speaking (e.g., indicating the direction of the gaze), in place of an explicit identification of the character.

The audio presentation of text in videos with a voice-over effect (i.e., the original voice is heard underneath) also contributes to character identification.

7.6 Guidance on text and speech adjustments

When numerous texts are present on screen, rendering all of them literally might cause synchronization problems.

The spoken caption/subtitle should not be longer than the span of time during which the caption/subtitle is presented on screen, to avoid overlapping with the subsequent caption/subtitle. Overlaps with the audio description should be avoided. If the audio presentation of text is continually longer than the text presented on screen, the lack of synchronization will accumulate and soon affect comprehensibility.

Fade in and fade out should be taken into consideration when calculating the reading time available for synchronisation purposes.

NOTE 1 – Numbers (for instance, years) and abbreviations normally take longer to read out and can cause synchronization issues.

Strategies to overcome synchronization issues include:

a) increasing the reading speed;

   NOTE 2 – While higher reading speeds can negatively affect some users' comprehension, other users might be used to faster reading speeds.

   NOTE 3 – Synthetic voices differ in their reading speeds, and the reading speeds of each voice can be altered. Ongoing studies indicate that adjusting the reading speed of spoken captions/subtitles only in longer utterances so that they fit in the time slot might be a suitable solution.

b) altering the text: text in videos can be shortened, rephrased or expanded; it can even be omitted if the content is already transmitted via other means.

   NOTE 4 – When converting captions/subtitles into spoken captions/subtitles, text can also be adapted to resemble oral language wording.

   NOTE 5 – In some cases, it is not possible or even desirable to provide a literal and synchronized rendering of the text in videos.

EXAMPLE – Users might prefer to listen to a song rather than to its audio presentation. Possible solutions would be omitting the song text or fitting adjusted spoken captions/subtitles when convenient.

Bibliography

[b-ITU-T T.701.21]	Recommendation ITU-T T.701.21 (2022), *Guidance on audio description.*
[b-IEC 62731]	IEC 62731:2013, *Text-to-speech for television – General requirements*.
[b-ISO/IEC TS 20071-25]	ISO/IEC TS 20071-25:2017, *Information Technology – User interface component accessibility – Part 25: Guidance on the audio presentation of text in videos, including captions, subtitles and other on-screen text*.

SERIES OF ITU-T RECOMMENDATIONS

Series A	Organization of the work of ITU-T
Series D	Tariff and accounting principles and international telecommunication/ICT economic and policy issues
Series E	Overall network operation, telephone service, service operation and human factors
Series F	Non-telephone telecommunication services
Series G	Transmission systems and media, digital systems and networks
Series H	Audiovisual and multimedia systems
Series I	Integrated services digital network
Series J	Cable networks and transmission of television, sound programme and other multimedia signals
Series K	Protection against interference
Series L	Environment and ICTs, climate change, e-waste, energy efficiency; construction, installation and protection of cables and other elements of outside plant
Series M	Telecommunication management, including TMN and network maintenance
Series N	Maintenance: international sound programme and television transmission circuits
Series O	Specifications of measuring equipment
Series P	Telephone transmission quality, telephone installations, local line networks
Series Q	Switching and signalling, and associated measurements and tests
Series R	Telegraph transmission
Series S	Telegraph services terminal equipment
Series T	**Terminals for telematic services**
Series U	Telegraph switching
Series V	Data communication over the telephone network
Series X	Data networks, open system communications and security
Series Y	Global information infrastructure, Internet protocol aspects, next-generation networks, Internet of Things and smart cities
Series Z	Languages and general software aspects for telecommunication systems

Printed in Switzerland
Geneva, 2022