Universal Access in the Information Society (2024) 23:75–98
https://doi.org/10.1007/s10209-023-01048-0

LONG PAPER

Automatic captions on video calls: a must for the older adults

Eduardo Nacimiento-García[1,2] · Carina S. González-González[1,2] · Francisco L. Gutiérrez-Vela[3]

Accepted: 18 September 2023 / Published online: 13 October 2023

Abstract
In recent years, the use of video call or video conference tools has not stopped increasing, and especially due to the COVID-19 pandemic, the use of video calls increased in the educational and work spheres, but also in the family sphere, due to the risks of contagion in face-to-face meetings. Throughout the world, many older people are affected by hearing loss. Auditory functional diversity can make it difficult to enjoy video calls. Using automatic captions might help these people, but not all video calling tools offer this functionality, and some offer it in some languages. We developed an automatic conversation captioning tool using Automatic Speech Recognition and Speech to Text, using the free software tool Coqui STT. This automatic captioning tool is independent of the video call platform used and allows older adults or anyone with auditory functional diversity to enjoy video calls in a simple way. A transparent user interface was designed for our tool that overlays the video call window, and the tool allows us to easily change the text size, color, and background settings. It is also important to remember that many older people have visual functional diversity, so they could have problems reading the texts, thus it is important that each person can adapt the text to their needs. An analysis has been carried out that includes older people to analyze the benefits of the interface, as well as some configuration preferences, and a proposal to improve the way the text is displayed on the screen. Spanish and English were tested during the investigation, but the tool allows us to easily install dozens of new languages based on models trained for Coqui STT.

Keywords Automatic captions · Automatic subtitles · Speech to text · Videocalls · Older persons · Auditory functional diversity

1 Introduction

The use of videoconferencing systems or video calls is becoming more widespread and its use continues to grow [1], and it is estimated that it will continue to do so in the coming years [2]. Especially in these times marked by the COVID19 pandemic, the use of some information and communication technologies, such as videoconferences or video calls [3], has become an essential tool in the day-to-day life of many people throughout the world [4]. These tools are used in the workplace, education [5], leisure, etc. Due to the pandemic, many people use these video call tools for their interpersonal relationships between family and friends, and thus avoid face-to-face relationships that can spread infections among them [6, 7].

If we focus on the group of older people, many of whom are prone to loneliness [8], these types of tools are currently doing a very important job. When older people interact with videoconference platforms, there are several challenges they may face. These challenges can be grouped into the following categories: technical difficulties, visual and auditory impairments, cognitive challenges, lack of familiarity with the platform, and physical limitations. Older people may have difficulty using the technology needed to set up and join video conferences. They may also encounter technical problems with their devices or internet connection, making it challenging to participate. Visual and auditory impairments can make it hard for older people to see or hear what is happening on the video conference, even with assistive

✉ Eduardo Nacimiento-García
enacimie@ull.es

Carina S. González-González
cjgonza@ull.es

Francisco L. Gutiérrez-Vela
fgutierr@ugr.es

1 Universidad de La Laguna, Canarias, Spain

2 Instituto Universitario de Estudios de Las Mujeres, Universidad de La Laguna, Canarias, Spain

3 Universidad de Granada, Granada, Spain

technologies. This can lead to difficulty following the conversation and remembering names and topics. Cognitive challenges can also arise, such as difficulty multitasking or following the conversation. Additionally, lack of familiarity with the platform may make it hard to navigate, leading to confusion or frustration. Physical limitations can impact the ability of older people to use a computer or mobile device, making it hard to participate in the conference. For example, mobility issues or arthritis may make it difficult to use controls or type on a keyboard. In this work, we have taken into account that for people over 60 years of age [9, 10], there is a high number of people with auditory functional diversity [11], which means that they cannot enjoy all the advantages offered by video conferencing or video call systems. Obviously, this handicap also applies to any person with auditory functional diversity, regardless of their age.

From birth, a person continuously suffers deterioration of their hearing capacity, so older people with different levels of auditory functional diversity represent a very significant number of people in this age range. It is estimated that at least a third of people over 70 and half of those over 80 need to use a hearing aid to hear correctly [11].

On the one hand, video call systems offer advantages to older people, though on the other hand, there is a large number of older people who cannot use video calls correctly due to auditory functional diversity. At this point, a feasible solution to solve this problem may be the use of automatic subtitles.

Auditory functional diversity affects the quality of the perception of the multimedia content even if subtitles are used, and in some cases, the information of the multimedia content is not fully assimilated. The quality of perception includes the assimilation of the information present in the multimedia content, that is, both the visual part and the sound part. In addition, this quality of perception includes the level of enjoyment when interacting with that multimedia content [10]. One point in favor of the use of automatic subtitles in video calls could be that people are consuming more and more subtitled audiovisual content. Subtitled audiovisual content is consumed in many cases without the person having auditory functional diversity. Familiarity with subtitles makes people better able to keep up with the subtitles [12].

Currently we find that there are some video call tools that allow to activate subtitles, but not all of them have the possibility of activating automatic subtitles [5], such as Google Meet [13] or Zoom [14]. In some tools these live subtitles must be added manually, so in the case of the analysis we discard the tools with manual subtitling. We will only focus for now at the video call tools that offer the possibility of activating automatic subtitles using 'automatic speech recognition' techniques. Through automatic voice recognition, we convert speech into text and it is the system itself that automatically generates subtitles from the conversations during the video call.

When analyzing Google Meet and Zoom, we find some problems, on the one hand we have privacy problems, and on the other hand the subtitles offered by some of these tools do not meet the accessibility requirements for the older adults. Both platforms have several accessibility problems that can make it difficult for individuals with disabilities to participate fully in video conferences. They have inconsistent keyboard navigation, as some parts of their interfaces cannot be accessed with a keyboard, making it difficult for people with motor impairments to use the platform. Also, they have limited closed captioning support, although they offer automatic captioning, but it is not always accurate, and there is no way to request live captioning or provide your captions. Moreover, they have limited support for screen readers, some parts of the interface are not fully compatible with screen readers, making it difficult for users with visual impairments to navigate the platform.

It is important to have software adapted to the basic accessibility needs that the different people who use it may have, and in this case, in particular the older adults. As it is software that offers subtitling, restrictions related to different types of functional diversities must be taken into account, such as visual, reduced mobility or cognitive, apart from hearing [15].

In this way, it is necessary, for example, that from the perspective of the visual part, the size of the elements that appear on the screen, and their colors and contrasts, be taken into account. For blind people we would have problems with this type of software. From the perspective of reduced mobility, we must take into account that, for example, it is necessary to avoid the need to press more than one key or button simultaneously, it is also necessary to avoid the need to carry out actions in a specific period of time. Cognitive functional diversities can affect thought, memory, language, learning, perception, etc., but what is necessary is that the interface be as simple as possible, and that, for example, simply when opening the program this already works correctly, without the need for complex steps [15].

At this point we plan to develop an automatic captioning tool that is independent of the video call platform used and that also respects user privacy and is accessible to anyone, including older adults.

We also question if with a small modification when displaying the texts on the screen word by word, users could perceive the transcription as smoother and faster, than if we show the complete text fragment when transcribing it. It is important to clarify that we do not propose rolling subtitles, which move around the screen. In this case the words appear one by one, but in a non-mobile position.

For all these reasons, we developed an automatic subtitling tool using Coqui STT as a Speech to Text system, to generate automatic subtitles. In order to be able to evaluate the second question raised, the tool will be tested using two different models when displaying the text on the screen, in Model 1 the proposed "improvement" is used, where the text is displayed word by word, and in Model 2, the text is displayed directly at the time we have its transcription without post-processing, that is, the processed audio fragment will be shown fully transcribed. It is important to note that in Model 2, the size of the transcribed text will depend on the silences found during the conversation. These silences are what allow cuts to be made in the recording, to later be processed in the background while the conversation continues.

Section 2, presents some video call tools that allow automatic or manual subtitles, as well as some drawbacks related to them, and we focus on finding a free software-based video call tool, which we can use for our tests. In Sect. 3, we present our proposal for an automatic subtitling tool independent of the video call platform, analyzing its interface, the way the audio will be processed, the user's perception, the limitations of the prototype and the supported languages. Section 4 presents the validation of the proposals, while Sect. 5 the discussion. Finally, the conclusions are provided in Sect. 6.

2 Background

2.1 Related works

We have conducted a systematic review about the related works of this research to find other previous design proposals in the literature. A set of inclusion and exclusion criteria were defined in order to find the related works. To be precise, six inclusion criteria (IC) and the corresponding five exclusion criteria (EC) were described:

- IC1: the work deals with video conference tools AND automatic caption AND
- IC2: the work addresses issues related to accessibility AND
- IC3: the work is written in Spanish or English AND
- IC4: the work has been published in the last 5 years AND
- IC4: the work is published in journals, books or peer-reviewed conferences AND
- IC5: the work is available in open access or through the databases of the university AND

Concerning the exclusion criteria:

- EC1: the work does not deals with video conference tools OR automatic caption OR
- EC2: the work does not addresses issues related to accessibility OR
- EC3: the work is not written in Spanish or English OR
- EC4: the work is not published in journals, books or peer-reviewed conferences OR
- EC5: the work is not available in open access or through the databases of the university.

First and foremost, the database chosen according to the following requirements is Web of Science (WoS) due to it being a highlighted database in the field of research. The research terms were linked by boolean AND/OR operators, creating a search string used in the chosen databases. The base query employed was the following: ((TS = (videoconference)) OR TS = ("video conference systems") AND TS = (accessibility)).

A PRISMA flow has been used in order to explain the procedure followed in the data process. But first, we must point out that the data extraction process is an iterative and incremental technique that has been divided into multiple stages depending on the activity performed. In step one, the results obtained in the search string process were recorded and stored in a Google Sheet spreadsheet pointing out the title, author(s), abstract, keywords, publication year/date, source name, type of publication, academic database where it appears and DOI/URL of each paper. Step two consisted in reading each paper's abstract in order to accept or reject the work from the study. Now, candidates papers for full reading were downloaded and stored in a Google Drive folder. Once read, they were analyzed following the inclusion and exclusion criteria established. After carrying out this procedure, the results were displayed in a PRISMA flow (Fig. 1).

After applying the search strings, 2624 papers were collected from WoS. Once the inclusion and exclusion criteria were applied, 2616 papers were left. At the end, 8 research papers were selected to be analyzed.

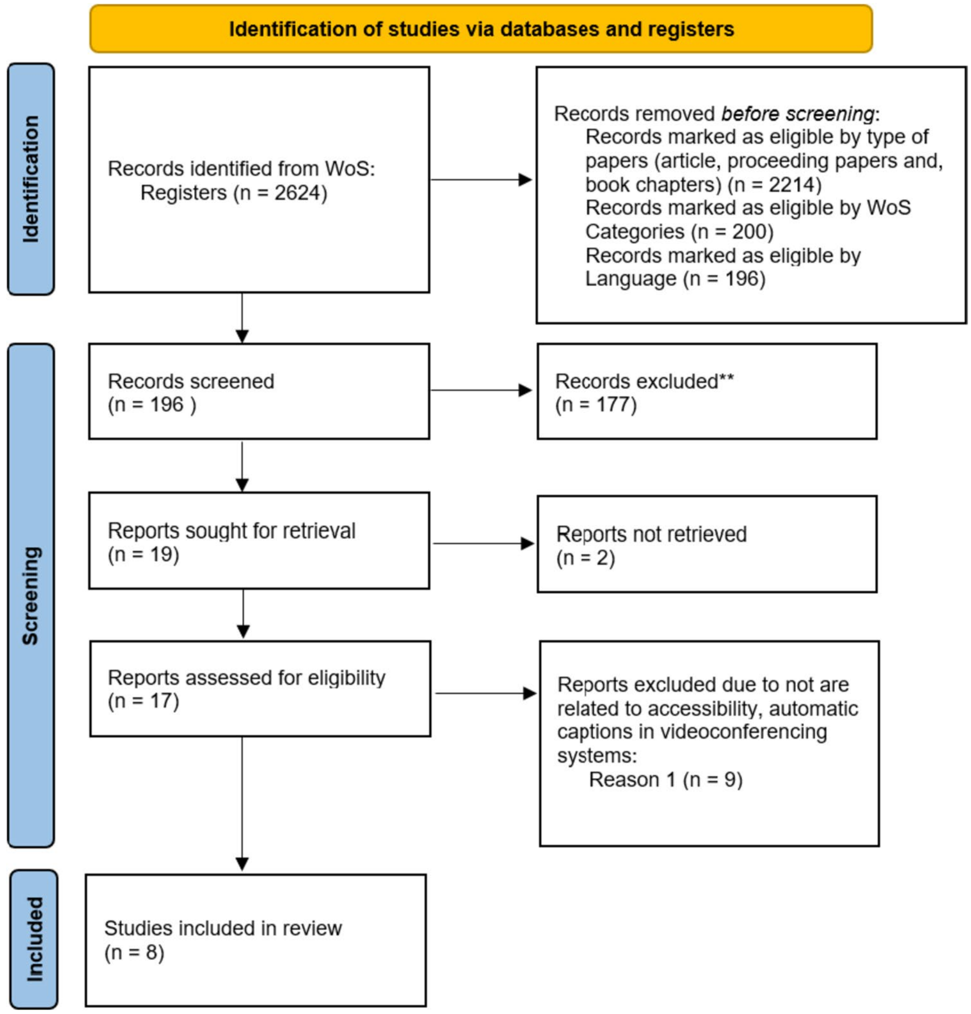
Table 1 illustrates the relationship among the reviewed works in terms of their accessibility requirements, audio caption features, and the key elements identified in the papers that are relevant to our proposal.

2.2 Video calling tools

Tools such as Google Meet or Zoom currently have automatic captioning options, but at the time of the research tests, Zoom did not yet have this option available [24], in addition to the fact that this option is not currently compatible with all versions. Automatic subtitling in Zoom is only available in English, with the possibility of manual subtitling [14].

By default, Google Meet and Zoom do not have automatic subtitles activated, but from settings one can activate automatic subtitles and in the case of Google Meet, it is possible to select the subtitle language among the available languages. The available languages in Google Meet are

Fig. 1 PRISMA flow of the SLR



English, Spanish (Mexico and Spain), French, German and Portuguese (Brazil) [13], though it is important to keep in mind that the availability of languages varies according to the region of the world from which it is accessed, since if it is accessed from Asia Pacific or European regions, you have access to those five languages (including the two variants of Spanish), but if you access from North America or Latin America, you only have access to English, German, Portuguese (Brazil), Spanish (Mexico). It is important to note that if the conversation is recorded, the subtitles are not recorded.

In the case of Google Meet, the settings only allow to enable or disable the subtitles, but they do not allow to change the properties of the text that will be displayed on the screen in the form of subtitles, such as the size or color of the text.

Many older people are also prone to visual impairment [25]. Because of this, many older people have difficulty or are unable to read subtitles correctly if a font that is too small to display the text on the screen is used. That is why it is necessary to take into account these accessibility needs, and for example allow changing the font size. The fact that we cannot modify the font size in Google Meet poses a problem for us from the point of view of accessibility for the older adults.

We consider that it is important to bear in mind that with the use of proprietary tools such as Google Meet, or others such as Zoom, we come across some problems derived from the use of proprietary software. Some of these problems are related to the privacy of the people who use that type of software. Many people worry about the loss of their privacy when using this type of tools, since with automatic captioning systems derived from automatic speech recognition, all the conversations of the participants are analyzed and transcribed by for-profit companies. All these conversations will be analyzed, without the user being able to know what will be the use that will be given to the information generated during the conversations. Also, with the privacy policies of most of these non-free tools, the videos can be used for face detection [26].

Among many other reasons, the right to privacy is one of the main reasons why many people decide to use free software tools instead of proprietary software tools [27].

Table 1 Related works

Authors	Accessibility requirements	Captions	Findings and elements relevant to our proposal
Neate, T; Kladouchou, V; Wilson, S; Shams, S [16]	The system is for people with Aphasia considering the difficulties in their verbal communication. Subtitles and gestures are proposed to improve the accessibility	Remark that the location on the screen of subtitles is crucial	Communication through body language and gestures is crucial, even in video conferences. This is an important aspect to consider when designing subtitles, as they need to be overlaid on the person speaking to avoid losing nonverbal and gestural information
Suduc, AM; Bizoi, M; Filip, FG [17]	–	Present the trends on using Automatic subtitling	Analyze the different video conferencing platforms and their usage statistics, which are steadily increasing, especially since the COVID-19 pandemic. Emphasize the challenges associated with nonverbal communication, as well as occasional issues with audio quality, among others. Highlight that there are platforms like Google Meet that offer automatic subtitling. One of the current trends in video conferencing platforms is the integration of AI, such as Speech to Text, which allows for the use of live automatic subtitles
Acosta-Vargas, P; Guana-Moya, J; Acosta-Vargas, G; Villegas-Ch, W; Salvador-Ullauri, L [18]	The paper focused on the WCAG accessibility remarking that sign language, subtitles and automatic transcriptions are key elements	Note that subtitles and automatic transcription are the weakness points in video conferencing platforms	The main video conferencing platforms are analyzed from the perspective of WCAG accessibility, and it is observed that the weak points, in terms of accessibility, in these platforms are related to sign language, subtitles, and automatic transcription
Seita, M; Andrew, S; Huenerfauth, M [19]	Focus on hearing difficulties and the importance of automatic subtitling in video calls	Note the importance of Automatic Speech Recognition (ASR) systems for generating subtitles	This article examines the importance of Automatic Speech Recognition (ASR) systems for generating subtitles during in-person conversations or video calls when one of the participants has hearing difficulties. It also emphasizes the significance of maintaining eye contact with the other person while speaking, which should influence the design of the automatic subtitling tool during video calls
Seita, M; Lee, S; Andrew, S; Shinohara, K; Huenerfauth, M [20]	The use of subtitles is essential to ensure accessibility for individuals with hearing impairments during video calls	Note the importance of Automatic Speech Recognition (ASR) systems for generating subtitles	The article discusses the significance of using ASR systems for generating automatic subtitles during video calls between hearing and hearing-impaired individuals, especially now with the increasing use of these tools for communication. It highlights the need for further research on user experience and user interface in this type of tool, as it is still limited. It proposes the strategy of co-designing these tools to better accommodate the needs of individuals

Table 1 (continued)

Authors	Accessibility requirements	Captions	Findings and elements relevant to our proposal
Arkhangorodsky, A; Chu, C; Fang, S; Huang, YQ; Jiang, DL; Nagesh, A; Zhang, BL; Knight, K [21]	The utilization of overlaid or modified subtitles creates a substantial cognitive burden in order to understand them properly	The system utilizes the retranslation strategy, occasionally resulting in flickering of the captions	The paper introduces a videoconferencing system that incorporates real-time translation captions displayed on the screen. The system aims to facilitate communication between individuals who speak different languages, thereby overcoming language barriers in multilingual settings. The current version of the system supports speech and captions in four languages and utilizes a combination of automatic speech recognition (ASR) and machine translation (MT) in a cascade fashion. To translate the speech in real-time, the system employs the retranslation strategy, resulting in occasional caption flicker. To enhance user experience and reduce cognitive load, the system incorporates features such as smooth scrolling captions and mitigates caption flicker. Its modular architecture allows for integration with various ASR and MT services in the backend
Spelter, B; Corsten, S; Diehlmann, L; Plath, A; Leinweber, J [22]	This article examines the requirements for developing video conferencing tools for individuals with aphasia	–	The paper highlights the need to implement a distraction-free screen and consider aspects such as font size and button design
Geislinger, R; Milde, B; Baumann, T; Biemann, C [23]	The system uses ASR technology to enhance accessibility for individuals with hearing difficulties or deafness	The paper presents a plugin for BBB for automatic subtitling	Considering the increasing demand for the use of open-source video conferencing platforms such as BigBlueButton, and recognizing the importance of having an automatic subtitling system using ASR technology to enhance accessibility for individuals with hearing difficulties or deafness, the development of an automatic subtitling module for BBB utilizing the open-source ASR system Kaldi is proposed

We have carried out a search for automatic subtitling tools that we can find useful, that allow automatic subtitling and that offer an interface that can be overlaid on the video call screen. The Ava tool [28] is the most similar system to the one proposed in this work. However after analyzing the tool, we have had to discard it because it does not seem appropriate for use by older people due to the difficulty of configuration and the lack of customization that would allow it to be adapted to the needs of each person. In addition, from the point of view of privacy it is less secure as it processes all conversations in the cloud and not locally. With Ava it is necessary to first register in order to use the app. There is a desktop application, but it is necessary to authenticate via the web to generate a token, which will then have to be pasted into the application and thus authenticated. The desktop version offers the possibility to display the text in a semi-transparent window on the video call screen, but it does not offer the possibility to adapt the characteristics of the text to the needs of the person. In the web version we can change the size of the text, but we could not enjoy the text overlay in the semi-transparent window. Ava offers the possibility of identifying the different people participating in the conversation, but for this, each of the participating people must authenticate and use Ava on their computer. Ava is designed for people with functional hearing diversity, but not for the older adults.

Currently we can find a wide variety of free software tools that allow to make video calls or video conferences, either configuring our own server or using a service offered by third parties. In this field, two of the main free software tools for video calls are Jitsi [29] and BigBlueButton [30], although with somewhat different main objectives between both applications, as we will see later.

We find developments in the testing phase to implement automatic subtitles using Speech to Text for these tools [31, 32], but it is not included by default in the officially distributed versions of these video call programs, so it is not available for use by any non-developer user.

Currently there are different free software tools for Speech to Text, which facilitate the work in terms of creating automatic subtitles for these platforms or free software tools for video calls. Among the free software tools available, we highlight Coqui STT [33], recently forked from Mozilla DeepSpeech [34, 35], by former developers who worked on this Mozilla project.

2.3 Free video calling tools

In this section we briefly review the free software tools for video calls Jitsi Meet and BigBlueButton, as they are two of the most used in this field [36].

Jitsi is a free software project licensed under the Apache 2.0 License, made up of a set of subprojects, also free [37]. Like many free software tools of this type, we can use Jitsi by installing it on our own server [38], but we can also make use of the Jitsi Meet service that the developers of the project offer for free at https://meet.jit.si.

Jitsi is designed with video calls or video conferencing in mind, so it has a fairly simple interface, which may be adequate for our use case [39]. In addition, we can participate directly in a video call by accessing a URL from a web browser, without the need to install any client on our computer or mobile device, although if we want, we also have a desktop application for a computer and a mobile application.

BigBlueButton [40] is a free software video conferencing tool, licensed under the GNU Lesser General Public License 3.0. An important difference between Jitsi and BigBlueButton is that BigBlueButton, in addition to being a videoconferencing system, is designed to be used as an online learning system [41]. Due to this, BibBlueButton includes some basic functions focused on this objective, such as digital whiteboards or the fact that it is possible to integrate it with the main Learning Management Systems (LMS) such as Moodle [42].

As in the case of Jitsi, we can install BigBlueButton [38] on our servers, personal computers or use the demo available on the project's servers completely free of charge.

For our research we finally need to choose one of the two tools to be able to carry out the test sessions with the volunteers. Let us remember that our application is independent of the video call platform used, so we are really interested in using a video call tool that is easy to use, so that the difficulty of using that video call tool does not influence the results that we can obtain on the complexity of using our automatic captioning tool.

Although both projects are free software, both allow us to make video calls, and our automatic captioning tool is independent of the video call platform, we chose to use Jitsi Meet for the sessions with the volunteers. By choosing Jstsi as our video call tool, we try to prevent the use of a more complex video call tool or one with additional features from interfering with the results of the automatic captioning tool that we have developed and are testing. For this investigation and to simplify, we are not considering the possibility of installing our own server with the video call tool, but we will use the free service in the cloud.

3 Proposal

One option would have been to propose the development of plugins or add-ons for each of the free video call tools, and in that case, although much of the code of our add-ons could be reused for one video call or another, we would be forced to maintain these add-ons continuously updated based on the updates that could be made in each of the video call tools.

The basic idea of our proposal is to create a tool that allows to have automatic subtitles when we make a video call, regardless of the video call software that we use for it, and allow our tool to be superimposed on the video call window. In addition, this automatic subtitling tool must meet some basic accessibility requirements, bearing in mind that the motivation for carrying out this project is that older people can enjoy video calls. For this reason it is important to bear in mind that many elderly people have hearing problems, but also vision problems, so it is necessary to allow options that adapt the text to personal needs. Another consideration to take into account is that the interface used must be simple so that it can be used by people without much computer knowledge, as is the case with most older people.

To carry out the work sessions with the volunteers and check if the proposal is adequate, a fully functional prototype was made, whose graphic interface is developed in pyQt5 [43] and the automatic translation is carried out using Coqui STT [33].

3.1 The interface

The graphical interface was designed with the idea of being as non-invasive as possible, that is, the interface should not bother the user during a video call. Basically, the interface consists of a transparent and "always visible" window superimposed on the other windows, but which in turn allows us to make the necessary adjustments in a simple way.

When we position the mouse cursor over the transparent window area, a settings button appears. If we click on the settings button, other buttons are displayed that allow us to modify the essential settings for a tool of this type. Through these buttons we can easily and intuitively adjust the options related to the text that is shown in the subtitles, such as the font size, the font color, the background color of the text.

Despite being a transparent window without a status or title bar, we have the possibility of moving it from position simply by clicking and moving the mouse without releasing it. We also have the options to maximize and restore the area of the transparent window.

If we have more than one language model installed for our tool, we can easily change it from the menu itself, simply by clicking on the language options and a list of available languages is displayed. By default, the tool does not incorporate any pre-installed language model, and for this reason, when we run the tool without available languages, it shows an advanced options window from where we can download the language we want before starting. In the advanced options screen, which is accessible by means of a button, we can manage the different languages in our tool, being able to install or uninstall the different language models with just one click. In addition, on this screen we can see information about the different languages installed and available, as well as access help.

Another important adjustment option offered by the tool is the possibility of changing the source from which the audio is captured, that is, we can switch between the system sound or the microphone sound. By default, the option to capture the system sound is activated, that is, the sound that is played and heard through the speakers or headphones, because in the case of a video call, this is the main option. We also have the option, with a simple click, to change the capture source to the microphone, so that we can use this tool with other features, such as during a presentation where what the speaker speaks is automatically subtitled. Also, it is important to keep in mind that by using the system sound as the audio source that we want to process, it means that we will be able to subtitle any audio source that we play on our computer, such as podcasts, videos, etc.

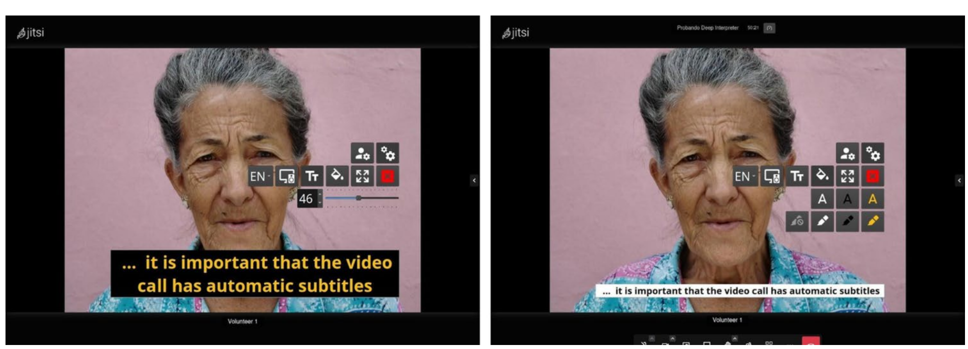
To make subtitles easier to read on screen, the automatic subtitles tool sets a maximum of 70 characters per line, as this value is considered in some research to be the recommended maximum [44, 45]. This maximum value of characters per line only applies when the user uses a small font size, because the text cannot exceed the size of the window. Because of this, when the font size is increased, the text overflows onto another line and the number of characters per line is less. The user can easily change the size of the font and the window, which allows the user to adapt the displayed text to the needs or tastes of it. To increase the font size the user only needs to use a slider.

In the Google Meet tool, live captions are added in a dark frame outside the video call area. With the use of this technique, the perception of all audiovisual information is damaged, since it is necessary to look away from the image to read the subtitles [46]. In general, often, in video player programs or online services such as YouTube, the subtitles are displayed on top of the video. With the transparent interface that we propose in our software, we do not have to worry about the problems related to using text inside or outside the image. With this tool we can move the window to the place we want on the screen, and thus the user decides which is the way that best suits their tastes and needs when reading the subtitles.

In Figs. 2 and 3, we see some screenshots where we can appreciate the simple interface of the tool. Figure 4 shows screenshots of how to manage installed languages or add new ones.

Fig. 2 Screenshots using Jitsi Meet and the automatic captioning tool. **a** When the cursor is not over the window. **b** When the cursor is over the window. **c** Settings options after clicking the settings button

Fig. 3 Text configuration options. **a** Modification of font size. **b** Modification of the color of the text and the background color of the text



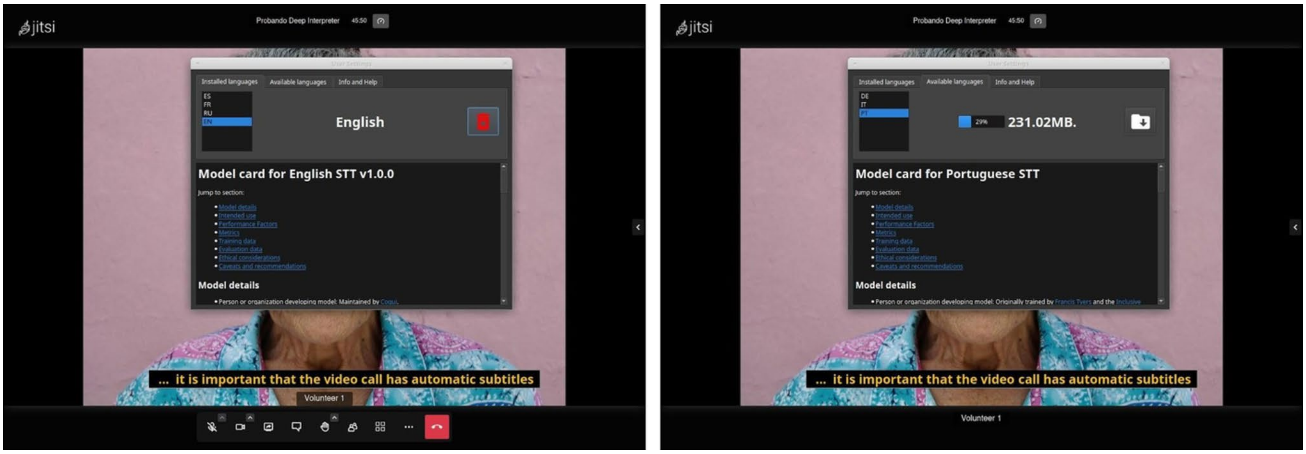


Fig. 4 Simple interface to install or uninstall languages. **a** Selected English language. **b** Downloading and installing the Portuguese language

3.2 Processing the audio

For the proper functioning of our tool, it is necessary to process the audio that we capture, either from the system or from the microphone. Bearing in mind that we need the subtitling to appear on the screen as the conversation takes place, and with as little delay as possible, we need to establish some way of being able to process the audio in fragments, and automatically.

The procedure to be able to process the audio is based on the fact that we will fragment the input audio based on cutting the audio with respect to the silences that are produced. It is important at this point to work in two separate threads: in the first thread we are making the audio cuts based on the silences, and in the second thread we will be processing each audio fragment through the automatic speech recognition provided by Coqui STT. It is necessary for these two processes to work in parallel and continuously, on the one hand to avoid unnecessary delays and on the other hand to avoid losses due to audio that was not processed.

At this point, we come across an important detail when analyzing a conversation in real time: it is necessary to properly adjust the parameters related to silence, that is, how much silence implies a cut in the audio. This is important, on the one hand so that the audio fragments that are analyzed are not too small, which could divide words or make it difficult for the STT system to understand them, and on the other hand, so that the audio fragments are not are too large. The latter, causes a large gap between the moment in which the conversation takes place and the moment in which said conversation appears in the subtitles, since it must be taken into account that no matter how fast the STT process could be, if one has to wait a long time to start processing a conversation, this will be perceived by the user as a bad automatic captioning system in real time.

In the case of CoquiSTT, we have to consider that the level of precision or error when transcribing a conversation will depend on the model used for it. In CoquiSTT we can use a multitude of pre-trained models [47] and each of them will have a different rate of successes and errors. Some possible measurements to take into account are the WER (Word Error Rate) and the CER (Character Error Rate) [48]. In the case of the model used in English, we have a WER of 4.5% and a CER of 1.6%. In the case of the model used in Spanish, we have a WER of 16.5% and a CER of 7.6% [49, 50].

The cutting of the audio fragments is done based on the detection of silences. This is a simple way to detect the separation between words, in a linear process, without the need to continuously reprocess the data.

If a fixed processing time were set, in all probability at some point we would cut words, thus causing an erroneous transcription. This could probably be improved in post-processing, but we are interested in a fast translation and that it does not have to be corrected later, even after being displayed on the screen. These subsequent corrections are made by some automatic captioning systems, such as the one available in Google Meet [13].

In our case, when using silence-based cuts, it is necessary to establish what would be considered a silence, that is, how many milliseconds there must be without capturing voice so that the system interprets it as silence and makes a cut in the audio. If the silence time is too small, it may happen that the system detects silence within the same word, and if the silence time is too long, the audio fragments to be processed will also be very large, which would require process a large amount of dialogue at once, thus causing the transcription to take too long, and may even reach the point where it is not cut off until the speech is completely finished.

We carry out some tests to establish an adequate silence value, starting from some studies that establish this threshold between 200 and 1000 ms, and that in some cases it is established directly at 300 ms [51, 52] or 400 ms [53]. In our case, after performing the tests with intervals between 100 and 1000 ms, we found that the number of erroneous words from 300 ms onwards was very low and practically equivalent to the WER and CER parameters of the system, but instead, for values lower than 300 ms these values increased. The value of silence could have been extended to more than 300 ms to ensure greater precision in all cases, but this would cause a greater delay in the transcription, and the intention is to make it as live as possible.

3.3 User perception

As a general rule, the time it takes to process an audio fragment with Coqui STT is less than the time that audio fragment takes. This makes it easier for us to use Coqui STT for live audio transcription without too much lag, that is, without too much time passing between speaking, the audio being processed and the text being displayed on the screen. Obviously, the silence parameters established in the previous point would have an influence here so that this delay does not increase unnecessarily.

While the audio snippet analysis is fast, the user's perception of the transcription speed might not be as good, if the audio snippets being analyzed are too large. That the analyzed fragments are too large could happen because the user speaks very quickly and without pauses, or because we have not properly adjusted the parameters related to the silence time necessary to make the cuts in the audio. If the silence parameters are poorly adjusted, we could also have very small fragments, which could become difficult to understand and transcribe. In our case, using a time of 300 ms, we consider that we obtained very good results from both perspectives, that is, with an acceptable response time, but without perceptibly affecting the quality of the transcription.

At this point, it is also proposed that the perception of the transcription of the subtitles improves if a small modification is applied in the way the transcribed text is presented on the screen.

By default, if we carry out the subtitling process as we have seen so far, we would have to cut a fragment, then that fragment is processed by Coqui STT and finally the transcribed text is displayed in the subtitles. The proposed modification is that, once an audio fragment has been analyzed and the text corresponding to the transcription has been generated, instead of writing that string of text directly on the screen in its entirety, that text will be divided word by word, so that that these words will be written on the screen one by one and adding a small pause between each one, for example 0.1 s.

It is estimated that the standard reading speed of subtitles is about 120 words per minute, which would be equivalent to 0.5 s per word [54], if we simplify, because obviously not all words are of the same size. The reading speed of a person with reading habits can be around 300 words per minute, that is, the person will take about 0.2 s per word. It is considered difficult to exceed 600 words per minute or 0.1 s per word, if you want to understand what you read [55]. Based on this data, the value of 0.1 s was chosen, which means that the words are displayed on the screen close to the reading speed, but without unnecessary waiting.

The main idea behind this small modification when displaying the text on the screen is that in this way the user's perception is that the transcription is being more fluid than if the text is displayed completely at once. However, the reality is that it is really being subtitled more slowly, because more time passes between when a word is pronounced and when it is written in the subtitles. Obviously, to avoid unnecessary waiting, the first word of a string is not delayed.

The true fluency effect occurs because while one analyzed audio fragment is printed word by word on the screen, the next audio fragment is analyzed in parallel, to generate a new transcribed text. By displaying the transcribed texts on the screen, the time that elapses without writing on the screen between one text of one processed audio fragment and the next is considerably reduced. Therefore, we can say that in this way, the subtitles produce an effect of greater fluidity in the perception of the text transcribed on the screen. In addition, based on the data shown above regarding the reading speed, it is considered that with this improvement it may be easier to read the subtitles. This improvement is due to the fact that the words appear at a rate close to the common reading rate, although somewhat faster to avoid unnecessary waiting for the user to read the subtitles.

3.4 A limitation of the prototype

When using a subtitle generator that is independent of the video call or video conference platform, such as the one presented in this article, we must take into account a clear inconvenience with respect to platforms that directly incorporate subtitling, or if it is done through a plugin or complement. The drawback is that we cannot access information that the video call tool could easily provide us. For example, we could access the identification of the different people participating in the video call. By not being able to differentiate between the different people participating in the conversation, the transcript of the conversation can be more difficult to understand for a person with total deafness, in the event that more than two people participate in the video call.

In this prototype, only the perception of automatically generated subtitles in conversations between two people is analyzed. In addition, only the other person's conversation is being transcribed, from the point of view of the person running the automatic translator. This makes it easier to understand the subtitles, in addition to the fact that it seems illogical to double our own conversation in a video call, although in this case it would be easier for us to differentiate between the voice that enters through the user's microphone and the voice of the other user reproduced by the system.

We focus only on the transcription of the conversation between two people, because the tool is designed to be as easy as possible to use during a video call. It is intended that the user is not required to register on any online platform together with the rest of the participants for the system to identify them, as is the case with Ava [28]. Remember that it is a subtitling tool independent of the video call tool that works offline and does not send data to the cloud to be processed, in this way we protect the user's privacy.

To enhance this automatic subtitling tool and make it suitable for video conferences involving more than two people, we can explore various tools that enable the detection of different individuals or voices. Some examples of free tools that would allow us to work to add this functionality could be Resemblyzer [56] or pyAudioAnalysis [57].

3.5 Supported languages

Another very important issue to take into account is the ability of our automatic subtitling tool to support different languages. In our case, for a language to work correctly in the Automatic Speech Recognition and Speech to Text process, we will use models trained to recognize speech in each language.

In a previous section, we could see that, for example, the Google Meet video call service works in up to five different languages, depending on the geographical region where we are. In our case, by using Coqui STT, and this being a fork of Mozilla DeepSpeech, we can use any language that is trained and available for any of these systems.

Coqui STT, like Mozilla DeepSpeech, makes use of Mozilla's Common Voice corpus [48], which is constantly growing thanks to the contributions of the great community behind it, more than 200,000 voices in version 8.0 of Common Voice corpus. At the time of writing, 93 languages are published, that is, they are fully localized and have enough compiled phrases [58]. In addition, in the process of creation, with different levels of progress, we find another 68 languages. Apart from Common Voice, it is common to also use content available in LibriVox [59, 60] to train the models, among other data sources.

If we so desire, we have the possibility to train models for our Speech to Text system, using each of the languages available in the Mozilla Common Voice project as a base, as long as that language has enough voice contributions. But, if we want, we can also make use of the already trained systems, which are available for download and use directly by our system based on Coqui STT. On the Coqui project website, we have a list with 85 pre-trained models from 51 different languages [47], that we can download and use in our automatic subtitling tool project, since they are licensed under free licenses. These models available in 51 different languages, include the 5 offered by Google Meet.

It is important to keep in mind that the quality of the dialogue transcription will depend on the model used for it. As we saw earlier, in the case of the model used for English we have a WER of 4.5% and a CER of 1.6% [49]. In the case of the model used for Spanish, we have a WER of 16.5% and a CER of 7.6% [50].

In our automatic subtitling software, it is very easy to add a new language, and we have several options for it. The easiest and most recommended way is by accessing the advanced settings of the application, where we will be shown a window with several tabs, where in the first we will see the installed languages, and in the second tab we will see all the languages available for download and installation automatically, and which are currently the different languages shown on the Coqui project page that we saw in the previous paragraph. In our case, in order not to add repeated languages, we have only added one model per language, although on the Coqui project website for some languages we can find more than one pre-trained model, we have chosen the model that we consider best in each case. If we want to add a new language that is not available in the list of languages available for download, we can also install it easily: we only have to create a directory with the language code inside the "langs" directory and place there the files of the pre-trained model. For example, to add a Spanish language model, we create a directory called "es" inside the "langs" directory, with the content of the module downloaded or trained by us. This content will have a ".tflite" file with the model, and probably also another ".scorer", which are the files that the STT system will use to transcribe the speech.

When we add a new language, it automatically detects it, and from the configuration options that we see in Fig. 5, we can easily change it. The figure shows "EN", because in that case we were using a model for the English language.

4 Validation

4.1 Participants

During the validation process, we had to deal with the restrictions that existed at that time in the Canary Islands

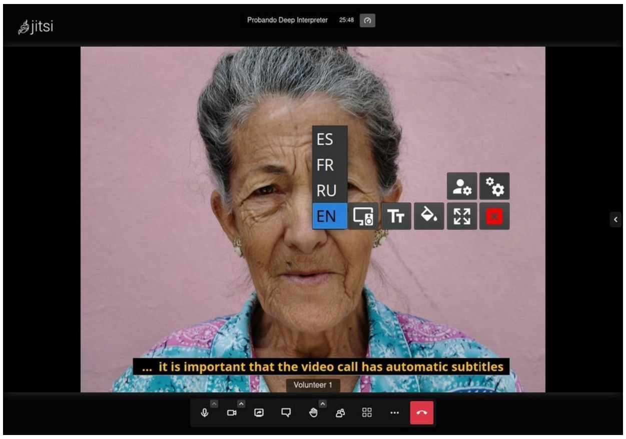


Fig. 5 Installed language selector

[61] due to the COVID-19 pandemic, so we were forced to reduce the verification tests that were carried out, especially with the tests carried out with the older adults.

The users participating in the study were chosen so that there was relative parity in terms of gender, and that there were people over 60 years of age as volunteers, even though the population sample was intergenerational.

The research and the participation of the volunteers were carried out in compliance with the Declaration of Helsinki, with informed consent at all times and with the utmost respect for the participants. In addition to this, taking into account the special circumstances derived from the pandemic, the experiments were carried out with the maximum guarantees to avoid contagion [6], using masks, hydroalcoholic gel and maintaining the greatest possible safety distances.

The validation process was carried out in two stages. In the first validation stage, eleven face-to-face sessions were held, taking into account the established restrictions and the appropriate sanitary measures. In these eleven sessions, the operation of our automatic subtitling tool was tested and the perception of the volunteers regarding the two proposed models was also measured.

Let us remember that the differences between the two proposed models consist basically in the fact that in the first model, the text resulting from the transcription of the conversation is post-processed and divided into words, so that it is then displayed on the screen word for word, with a pause between 0.1 s words. In the second model, the transcribed text is displayed on the screen without the need for post-processing to divide by words, so that the text is displayed as soon as it is available.

The automatic captioning tool was tested by eleven volunteers of different ages and genders. Of the participants, three users were over 60 years of age and two were over 70. Of these three people, two have very basic computer skills, and the third has almost no knowledge of the use of new technologies.

Due to the lack of computer knowledge on the part of the older adults volunteers who participated in the research, it was decided to carry out the tests in person, so that we can correctly evaluate the use of the automatic subtitling tool, and that the results were not affected by other problems derived from little or no computer knowledge. These problems derived from third parties could have been especially complex if we had decided to carry out the entire process remotely. In relation to all this, it is important to point out that many older people have few computer skills [62]. That is why the interface of our tool is as simple and intuitive as possible, but to evaluate it correctly we cannot do it adequately in a way that we do not control.

The group of people participating in the research speak Spanish as their mother tongue. Of all the participants, only two (under 60 years old) said they had a lot of previous experience in the use of subtitles, the rest of the participants said that their experience with subtitles is very punctual. For example, when a scene from a movie is subtitled on television, interviews on the news, among other resources. Three participants, the three users over 60 years of age, consider that they have Mild hearing loss or Moderate hearing loss [63]. All of them have visual problems.

The design of the experiment consisted of setting up two computers in two different rooms, where the user had to have a video call with one of the researchers using Jitsi Meet and our automatic captioning tool. The computers had the Xubuntu GNU/Linux operating system installed, and the Mozilla Firefox web browser, where the video call link was accessed.

The test lasted approximately 30 min in total, and the two models presented were tested, so that simple conversations were held between the user and the researcher. The same researcher participated in all sessions to ensure greater homogeneity between all sessions. During the work sessions, Model 1 was used for about 10 min, then Model 2 for another 10 min. After these tests, users were again allowed to evaluate Model 1 and Model 2, so that they could evaluate better what was their perception regarding each model.

Each work session was carried out independently on different days to prevent fatigue from affecting the users or the researcher.

After making modifications to the initial prototype, based on the results obtained in the first phase, a second work session was held again with the volunteers. On this occasion, 9 of the 11 people who participated in the first evaluation were present in this new test, including the 3 people over 60 years of age. The session was carried out in the same way as on the first occasion, but only using transcription Model 1, but with the two variants of the interface, on the one hand the interface of the first prototype, and on the other hand, the new interface of the second prototype with the accessibility improvements implemented.

In the case of the second test, the methodology used is similar to the one used during the first experiment previously exposed.

4.2 Methods and instruments

The entire evaluation process in this first phase went smoothly. Then, once the tests were finished, each of the volunteers was given the corresponding evaluation questionnaire. To evaluate the usability of the automatic subtitling tool, the System Usability Scale (SUS) [64] was used, with a Likert-type scale (1–5) [65]. In this part we evaluate the tool in a generic way, without questioning the perception of fluidity and speed of one model or another.

The items used in the System Usability Scale (SUS) were the following:

- I think I would use this app;
- I find this application unnecessarily complex;
- I think the application was easy to use;
- I think you would need help from a person with technical knowledge to use this app;
- The functions of this application are well integrated;
- I think the application is very inconsistent;
- I imagine most people would learn to use this application very quickly;
- I found the app very difficult to use;
- I feel confident using this app;
- I needed to learn many things before being able to use this application.

In addition to the ten items of the System Usability Scale (SUS), the volunteers participating in the research answered two statements related to the two proposed transcription models. Were intended to be able to identify the perception of these people about the use of Model 1 and Model 2, which we defined earlier. In these two items, as well as in the previous ones, the Likert scale (1–5) was used. The two statements were the following:

- The first mode of the application runs smoother than the second mode;
- The first mode of the application transcribes faster than the second.

We chose these two statements about the perception of greater fluidity or speed of one model compared to the other, because in this case what interests us is precisely knowing the user's perception. Although we know that the new proposed model is actually clearly slower by introducing artificial pauses.

The parameters used in the Likert scale in the questionnaires were the following:

- Strongly disagree;
- Disagree;
- Neutral;
- Agree;
- Strongly agree.

However, before being able to evaluate the answers obtained, it is necessary that we first normalize the data using the following rule [64]:

- Odd numbers are subtracted by 1;
- The result of the even answers is subtracted from 5;
- Add all responses once converted according to the steps above. The result of the sum is multiplied by 2.5.

In order to correctly evaluate the results obtained through the System Usability Scale (SUS), we must take into account the data shown in Table 2 [64], since when using this system, it is considered that a result with a value of 60 is approximately equivalent to a 50% valuation, and that in the event that we obtain a result greater than 80.3, that valuation is greater than 90% [66].

With the two extra questions, we have decided to evaluate the user's perception regarding the differences between Model 1 and Model 2. We had proposed that by adding the modification of Model 1 with respect to Model 2, an improvement is added when it comes to display the transcribed text on the screen, although we are really adding some delay when displaying the text on the screen, because we are dividing the text into words and adding a pause of 0.1 between each word.

Currently, for the development of a second prototype, we have already made some improvements related to the accessibility of subtitles, especially focused on the WCAG recommendations for captions [67], and especially for captions in real time [68]. In this sense, according to these specifications, two basic design recommendations should

Table 2 Interpreting SUS score

SUS Score	Grade	Adjective Rating
>80.3	A	Excellent
68–80.3	B	Good
68	C	Okay
51–68	D	Poor
<51	F	Awful

be fulfilled: the first one establishes that it is necessary to ensure that there is sufficient color contrast between the text of the subtitles and the background, and the second recommendation establishes that the Subtitles can be resized without loss of readability. In our first prototype we already had the option to change the color and size of the letters, but to make sure that we faithfully adhere to these recommendations, we added the option for the background color of the text, which until now was always transparent. Because the background of the video call may change, a transparent text background may make it difficult to read subtitles clearly at certain times.

In the subtitle configuration options the user can choose between a yellow, black or white text color, and regarding the background color of the text, the user can choose between yellow, black, white or transparent colors. These three colors are the ones that are normally used in subtitles.

Also, some modifications were made to the settings icons, using a dark palette with white icons, which generates a better contrast with the background of the screen.

Another of the points on which we had to work according to the data obtained in the questionnaires, was to facilitate the use and configuration. This was observed at the time of having to install new languages, since in the first prototype we did not have a language manager, and the way to install new languages was by adding the pre-trained model files to the corresponding directory within the "langs" directory. Although this was a simple task, for people with little computer skills, it was a complicated task. For all these reasons, a language manager was developed that allows us to install or uninstall any of the 51 languages available for Coqui STT with a simple click.

In the end, all the volunteers who participated in the second work session, using the new prototype, were given a questionnaire with the following questions, with the same Likert Scale used throughout the investigation:

- Do you consider that it improves the ease of use of the new prototype compared to the previous one?
- Do you consider that the new icons and the dark interface improve the accessibility of the prototype?
- Do you think that the subtitles are better readable in the new prototype than in the previous one by being able to add a background to the text?
- In general, do you consider that the new prototype is better than the previous one?

4.3 Observation

To evaluate the tool, video calls were made between the volunteers who were part of the study and one of the researchers. All the tests carried out were through video calls between two people. The video calls were made mainly using the Spanish language, although in one of the cases, in the middle of the conversation they switched to the English language, continuing the video call with the automatic subtitling in English instead of in Spanish. Because in our tool it is easy to change from one language to another, and because the model is available in English, it was very easy for the user to change the language from Spanish to English and continue the conversation naturally in another language.

4.4 Results

In Fig. 6 we can see the results obtained with respect to the questions of the System Usability Scale (SUS), the responses of each person to each of the 10 SUS questions are shown. In Fig. 7 we can see the means and medians of the responses for each of the SUS questions.

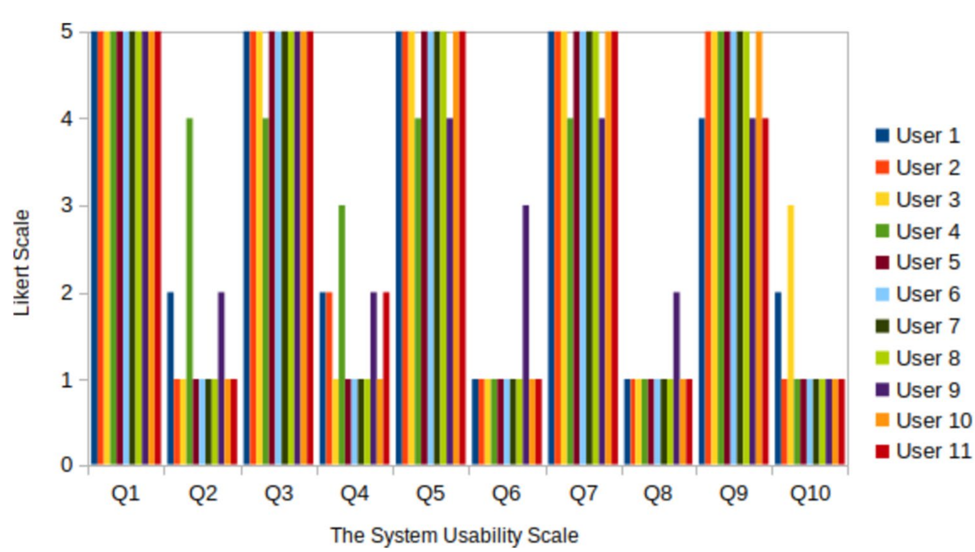


Fig. 6 System usability scale results: results by question and user

Fig. 7 System usability scale results: mean and median by question

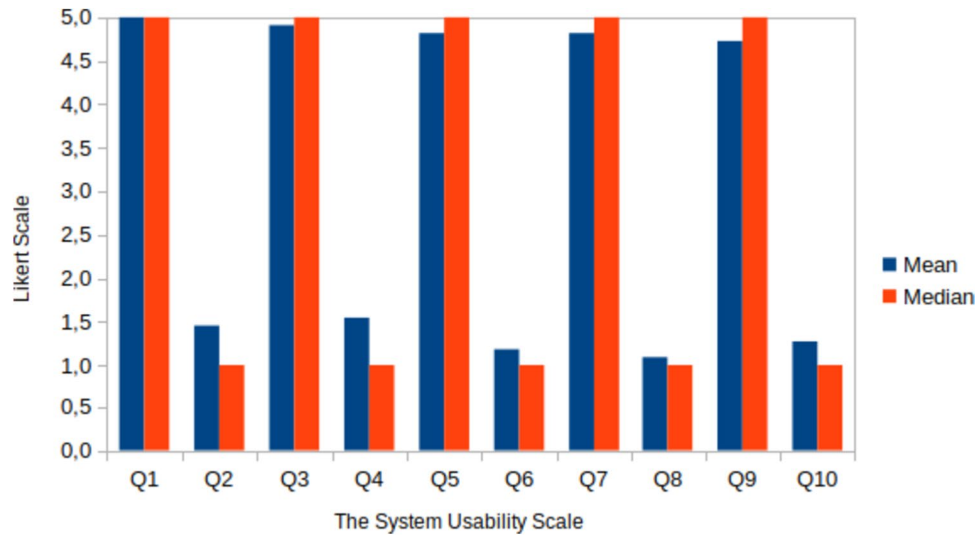


Table 3 SUS responses

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
User1	5	2	5	2	5	1	5	1	4	2
User2	5	1	5	2	5	1	5	1	5	1
User3	5	1	5	1	5	1	5	1	5	1
User4	5	4	4	3	4	1	4	1	5	3
User5	5	1	5	1	5	1	5	1	5	1
User6	5	1	5	1	5	1	5	1	5	1
User7	5	1	5	1	5	1	5	1	5	1
User8	5	1	5	1	5	1	5	1	5	1
User9	5	2	5	2	4	3	4	2	4	1
User10	5	1	5	1	5	1	5	1	5	1
User11	5	1	5	2	5	1	5	1	4	1

Table 4 SUS results

	Normalized results
User 1	90
User 2	97.5
User 3	100
User 4	75
User 5	100
User 6	100
User 7	100
User 8	100
User 9	80
User 10	100
User 11	95
Mean	**94.32**

In Table 3 we can see the data obtained after processing the System Usability Scale (SUS) questionnaires. After calculating the scores obtained, we have the following values, which we can see in Table 4.

Next, in Tables 5 and 6, we can see the responses of people over 60 years of age to the System Usability Scale (SUS), as well as those normalized results and their interpretation.

After processing the questionnaires, in the answers related to the differences in perception between the two models, we have that the results obtained for these questions can be seen normalized in Table 7 in general, and for people over 60 years in Table 8. Similarly, in Figs. 8 and 9, we can see these results represented for these two questions.

The answers to the questionnaires, in relation to the accessibility improvements and ease of use proposed in the new prototype, can be seen in Table 9.

We also see in Fig. 10 the score obtained for each of the questions, while in Fig. 11 we can see the results of calculating the mean and median of the results obtained.

Table 5 SUS responses > 60 years old

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
User1	5	2	5	2	5	1	5	1	4	2
User2	5	1	5	2	5	1	5	1	5	1
User3	5	1	5	1	5	1	5	1	5	1

Table 6 SUS results > 60 years old

	Normalized results
User 1	90
User 2	97.5
User 3	100
Mean	**95.83**

Table 7 Two models—responses

	Q1	Q2	Q1 normalized	Q2 normalized
User1	5	5	100	100
User2	5	4	100	75
User3	5	5	100	100
User4	4	4	75	75
User5	5	5	100	100
User6	5	5	100	100
User7	5	5	100	100
User8	5	5	100	100
User9	2	3	25	50
User10	5	5	100	100
User11	5	5	100	100
Mean			86.36	90.91

Table 8 Two models—responses > 60 years old

	Q1	Q2	Q1 normalized	Q2 normalized
User1	5	5	100	100
User2	5	4	100	75
User3	5	5	100	100
Mean			100	91.67

5 Discussion

As we can see, the general results obtained from the first phase of our study after normalizing and evaluating the SUS results, are 94.32 (Table 4), which is much higher than 80.3 because the assessment is much higher than 90%. Based on these data, we can consider that very good results were obtained, although it will be necessary to evaluate the weak points of the tool from this point of view.

After analyzing the results obtained in the questionnaires related to the two transcription models, we find that, in all cases, Model 1 behaves more fluently than Model 2. We obtained “totally agree” in all responses. In the case of the question of whether Model 1 is faster at transcribing the conversation than Model 2, the answers obtained were not unanimous. The results obtained reflect a rating of 93.33, so this rating is also very high.

In general, in relation to the new prototype, very good results were obtained for all the answers, especially to the question about whether, do you consider that the subtitles are better read in the new prototype than in the previous one, since it is possible to add a background to the text?, where all the people valued the maximum score for this answer, obtaining a 100% assessment. If we review each of the other questions, we see that we have also obtained very good results and with a score above 90% in all cases. In particular, we obtained a score of 93.33% for questions 2 and 4, where we asked, respectively, if you consider that the new icons and the dark interface improve the accessibility of the prototype? And if, in general, do you consider that the new prototype is better than the previous one? To the question about whether, do you consider that the ease of use of the new prototype improves compared to the previous one?, a rating of 91.11% was obtained.

In Fig. 12 we can observe some of the modifications made as a result of the evaluations made in the first and second phases of the investigation. It is important to keep in mind that all the other screenshots seen in the other figures belong to the new prototype with the accessibility modifications made.

If we disaggregate the data analyzed above into two groups, depending on whether the volunteers are over or under 60 years of age, and we analyze the results obtained from people over 60 years of age, the result obtained is better than in the general case. In the case of people over 60 years of age, the result of calculating the responses to the System Usability Scale (SUS) is 95.83. Regarding the questions about the two models, Model 1 is perceived as more fluid than Model 2 with a score of 100, and this score is 92.73 with respect to speed. As we can see, the evaluations regarding fluency and speed are also higher in the case of people over 60 years of age than in the general case. When analyzing the data related to the interface and simplicity improvements proposed in the new prototype, we find that in

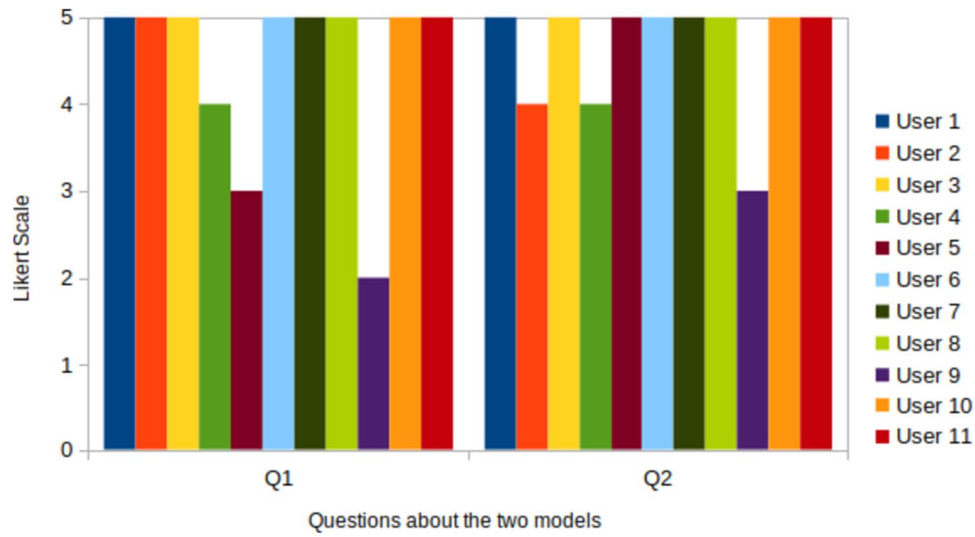


Fig. 8 Results of the questions about the two models: results by question and user

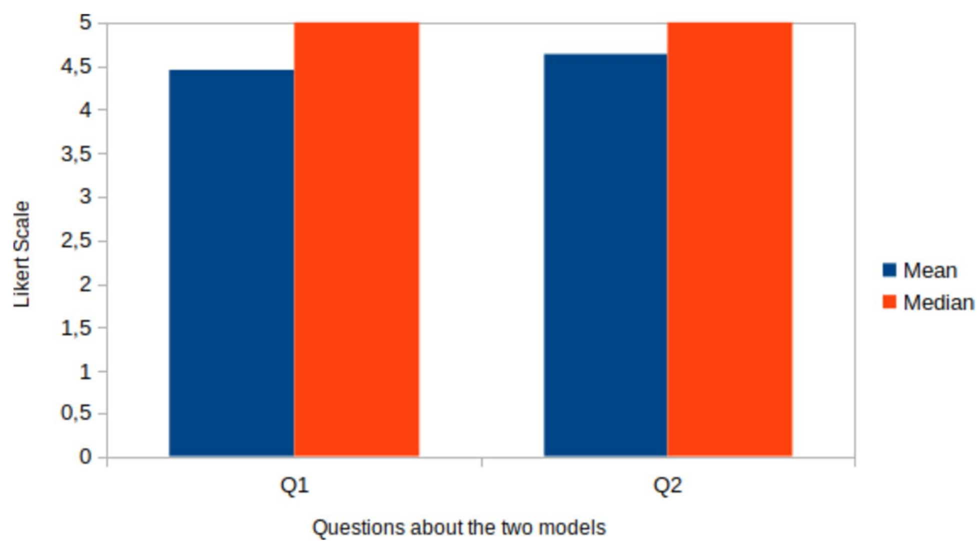


Fig. 9 Results of the questions about the two models: mean and median by question

Table 9 New prototype results

	Q1	Q2	Q3	Q4
User1	5	5	5	5
User2	4	5	5	4
User3	5	4	5	4
User4	5	5	5	5
User5	5	5	5	5
User6	3	4	5	4
User7	5	5	5	5
User8	5	5	5	5
User9	4	4	5	5

the case of people over 60 years of age, we have unanimity of criteria with a positive assessment of 100%.

In our research, we have disaggregated people who are over 60 years of age, in order to analyze the data separately. We have specifically chosen the age of 60 years, because that is when the World Health Organization begins to speak of aging [69, 70].

When disaggregating the data collected by gender and analyzing it, we find some small differences between the perceptions of women and men regarding the automatic subtitling tool. In the analysis of the data from the System Usability Scale (SUS) questionnaire, we find that women value the tool more positively than men. The mean for women in this case is 97.50, and in the case of men we have that the mean of the assessment is 90.50, as we can see in Table 10. Similarly, as we see in Table 11, women perceive that Model 1 works more smoothly and quickly than men. In the case of the questionnaire related to the improvements of the new prototype, no significant differences were found based on the gender of the people participating in the research.

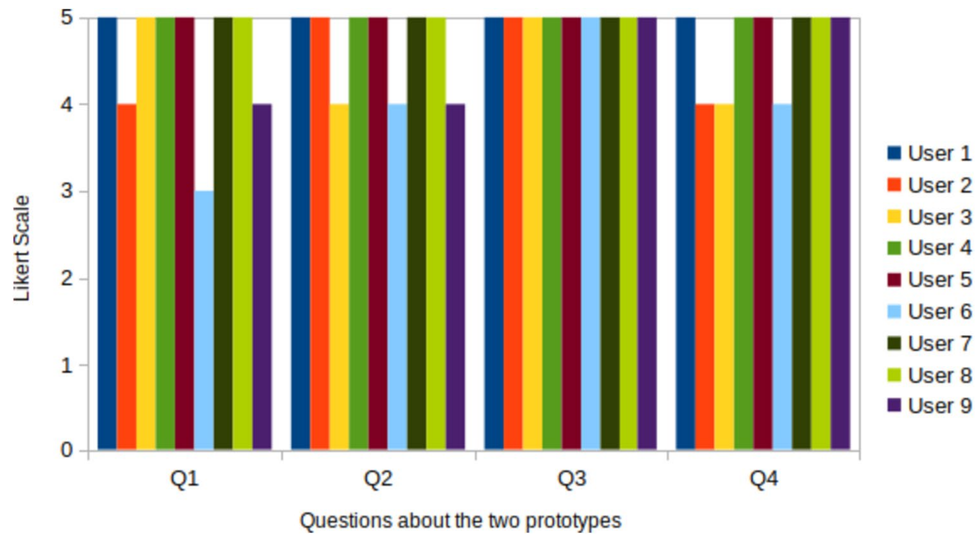


Fig. 10 Results of the questions about the two prototypes: results by question and user

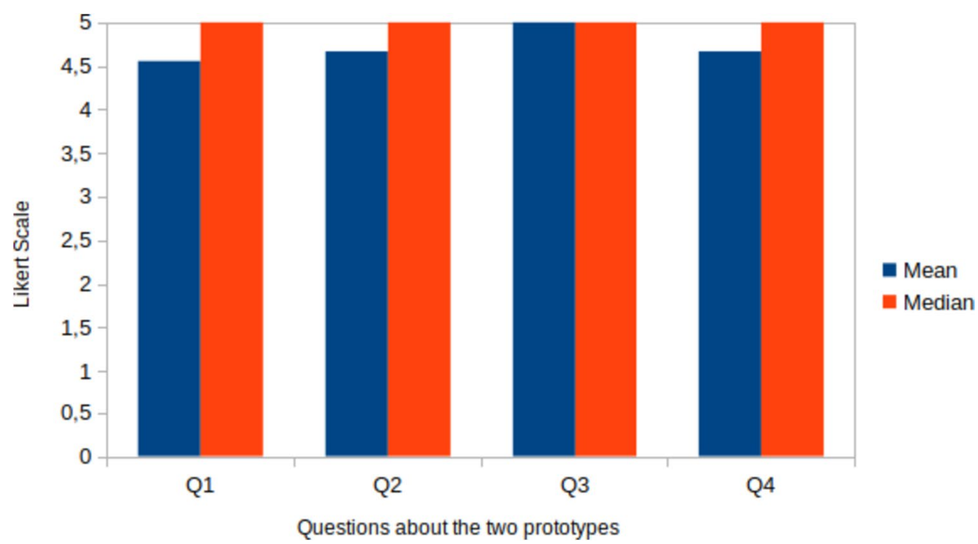


Fig. 11 Results of the questions about the two prototypes: mean and median by question

Based on our findings from the Systematic Literature Review (Sect. 2.1), it has become evident that a significant drawback of leading video calling tools lies in their inadequate support for subtitles and automatic conversation transcription [18]. Furthermore, it is important to acknowledge that when subtitles are altered or modified on the screen as a result of adjusting translated text, it introduces a higher cognitive load, ultimately impeding the comprehension of subtitles [21]. Another aspect to consider when designing a tool like the one proposed, which will be used during a video call, is the need for a user interface that is free from distractions and with sufficiently large and simple buttons [22]. This last point is important in attempting to mitigate, as much as possible, the loss of nonverbal or gestural language, as well as the need to maintain eye contact with the other person while speaking. Therefore, the arrangement of subtitles within the video call area is crucial, especially when considering a tool that will be used by individuals with hearing impairments [16, 19].

Based on these previous findings, the main design decisions for the proposed automatic subtitling tool are justified, such as offering a simple, transparent interface without distractions, and with large and simple buttons, which overlays the video calling tool. This allows the person to place the subtitles over the other person in the conversation, thus attempting to mitigate the loss of nonverbal language and maintaining visual contact as much as possible. Similarly, it is justified that once the subtitles are displayed on the screen, they should not be corrected, if we want to facilitate the use of these types of tools for elderly individuals.

Fig. 12 Comparison between both prototypes. **a**, **b** New prototype. **c**, **d** Old prototype

Table 10 SUS results by gender

Women	Normalized results
User 1	90
User 3	100
User 5	100
User 7	100
User 8	100
User 11	95
Mean	**97.50**
Men	Normalized results
User 2	97.5
User 4	75
User 6	100
User 9	80
User 10	100
Mean	**90.50**

6 Conclusions

Literature review highlights a significant drawback in leading video calling tools, namely their insufficient support for subtitles and automatic conversation transcription. This limitation negatively impacts effective communication, and when subtitles are modified on the screen due to translated text adjustments, it introduces a higher cognitive load and hampers subtitle comprehension. To address these challenges, we designed and developed an automatic subtitling tool that prioritizes key design decisions based on previous studies, offering a distraction-free user interface with large, simple buttons that overlay the video calling tool. Therefore, we solely focus on transcribing conversations between two people. Our system does not require the user to register on any online platform along with other participants for identification purposes. Additionally, our subtitling tool is independent of the video call tool and functions offline, ensuring

Table 11 Two models by gender

Gender	Fluency	Speed
Women	93.33	100
Men	84	84

that user data is not sent to the cloud for processing, thereby safeguarding privacy. After analyzing the data obtained in the different questionnaires, we can conclude that this tool and its interface are adequately adapted to the needs of the older people.

The results obtained can be divided into three blocks: in the first block we would have the answers to the System Usability Scale (SUS), in the second block, the questions related to the comparison between the two models with or without the improvement when showing the transcriptions, and in the third block we would have the comparison between the initial interface prototype and the new modified prototype based on the previous results obtained.

At the time of starting our research, we were in one of the moments with the greatest restrictions derived from the COVID-19 pandemic, so it was very difficult to carry out the evaluation only with older people and comply with current regulations. Due to these restrictions, at that time it was decided to opt for a group of volunteers of different age ranges, and thus be able to evaluate the automatic subtitling tool.

After analyzing the System Usability Scale (SUS), the score obtained in general was 94.32, which is a very high value and implies a very high approval by the people evaluated. Regarding the questions asked about the two proposed transcription models, we see that we obtained a score of around 90% both for the question about fluency and for the question about perceived speed, with these values being 89.09 and 92.73 respectively.

With these results in hand, we were able to conclude that Model 1 is better suited to the user's needs than Model 2, and therefore, the user perceives that Model 1 transcribes the text on the screen more fluidly and in addition to faster than Model 2. Therefore, we can also conclude that it compensates for the small modification made in Model 1 with respect to Model 2, in which the transcribed text is shown word by word, although in reality the text is being transcribing more slowly by artificially adding a pause of 0.1 s between each of the transcribed words.

The modifications made to the new prototype based on the results obtained in the previous phases of the investigation, allowed us to modify the user interface. Looking at the results obtained in the questionnaires in this part, we can also conclude that these modifications considerably improved the accessibility of the tool, in addition to allowing us to improve its ease of use.

In the final questionnaires comparing the two prototypes, we observed that in all questions we exceeded 90% approval, and especially in the case of improving accessibility by facilitating the reading of subtitles, where we obtained an affirmative assessment of 100%, leaving no doubt that this improvement of the tool is favorable.

In summary, the main contributions and conclusions of this work are as follows: (a) the development of a tool designed to provide an effortless experience for users during video calls, with a focus on transcribing conversations between two people; (b) this tool is user-friendly, independent of the video call tool, and functions offline, ensuring that user data is not sent to the cloud for processing, thereby safeguarding privacy; (c) the evaluation of the tool using the System Usability Scale (SUS) showed a very high approval rating (94.32), indicating that the tool is well adapted to the needs of older people; (d) the research compared two proposed transcription models, with Model 1 being better suited to the user's needs than Model 2, and showed that the modifications made to the new prototype based on the results obtained in the previous phases of the investigation allowed for improved accessibility and ease of use; (e) the research suggests that the tool can be of great help for people with hearing and/or visual problems when making video calls and can help eliminate existing technological barriers.

Finally, the tool can also be of great help when using any type of software or web service that produces an audio output in the form of a voice. We believe that it is important to continue researching and improving tools of this type to increase accessibility for all people, and to help eliminate existing technological barriers.

We believe that our work contributes to improving the user experience for older individuals and those with hearing impairments in the use of video calling tools. Furthermore, in the field of automatic captioning for video calls, there is a recognized need for more research from the perspective of user experience and user interface [20].

Acknowledgements Work co-financed by the Canary Islands Agency for Research, Innovation and Information Society of the Ministry of Economy, Knowledge and Employment and by the European Social Fund (ESF) Integrated Operational Program of the Canary Islands 2014-2020, Axis 3 Priority Topic 74 (85%). This work has been supported by the PERGAMEX ACTIVE project, Ref. RTI2018-096986-B-C32, funded by the Ministry of Science and Innovation. Spain.

Authors' contributions Writing, E.N-G; Validation, E.N-G; Investigation, E.N-G; Software, E.N-G; Supervision C.S.G-G and F.L.G-V.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. Work co-financed by the Canary Islands Agency for Research, Innovation and Information Society of the Ministry of Economy, Knowledge and Employment and by the European Social Fund (ESF) Integrated Operational Program of the Canary Islands 2014–2020, Axis 3 Priority Topic 74 (85%). This work has been supported by the PERGAMEX ACTIVE project, Ref. RTI2018-096986-B-C32, funded by the Ministry of Science and Innovation.

Spain. Agencia Canaria de Investigación, Innovación y Sociedad de la Información (TESIS2020010113).

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Availability of data and materials Not applicable.

Code availability Not applicable.

References

1. Beauford, M.: The State of Video Conferencing in 2022 (2022). https://getvoip.com/blog/2020/07/07/video-conferencing-stats/. Accessed 202204-07
2. Fortune Business Insights: Video Conferencing Market Size, Share, Trends & Growth [2028] (2021). https://www.fortunebusinessinsights.com/industry-reports/video-conferencing-market-100293. Accessed 202204-07
3. Mendes, P.R.C., Vieira, E.S., Freitas, P.V.A.D., Busson, A.J.G., Guedes, A.L.V., Neto, C.D.S.S., Colcher, S.: Shaping the Video Conferences of Tomorrow With AI. In: Anais Estendidos do Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia), pp. 165–168. SBC, Porto Alegre (2020). https://doi.org/10.5753/webmediaestendido.2020.13082
4. Juvonen, J., Schacter, H.L., Lessard, L.M.: Connecting electronically with friends to cope with isolation during COVID-19 pandemic. J. Soc. Personal Relations. **38**(6), 1782–1799 (2021). https://doi.org/10.1177/0265407521998459
5. McCarron, L.: Creating accessible videos: captions and transcripts. Commun. Assoc. Inf. Syst. **48**(1), 140–148 (2021). https://doi.org/10.17705/1CAIS.04819
6. Tan, L.F., Chua, J.W.: Strategies to stop and prevent COVID-19 transmission in long-term care facilities (LTCFs). QJM Int. J. Med. **114**(2), 151–152 (2021). https://doi.org/10.1093/qjmed/hcaa265
7. Ozaslan, M., Safdar, M., Kilic, I.H., Khailan, A., R.: Practical measures to prevent COVID-19: a mini-review. J. Biol. Sci. **20**(2), 100–102 (2020). https://doi.org/10.3923/jbs.2020.100.102
8. O'Sullivan, R., Burns, A., Leavey, G., Leroi, I., Burholt, V., Lubben, J., Holt-Lunstad, J., Victor, C., Lawlor, B., Vilar-Compte, M., Perissinotto, C.M., Tully, M.A., Sullivan, M.P., Rosato, M., Power, J.M., Tiilikainen, E., Prohaska, T.R.: Impact of the COVID-19 pandemic on loneliness and social isolation: a multi-country study. Int. J. Environ. Res. Public Health **18**, 19 (2021). https://doi.org/10.3390/ijerph18199982
9. Holt, J., Hotto, S., Cole, K.: Demographic aspects of hearing impairment. https://research.gallaudet.edu/Demographics/factsheet. php (1994). https://research.gallaudet.edu/Demographics/factsheet.php. Accessed 2022-04-30
10. Gulliver, S.R., Ghinea, G.: How level and type of deafness affect user perception of multimedia video clips. Univ. Access Inf. Soc. **2**(4), 374–386 (2003). https://doi.org/10.1007/s10209-003-0067-5
11. Hickish, G.: Hearing problems of elderly people. BMJ : British Medical Journal **299**(6713), 1415–1416 (1989). https://doi.org/10.1136/bmj.299.6713.1415
12. Szarkowska, A., Gerber-Morón, O.: Viewers can keep up with fast subtitles: evidence from eye movements. PLoS ONE **13**(6), 0199331 (2018). https://doi.org/10.1371/journal.pone.0199331
13. Google: Use captions in a video call-Computer-Google Meet Help. https://support.google.com/meet/answer/9300310. Accessed 2021-04-24
14. Zoom Video Communications, Inc.: Enabling or disabling closed captioning (2022). https://support.zoom.us/hc/en-us/articles/4409683389709-Enabling-or-disabling-closed-captioning. Accessed 2022-04-30
15. Kavcic, A.: Software accessibility: recommendations and guidelines. In: EUROCON 2005—The International Conference on "Computer as a Tool", vol. 2, pp. 1024–1027. https://doi.org/10.1109/EURCON.2005.1630123
16. Neate, T., Kladouchou, V., Wilson, S., Shams, S.: "Just not together": the experience of videoconferencing for people with aphasia during the Covid-19 Pandemic. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. CHI'22, pp. 1–16. Association for Computing Machinery, New York (2022) https://doi.org/10.1145/3491102.3502017
17. Suduc, A.-M., Bizoi, M., Filip, F.G.: Status, challenges and trends in videoconferencing platforms. Int. J. Comput. Commun. Control **18**(3), 234 (2023). https://doi.org/10.15837/ijccc.2023.3.5465
18. Acosta-Vargas, P., Guana-Moya, J., Acosta-Vargas, G., Villegas-Ch, W., Salvador-Ullauri, L.: Method for assessing accessibility in videoconference systems. In: Russo, D., Ahram, T., Karwowski, W., Di Bucchianico, G., Taiar, R. (eds) Intelligent Human Systems Integration 2021. Advances in Intelligent Systems and Computing, pp. 669–675. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-68017-699
19. Seita, M., Andrew, S., Huenerfauth, M.: Deaf and hard-of-hearing users' preferences for hearing speakers' behavior during technology-mediated in-person and remote conversations. In: Proceedings of the 18th International Web for All Conference. W4A'21, pp. 1–12. Association for Computing Machinery, New York (2021) https://doi.org/10.1145/3430263.3452430
20. Seita, M., Lee, S., Andrew, S., Shinohara, K., Huenerfauth, M.: Remotely co-designing features for communication applications using automatic captioning with deaf and hearing pairs. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. CHI'22, pp. 1–13. Association for Computing Machinery, New York (2022) https://doi.org/10.1145/3491102.3501843
21. Arkhangorodsky, A., Chu, C., Fang, S., Huang, Y., Jiang, D., Nagesh, A., Zhang, B., Knight, K.: MeetDot: Videoconferencing with Live Translation Captions (2021). https://doi.org/10.48550/arXiv.2109.09577
22. Spelter, B., Corsten, S., Diehlmann, L., Plath, A., Leinweber, J.: The user-centred design in the development of a platform for teletherapy for people with aphasia. In: Antona, M., Stephanidis, C. (eds) Universal Access in Human-Computer Interaction. Novel

Design Approaches And Technologies. Lecture Notes in Computer Science, pp. 342–359. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-05028-223
23. Geislinger, R., Milde, B., Baumann, T., Biemann, C.: Live Subtitling for BigBlueButton with Open-Source Software (2021)
24. Lyons, K.: Zoom now has auto-generated captions available for free accounts (2021). https://www.theverge.com/2021/10/25/22744704/zoom-auto-generated-captions-available-free-accounts-accessibility. Accessed 2022-04-30
25. Hakkinen, L.: Vision in the elderly and its use in the social environment. Scand. J. Soc. Med. Supplementum **35**, 1–60 (1984)
26. John, A.S.: It's not just zoom. Google Meet, Microsoft Teams, and Webex Have Privacy Issues, Too. https://www.consumerreports.org/video-conferencing-services/videoconferencing-privacy-issues-google-microsoft-webex/. Accessed 2021-04-24
27. Coleman, E.G.: Coding freedom: the ethics and aesthetics of hacking. Princeton University Press, Princeton (2012). http://bdjc.iia.unam.mx/items/show/28
28. Ava: Professional & AI-Based Captions for Deaf & HoH | Ava. https://www.ava.me/. Accessed 2022-10-20
29. Asfar, A.M.I.T., Asfar, A.M.I.A.: How to using online meeting on jitsi meet application (2020). https://doi.org/10.13140/RG.2.2.28536.80644
30. Grimaldo, F., Fuertes, A., Cobos, M., Arevalillo, M., Pérez, C., Moreno, P., Ferrís, R.: BigBlueButton, Una Alternativa de Código Abierto Para la Comunicación Interactiva en Actividades Educativas. Universidad de Oviedo. Escuela de Ingeniería Informática, Oviedo (2014). https://upcommons.upc.edu/handle/2099/15454
31. BigBlueButton: bigbluebutton/deepspeech-web. BigBlueButton (2021). https://github.com/bigbluebutton/deepspeech-web. Accessed 2021-04-25
32. Vaessen, N.: Speech-to-Text in Jitsi Meet. FOSDEM VZW (2018). https: //doi.org/https://doi.org/10.5446/41630
33. Coqui: coqui-ai/STT. coqui (2021). https://github.com/coqui-ai/STT. Accessed 2021-04-25
34. Mozilla: mozilla/DeepSpeech. Mozilla (2021). https://github.com/mozilla/DeepSpeech. Accessed 2021-04-25
35. Nacimiento-García, E., González-González, C.S., GutiérrezVela, F.L.: Free software virtual assistants for designing pervasive gaming experiences to promote active aging: State of the art. In: CEURWS.org/Vol-2747—VI Iberoamerican Conference of Computer Human Interaction, 10 (2020)
36. Das, A.: 5 Best open source video conferencing tools for group chat (2021). https://itsfoss.com/open-source-video-conferencing-tools/. Accessed 2022-04-30
37. Jitsi: Architecture Jitsi Meet Handbook (2021). https://jitsi.github.io/handbook/. Accessed 2021-05-02
38. Iswara, I.B.A.I., Yasa, I.P.P.K.: Analisis dan perbandingan quality of service video conference Jitsi dan BigBlueButton pada virtual private server. Jurnal RESISTOR (Rekayasa Sistem Komputer) **4**(2), 192–203 (2021). https://doi.org/10.31598/jurnalresistor.v4i2.794
39. Octaberlina, L.R., Muslimin, A.I.: Perceived usefulness of jitsi meet in an EFL classroom. Pegem J. Educ. Instruct. **12**(1), 177–181 (2022). https://doi.org/10.47750/pegegog.12.01.17
40. BigBlueButton Inc: BigBlueButton (2021). https://docs.bigbluebutton.org/. Accessed 2021-05-02
41. Ukoha, C.: As simple as pressing a button? A review of the literature on BigBlueButton. Procedia Comput. Sci. **197**, 503–511 (2022). https://doi.org/10.1016/j.procs.2021.12.167
42. Coelho, R.J.d.S.: Integra¸c˜ao e avalaia¸c˜ao de QoE da ferramenta bigbluebutton para webconferˆencia integrada a plataforma moodle (2018). http: //repositorioinstitucional.uea.edu.br//handle/riuea/857. Accessed 202204-30
43. Riverbank Computing: PyQt5: Python bindings for the Qt cross platform application toolkit (2021). https://www.riverbankcomputing.com/ software/pyqt/. Accessed 2021-04-25
44. Dyson, M.C., Haselgrove, M.: The influence of reading speed and line length on the effectiveness of reading from screen. Int. J. Hum. Comput. Stud. **54**(4), 585–612 (2001). https://doi.org/10.1006/ijhc.2001.0458
45. Doherty, S., Kruger, J.-L.: Assessing quality in human- and machine generated subtitles and captions. In: Moorkens, J., Castilho, S., Gaspari, F., Doherty, S. (eds.) Translation Quality Assessment: From Principles to Practice. Machine Translation: Technologies and Applications, pp. 179–197. Springer, Berlin (2018). https://doi.org/10.1007/978-3-319-91241-79
46. Park, S.-B., Oh, K.-J., Kim, H.-N., Jo, G.-S.: Automatic subtitles localization through speaker identification in multimedia system. In: 2008 IEEE International Workshop on Semantic Computing and Applications, pp. 166–172 (2008). https://doi.org/10.1109/IWSCA.2008.28
47. Coqui: Coqui-Models (2022). https://coqui.ai/models. Accessed 2022-0412
48. Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F.M., Weber, G.: Common Voice: A Massively-Multilingual Speech Corpus. arXiv:1912.06670 [cs] (2020). http://arxiv.org/abs/1912.06670. Accessed 2021-04-26
49. Coqui: English STT v1.0.0 (Huge Vocabulary). https://coqui.ai/english/coqui/v1.0.0-huge-vocab. Accessed 2022-10-13
50. Coqui: Spanish STT v0.0.1 (Jaco-Assistant). https://coqui.ai/spanish/jaco-assistant/v0.0.1. Accessed 2022-10-13
51. Campione, E., Véronis, J.: A large-scale multilingual study of silent pause duration, 4 (2002)
52. Kahng, J.: Exploring utterance and cognitive fluency of l1 and l2 english speakers: temporal measures and stimulated recall. Lang. Learn. **64**(4), 809–854 (2014). https://doi.org/10.1111/lang.12084
53. Segalowitz, N., Freed, B.F.: Context, contact, and cognition in oral fluency acquisition: learning spanish in at home and study abroad contexts. Stud. Second. Lang. Acquis. **26**(2), 173–199 (2004). https://doi.org/10.1017/S0272263104262027
54. Carver, R.P.: Reading rate: Theory, research, and practical implications **36**(2), 84–95. Publisher: [Wiley, International Reading Association]
55. Thorn, F., Thorn, S.: Television captions for hearing-impaired people: a study of key factors that affect reading performance. Hum. Factors **38**(3), 452–463 (2016). https://doi.org/10.1518/001872096778702006
56. resemble-ai: resemble-ai/Resemblyzer. Resemble AI (2021). https://github.com/resemble-ai/Resemblyzer. Accessed 2021-04-26
57. Giannakopoulos, T.: tyiannak/pyAudioAnalysis (2021). https://github.com/tyiannak/pyAudioAnalysis. Accessed 2021-04-26
58. Mozilla: Common Voice by Mozilla (2022). https://commonvoice.mozilla.org/. Accessed 2022-04-19
59. Chesley, A.: A Brief History of Crowdsourced Digital Publishing at LibriVox.org. Computers & Writing Proceedings, 8 (2017)
60. LibriVox: LibriVox. https://librivox.org. Accessed 2022-04-19
61. Gobierno de Canarias: Coronavirus. https://www.gobiernodecanarias. org/principal/coronavirus/. Accessed 2022-04-10
62. Roupa, Z.: The use of technology by the elderly. Health Sci. J. **4**(2), 10 (2010)
63. Hear-it AISBL: Definition of Hearing Loss - Mild, Moderate & Severe Hearing Loss (2012). https://www.hear-it.org/Defining-hearing-loss Accessed 2022-10-26

64. Assistant Secretary of State for Public Affairs: System Usability Scale (SUS). Publisher: Department of Health and Human Services (2013). https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html. Accessed 2021-05-02
65. Joshi, A., Kale, S., Chandel, S., Pal, D.K.: Likert scale: explored and explained. Curr. J. Appl. Sci. Technol. **7**, 396–403 (2015). https://doi.org/10.9734/BJAST/2015/14975
66. Sauro, J.: MeasuringU: Measuring Usability with the System Usability Scale (SUS). https://measuringu.com/sus/. Accessed 2021-05-02
67. W3C WAI: Captions | How-To | WCAG 3 | Web Accessibility Initiative (WAI) | W3C. https://www.w3.org/WAI/GL/WCAG3/2021/how-tos/ captions/. Accessed 2022-04-12
68. W3C WAI: Understanding Success Criterion 1.2.4: Captions (Live). https://www.w3.org/WAI/WCAG21/Understanding/captions-live.html. Accessed 2022-04-12
69. World Health Organization: Integrated Care for Older People: Guidelines on Community-level Interventions to Manage Declines in Intrinsic Capacity. World Health Organization. https://apps.who.int/iris/handle/10665/258981
70. World Health Organization: Ageing and health. https://www.who.int/news-room/fact-sheets/detail/ageing-and-health. Accessed 2022-10-25