International Telecommunication Union

# ITU-T Technical Paper

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

(17 October 2019)

## FSTP-ACC-RCS
## Overview of remote captioning services

**Summary**

This Technical Paper describes remote captioning services. It defines reference model, requirements and functionality that facilitate, via an assistive intermediary (i.e., real time captioner or via voice recognition software), to enable the inclusive meeting participation of person either on site or remotely. The aim of this Technical Paper is to enable meeting organisers to provide attendees located even in different countries to access to real time captioning services and give comparable and equivalent understanding and participation experience. Depending on the type of meeting, it can be a one-to-one meeting, a group meeting of persons in different locations using conferencing tools, a teleconference call or in a physical conference setting.

Remote captioning is often displayed on a screen or TV monitor in the meeting room. Participants can also access remote captioning on their laptops, pads or smartphone using URL link provided to them.

This revision addresses some aspects of data protection.

**Keywords**

Remote captioning services (RCS); CART; respeaking; verbatim speech to text reporting; speech to text reporter.

**Change Log**

This document contains Version 2 of the ITU-T Technical Paper on "*Overview of remote captioning services*" approved at the ITU-T Study Group 16 meeting held in Geneva, 7-17 October 2019.

This version supersedes Version 1, which was approved at the ITU-T Study Group 16 meeting held in Ljubljana, 20 July 2018.

**Editor:** Lidia Smolarek-Best
European Federation of Hard of Hearing People (EFHOH)
UK
Email: smolarek-best@hotmail.co.uk

# CONTENTS

**Page**



## List of Figures

**Page**

# Technical Paper ITU-T FSTP-ACC-RCS

## Overview of remote captioning services

### 1 Scope

The scope of this document is to give an overview of and to describe requirements for remote captioning services. It describes some examples on the production side, including the use of automatic speech recognition. The document also contains examples of best practice key performance indicators (KPIs).

### 2 References


[ITU-T F.745] ITU-T H.745 (2016), *Functional requirements for network-based speech-to-speech translation services.* https://www.itu.int/rec/T-REC-F.745

[ITU-T F.791] ITU-T F.791 (2018), *Accessibility terms and definitions.* https://www.itu.int/rec/T-REC-F.791

[ITU-T H.625] ITU-T H.625 (2017), *Architecture for network-based speech-to-speech translation service.* https://www.itu.int/rec/T-REC-H.625

[ITU-T P.10] ITU-T P.10 (1998), *Vocabulary of terms on telephone transmission quality and telephone sets*. https://www.itu.int/rec/T-REC-P.10

[ITU-T FSTP-ACC-RemPart] Technical Paper ITU-T FSTP-ACC-RemPart (2015), *Guidelines for supporting remote participation in meetings for all*. https://www.itu.int/pub/T-TUT-FSTP-2015-ACC

[ITU-T FSTP-AM] Technical Paper ITU-T FSTP-AM (2015), *Guidelines for accessible meetings.* https://www.itu.int/pub/T-TUT-FSTP-2015-AM

[Arnott-et-al] Arnott J. L., Newell A. F., Downton A. C., *A comparison of palantype and stenograph for use in a speech transcription aid for the deaf*, J Biomed Eng. 1979 Jul;1(3):201-10.

[E-Mich] E-Michigan Deaf and Hard of Hearing People, *Communication Access Realtime Translation (CART)*. http://webcache.googleusercontent.com/search?q=cache:_QRjsUi6-BMJ:www.hearingloss-mi.org/2016/03/455/+&cd=1&hl=en&ct=clnk&gl=jp (visited 2018-08-30)

[NCRA] USA certification site for verbatim reporters guidance. https://www.ncra.org/certification (visited 2018-08-30)

[Noffz] *European variations in services on text transcription*. https://efhoh.org/wp-content/uploads/2017/04/Birgit-Nofftz-Germany-2014-Speech-to-Text.pdf

[Romero-Fresco] *The NER Model for accuracy in captioning.* https://link.springer.com/chapter/10.1057/9781137552891_3

## 3 Terms and definitions

### 3.1 Terms defined elsewhere

This Technical Paper uses the following terms defined elsewhere:

**3.1.1 automatic speech recognition (ASR) [ITU-T P.10]:** A system that can recognize continuous speech, often having phoneme-sized references, using lexical, syntactic, semantic, and pragmatic knowledge, and reacts appropriately (therefore having interpreted the message and found the corresponding action to be taken).

**3.1.2 captions/captioning [ITU-T F.791]:** Captions are a real time transcription of spoken words, sound effects, relevant musical cues, and other relevant audio information in live or pre-recorded events. They can be open, not adjustable by the user, or closed where they can be turned on and off by the users at will. See clause 3.13 of ITU-T F.791 for further explanation of open and closed accessible services. Any audio information presented at the meeting, including all spoken presentations and announcements, audio tracks of audio-visual presentations and questions from the audience should be captioned in real time into a synchronous text transcript (real time captions).

**3.1.3 machine translation (MT) [ITU-T-F.745]:** Text in a source language is converted by computers into text in a target language which has the same meaning as the original text in the source language.

**3.1.4 real time [ITU-T-F.791]:** Data or services (e.g., broadcasting) that are transmitted with virtually no delay.

**3.1.5 respeaking [ITU-T-F.791]:** A technique to produce captions where a person ("the respeaker") listens to the speech and re-speaks it, such that the respeaker's vocal input is processed by a speech recognition software which transcribes it and produces the captions.

**3.1.6 speech-to-speech translation [ITU-T-F.745]:** Speech in a source language is translated into speech in a target language.

### 3.2 Terms defined here

This Technical Paper defines the following terms:

**3.2.1 communication access real time translation (CART):** CART is a North American term for real time captioning which is used for meetings to enable participation for persons with disabilities as described in [E-Mich]. This service can be provided by means of either textual or graphical supplementary content. The captions and the dialogue are usually in the same language. The service is primarily to assist users having difficulty hearing the sound. They are also beneficial to those persons who do not understand the spoken language for other reasons, e.g., language is not that person's first language, e.g., sign language. (See "real time captioning".)

**3.2.2 captioner:** A verbatim reporter who records word-for-word using exactly the same words in the same order as the speaker's original statement, which should not be open to interpretation or editing with a transcription technique or device.

**3.2.3 palantype:** Verbatim Reporters use specially designed keyboards in order to type the spoken word verbatim. Unlike a QWERTY keyboard, not every letter in a word is pressed, but several keys will be pressed at once which represent whole words, phrases or short forms. Specially designed computer software will then convert these phonetic chords back into spoken language (e.g., English), which can then be displayed for someone to read.

**3.2.4 real time captioning:** A service that transmits captioned data with virtually no delay.

**3.2.5 speech to text reporter (STTR):** A European term for verbatim reporter providing access services to deaf and hard of hearing people using Stenotype or Palantype.

**3.2.6 stenography:** A form of shorthand typing done on a special machine which makes it possible to produce simultaneously a verbatim transcript as described in [Arnott-et-al].

**3.2.7 syllabic chord keyboard:** A keyboard where multiple keys are being pressed simultaneously, creating complete syllables or words at once, in contrast to the standard QWERTY keyboard, on which keys are being pressed one by one, and single characters appear. It is not a phonetic system, but orthographic. The Velotype keyboard is an example of chord-keyboard.

**3.2.8 transcription:** In the linguistic sense, it is the systematic representation of language in written form.

## 4 Abbreviations

| | |
|---|---|
| ASR | Automatic speech recognition |
| CART | Communication access real time translation |
| KPI | Key performance indicator |
| MT | Machine translation |
| RTC | Real time captioning |
| S2ST | Speech-to-speech translation |
| STTR | Speech to text reporter |
| WPM | Words per minute |

## 5 Conventions

None.

## 6 Background

### 6.1 What is remote captioning?

Remote captioning is provided by a qualified verbatim reporter often known and STTR or captioner at a remote location from the physical location of the meeting. A voice connection such as a telephone, cell phone, or computer microphone is used to send the voice to the captioner, and the real time text is transmitted back over a modem, Internet, or other data connection. Remote captioning is often displayed on a screen or TV monitor in the meeting room. Participants can also access remote captioning on their laptops, pads or smartphone using URL link provided to them by organizers.

### 6.2 Who needs remote captioning?

While remote captioning serves persons who are deaf, hard of hearing, deafblind persons and persons with reading disabilities, it is also useful for people whose first language is different from the language being used, to understand speakers with different voices and accents in various group situations (at work, in education, community events, etc.).

## 7 Streaming of transcribed text

Streaming of transcribed text is sent from remote captioner to be displayed at the event using an Internet-based platform. The text can also be accessed using individual devices both in the meeting room and remotely throughout the world. Different platforms are currently in use by service providers, including their own in-house system. See Figure 7-1.

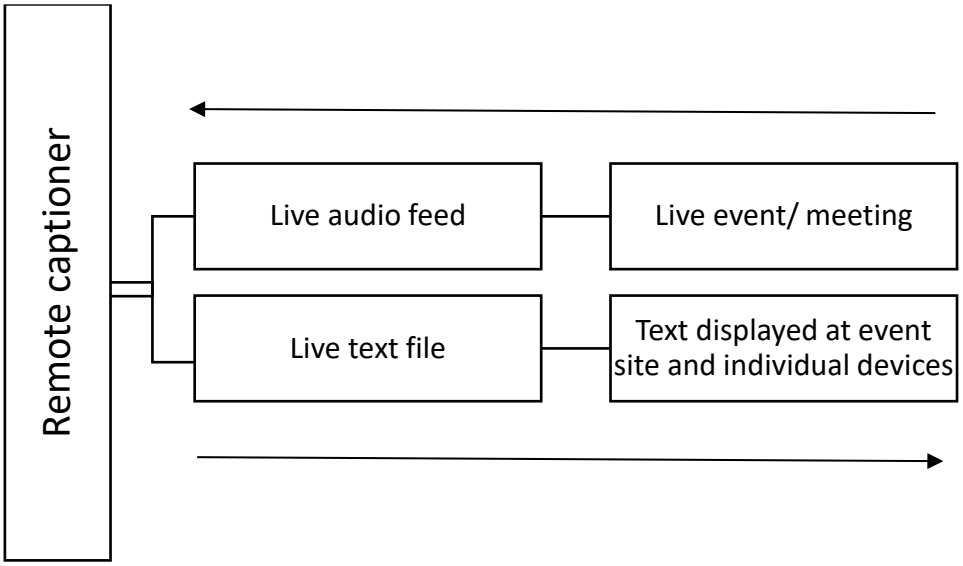


**Figure 7-1 – Simple diagram of the remote captioning process**

To ensure successful streaming the venue host needs to ensure correct set up before the event takes place with checking both audio input and text output.

It is the best to use a hard-wired Internet access at both ends, which will allow for a more stable connection, and also better security. Streaming via public Wi-Fi networks risks interruptions and consequently exclusion of the participants from full participation. Organisers should pay attention to potential firewall barriers.

# 8 Technical aspects of verbatim reporters' equipment

## 8.1 Use of automatic speech recognition for remote captioning service

The remote captioning service can use automatic speech recognition (ASR) for speech-to-text, and to transmit the resulting text to the PWD client. One way to use ASR is described in [ITU-T F.745] and [ITU-T H.625].

### 8.1.1 Service description of speech-to-speech translation for machine translation

The service defined in [ITU-T F.745] enables users to leverage network-based speech-to-speech translation (S2ST) for real time captioning using machine translation (MT). This service is generically realized by the following operations, which is taken from clause 6.3 *Service description* of [ITU-T F.745].

1. Speech is input by users to S2ST clients.
2. S2ST clients send device and speech information to ASR servers.
3. Speech in source language is recognized by ASR servers.
4. The recognized speech is encoded by ASR servers and transferred to MT servers.

An MT server in [ITU-T F.745] then converts text in source language A to text in target language B. If the source and target language are the same language, then MT is a trivial text-to-text conversion with no changes.

### 8.1.2 Service description of S2ST for real-time captioning or communication access real time translation

This S2ST process can be modified for real-time captioning (RTC) to produce a speech-to-text effect, in the following way (see Figure 8-1):

1. Speech is input by users to S2ST clients.
2. S2ST clients send device and speech information to ASR servers.
3. Speech in source language is recognized by ASR servers.
4. The recognized speech is encoded by ASR servers and transferred to RTC or communication access real time translation (CART) servers.

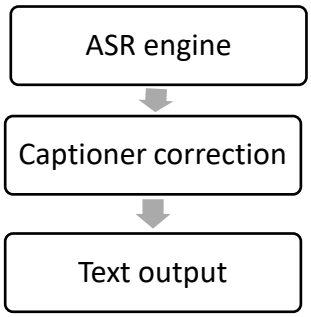


**Figure 8-1 – Simple diagram of delivering captioning using ASR engine**

## 9 Respeaking

In respeaking, a verbatim reporter is trained to repeat incoming audio word for word into an ASR engine which transcribes the words into text as described in clause 8. The transcribed text is sent to the user's terminal screen. See Figure 9-1.

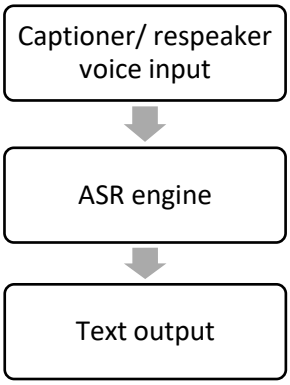


**Figure 9-1 – Simple diagram of delivering captioning using ASR engine in respeaking**

## 10 Security aspects of remote captioning

All remote captioners should follow the code of practice as set out in their own country/jurisdiction and must take all reasonable precautions to secure the privacy of meetings since they can contain confidential information that should not be disclosed e.g., outside the event place.

As a minimum, each remote captioned session must have a different URL link, even if the request for remote captioning is from same client. It is often possible that different sessions are organized in the same building and different participants may be involved.

Confidential information could be any of the following (non-exhaustive list):

– Sensitive information in meetings and/or conferences;
– Medical information on patients;
– Legal information on persons involved;
– Privacy of those involved;
– Medical sessions;
– Counselling sessions.

## 11 Compliance with data protection regulations

All remote captioning service providers need to follow the applicable privacy and data protection laws and regulations. An example for the European context can be found in [b-GDPR].

## 12 Encryption of transcribed text from a remote captioner to the PWD's terminal screen

The text streamed to PWD's terminal need to have protection in form of password to ensure confidential information is only shared with those who were invited to a meeting remotely. The remote captioning provider is responsible for security of the script and shall follow relevant data protection requirements. The transcribed text should be:

– Password protected – Both the captioner and the client using remote Internet-based CART services need to sign in with a username and password;
– SSL encryption – In addition to complex URLs (i.e., protection by obscurity), the link to a secure event will have an https:// address to avoid e.g. man-in-the-middle attacks.

## 13 Encryption of audio streams from the meeting place to the remote captioner

The audio streamed from the event to captioner needs to be protected.

## 14 Quality of remote captioner's transcribed text

The quality of remote captioning text can be established by standardised tests. For the English language, tests used to provide certified qualifications for court reporters described in [NCRA] can be used. [Noffz] describes adaptations that can be made for other European languages which have varied level of captioning service development. Additionally, captioners are able to measure text output with their software, and clients can use the NER Model described in [Romero-Fresco] to measure both captioning speed and quality. See examples of key performance indicators (KPIs) in Annex A.

For persons with visual difficulties and some deaf-blind persons, tablets or separate monitor may need to be provided by the organisers at the venue to allow them to adjust font size to their needs.

To aid organisers and remote captioners in achieving desirable levels of service, the following steps need to be taken:

1. Use of qualified reporters: whether booking directly with captioners or with agencies, captioners should be qualified and (if applicable) registered.
2. Identifying speakers:
    a. Provide the captioners with names of participants well in advance if possible.
    b. During the event, all speakers should be introducing themselves (e.g. Mr Smith) each time before speaking to enable the captioners to identify speaker name.

3. Documentation: advance provision of the meeting agenda and if possible copies of presentations and any relevant information to the captioners.
4. Audio clarity: all speakers must speak directly to the microphones. If the captioners cannot hear, (inaudible) will be typed.
5. Quality of audio: Internet connection needs to be controlled by designated staff at the meeting venue, and organisers must ensure their equipment is in working order.

## 15 Remote captioning services for meetings

See [ITU-T FSTP-ACC-RemPart] and [ITU-T FSTP-AM].

# Annex A

## Examples of minimum requirements for quality of transcribed text session

This annex gives elements for consideration as minimum requirements based on the NER Model described in [Romero-Fresco]. More detailed requirements are for future study.

– Timely and accurate with maximum 3 seconds delay dependent on the Internet speed;

– Minimum speed of text output of 180 words per minute (WPM) (900 characters) to be able to follow speakers speed;

– 98 % accuracy of live text output;

– High quality of service (good audio quality, room acoustics, etc.);

– Strong Internet connection at both ends of the service;

– Organisers should check with service provider if the text streaming protocol is compatible with all devices currently used at the venue such as Laptops, mobiles, tablets;

– Displayed text should be easily read by the service users from a distance;

– Displayed text should have strong contrast. Use of sans-serif font is advisable;

– Displayed text should have as a minimum two lines of simultaneous text, to allow users to read captions (see Figures A-1 and A-2);

– Crowd-sourced caption correction might be considered to aid the accuracy of the transcript, if agreed with captioner.

Figures A-1 and A-2 contain examples of presenting captioning on a screen in meeting rooms.


**Figure A-1 – Captioning using same screen with sign language interpretation and speaker presentation**


**Figure A-2 – Captioning presented on a separate TV monitor**

# Annex B

## Questions to service providers for achieving minimum requirements of service

The following are questions that hiring parties of remote captioning services can ask service providers to facilitate achieving expected levels of service:

1. What is your experience in providing the type of service I need?
2. If something goes wrong during my event, who do I call?
3. What are the qualifications of your technical staff?
4. Where are your headquarters and offices located?
5. Please provide details regarding the technology that you use to provide your services. Example: hardware and software used to provide the service, storage and transmittal of information, and security measures/infrastructure that are used by your firm.
6. What is your backup protocol if services are requested but your staff is not available?
7. How many employees have certifications, and which type of certification?
8. Describe your understanding of and familiarity with terminology and our industry.

# Appendix I
# Examples of special keyboards

The following figures illustrate various types of equipment used by captioners.

Figure I-1 illustrates a steno keyboard. The lower image shows the general layout of a steno keyboard. It has been designed for use with English language and is widely used in the USA, UK and during English speaking meetings.

Figure I-2 illustrates a European languages keyboard. The lower image shows the general layout of the keyboard. The keyboard and its software are adaptable to different languages.

Figure I-3 illustrates the palantype keyboard. The lower image shows the general layout of the keyboard. This keyboard is typically used in UK.


**Figure I-1 – Steno keyboard**


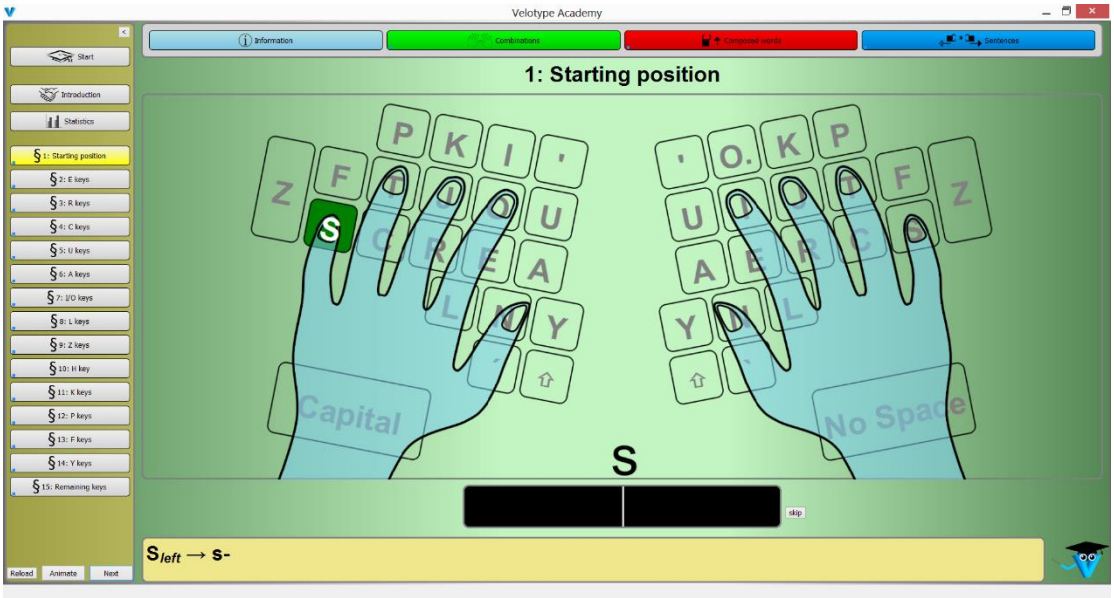


**Figure I-2 – Velotype**

**Figure I-3 – Palantype keyboard**

# Bibliography

[b-GDPR] Regulation (EU) 2016/679 of the European Parliament and of the Council on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), 27 April 2016. https://eur-lex.europa.eu/eli/reg/2016/679/oj