Universal Access in the Information Society (2024) 23:167–178
https://doi.org/10.1007/s10209-022-00926-3

LONG PAPER

# Comparative analysis between a respeaking captioning system and a captioning system without human intervention

Adrian Ruiz-Arroyo[1] · Angel Garcia-Crespo[1] · Francisco Fuenmayor-Gonzalez[1] · Roxana Rodriguez-Goncalves[1]

Accepted: 26 September 2022 / Published online: 11 October 2022


**Abstract**
People living with deafness or hearing impairment have limited access to information broadcast live on television. Live closed captioning is a currently active area of study; to our knowledge, there is no system developed thus far that produces high-quality captioning results without using scripts or human interaction. This paper presents a comparative analysis of the quality of captions generated for four Spanish news programs by two captioning systems: a semiautomatic system based on respeaking (system currently used by a Spanish TV station) and an automatic system without human interaction proposed and developed by the authors. The analysis is conducted by measuring and comparing the accuracy, latency and speed of the captions generated by both captioning systems. The captions generated by the system presented higher quality considering the accuracy in terms of Word Error Rate (WER between 3.76 and 7.29%) and latency of the captions (approximately 4 s) at an acceptable speed to access the information. We contribute a first study focused on the development and analysis of an automatic captioning system without human intervention with promising quality results. These results reinforce the importance of continuing to study these automatic systems.

**Keywords** Automated closed captioning · ASR · Automatic speech recognition · Live broadcasting

## 1 Introduction

The deaf and hard-of-hearing community want to have access to the same television content that hearing people can access. Full accessibility to audiovisual content is a right for deaf and hard-of-hearing people. The information provided by television, film, social media, and streaming video services must be accessible to this community. It is necessary to ensure the quality of captions, both open and closed captions, regardless of whether content is prerecorded or live.

✉ Angel Garcia-Crespo
angel.garcia@uc3m.es

Adrian Ruiz-Arroyo
adruiza@inst.uc3m.es

Francisco Fuenmayor-Gonzalez
ffuenmay@inst.uc3m.es

Roxana Rodriguez-Goncalves
roxrodri@inf.uc3m.es

[1] Institute for Technological Development and Innovation Promotion, University Carlos III of Madrid, Leganés, Spain

In addition, for this content to be considered fully accessible, these captions must meet 5 criteria [1]:

- Captions should be accurate (errors should be minimized);
- Captions should be in a style that allows a user to understand them;
- Captions should contain a full textual representation of the audio (speaker identification and nonspeech audio description);
- Captions should be displayed long enough to be easily read and should be synchronized with the audio (caption speed);
- Captions should preserve the meaning of the content and the intent of the content.

With this in mind, many television stations have started to caption their programming. However, these criteria are not always met, with live programs (news, sports, etc.) being the most affected by these issues. A study conducted in Spain focused on comparing the captions generated for live and semi-live programs indicated that there is a longer delay

and slower captioning speed in the captions generated live by respeaking [2]. For these reasons, some researchers have focused their studies on the generation of real-time captions or the synchronization of scripts (prewritten text) with speech. There is even research focused on the development of environments that allow the deaf and blind community to access broadcast television content in the United States [3] and Spain [4].

To date, there are four main alternatives that have been studied for the generation of real-time captions: stenographers, including fast typists using conventional QWERTY keyboards and speech-to-text reporters using palantype or stenograph keyboards; automatic speech recognition method (ASR) by means of respeakers (respeaker) with and without editors that correct errors in real time; and the automatic speech recognition method using the original audio of the live broadcast (without human intervention) as the input. The latter is not widely used by broadcasters as few studies have shown that this method achieves better results compared to respeaking [5]. In addition to these methods, many broadcasters also use closed caption synchronization methods based on preplanned scripts, which can be performed manually or automatically using ASR-based systems [6, 7]; however, this method often presents errors, either because the speakers paraphrase the script, because the blocks of the previously planned transmission are changed, or because content has been added or deleted at the last minute (not being able to make the changes in the script). This method can be of great benefit as long as there is an alternative method to generate real-time captions for spontaneous speech moments. Regarding hybrid captioning (simultaneous use of script synchronization and automatic captioning), we are currently documenting a comparative study conducted in parallel to the study presented in this manuscript but with another Spanish television station. This TV station used a hybrid method for live captioning (scripts and respeaking); however, we also used hybrid captioning including a script synchronization module in our automatic captioning system without human interaction. The scripts were provided by the TV station and covered some blocks of programming. The rest of the programs were captioned automatically. The results will be published in a future manuscript.

Many studies have focused on the development and analysis of captioning systems for live television programs. Most of these focus on the respeaking method [8–10] or ASR methods that mix the respeaking and the original audio as input to the system [11]. These respeaking methods are used to prevent possible problems that may occur in the original audio (noise, mispronunciation of words, speaker changes, etc.). Respeaking is a method for generating captions in real time. In respeaking, a professional in a noise-free environment repeats or paraphrases a broadcast to be captioned that they are listening to using a microphone connected to speech-to-text software (the person must listen to the broadcast and speak at the same time without being distracted by his or her own voice). The software generates the transcript, which must be observed by a person so that he or she can correct any errors made by the speech recognition software. This text correction work is performed by the speaker him/herself when the software is already sufficiently trained so that most of the errors made are predictable and manageable by this person [12]. This method has been an alternative to manual captioning (stenography) since the professional staff needed to generate this type of captioning (stenographers or expert writers on special keyboards) is becoming increasingly scarce or simply cannot cope with the amount of programming that is currently broadcast live, and the training of new staff takes a long time and is costly. Using respeaking methods often implies a reduction of the original content (keeping as much of the context of the original content as possible). Therefore, these captions may not cover entire sentences or a speaker's original ideas since parts of the speech (the part considered less relevant) is often discarded, especially to ensure the correct latency of the captions [9]. However, if editing is truly necessary, this must be done with great care as edits are not well received by deaf or hard-of-hearing users who can read lips, and it can be frustrating for them if the captions do not show exactly what the speaker is saying. One of the reports made by Ofcom indicated that for some deaf users, "it is seen as a form of censorship and 'denying' deaf people full access to information available to the hearing population" [13].

## 2 Previous work

The previous works that we have highlighted for our study focused on the quality of captions (standards and quality measurement). First, we indicate the limits of the technical parameters that must be measured in the generated captions and detect which are the minimum values that these must present in order to consider that good quality captions have been obtained. In addition, some previous studies focused on the measurement of the quality of captions for different captioning systems, and different types of television programs are highlighted.

### 2.1 Technical parameters

The basic technical parameters that should be considered to measure the quality of the captions are as follows:

- Caption accuracy: Errors that occur in text relative to the spoken content are usually measured by WER, weighted word error rate (WWER) [14], NER [15], etc.;

- Latency delay between the speech and the appearance of the captions;
- Caption speed: number of characters per second, or words per minute, displayed on the screen.

Some research has focused on studying the preference of viewers with respect to the three parameters described above. Regarding the captioning speed, most of these studies conclude that the optimal speed should be between 150–180 words per minute (wpm). In addition, the captions should be presented in blocks; and the speed can be affected by different factors, such as the person's reading fluency and linguistic characteristics [16].

Regulatory organizations in different countries have also established best practice guidelines for television captioning, including live programming. In Spain, the UNE 153,010:2012 standard [17] states that it is not recommended to exceed a captioning speed of 15 characters per second (cps), which is equivalent to approximately 180 wpm; and the latency must be less than 8 s to be considered acceptable in Spain. Furthermore, the UK communications regulator Ofcom indicates in its best practices that the captioning speed should not exceed 160–180 wpm (more than 200 wpm would be impossible to read); in addition, it indicates that the latency should not exceed 3 s, even for live captions [18]. However, some studies have found that live captioning for live programming does not currently achieve this latency [19].

## 2.2 Caption quality measurement

Previous work has indicated that respeaking methods have better accuracy (measured by WER) than ASR methods without human intervention, especially for spontaneous speech programming, such as sports programs, interviews, and live events [8].

Other previous studies have recorded good accuracy in caption generation with respeaking techniques. One of the studies recorded an average NER of 98.38%, with a latency of 4.7 s [9]. The most complete study of closed caption quality thus far was conducted with the collaboration of Ofcom in the U.K. This study analyzed 300 programs and 78,000 closed captions (news, entertainment and interview programs). The study reported an average accuracy of 98.38%, where 58% of the errors were minor, 39% of the errors were standard and 3% of the errors were serious (in terms of the NER). Furthermore, the study reported a captioning speed of 139 wpm and latency of 5.3 s (acceptable in Spain but not in the U.K.), obtaining a latency of up to 4.6 s and 4.7 s in news programs, since the use of scripts and hybrid captioning methods (e.g., combination of prerecorded captions and live captions) is becoming more common [19, 20].

Regarding caption generation using ASR methods without human intervention (without respeaking), one study recorded average accuracies in terms of the WER of 12%, with a latency of 6.5 s [21]. However, it is not possible to compare the results of these previous works since the captions have not been performed using the same audio, tested under the same conditions in general, or assessed using the same evaluation method.

To the best of our knowledge, no previous works have focused on improving the generation of automatic captioning without human intervention in order to compare it with the respeaking method currently used by some television stations as a stable method for the automatic generation of captions. The present study aims to compare two captioning methods currently used in Spain to generate captions, allowing to determine whether either of the methods better performs the task of transcribing the captions as accurately as possible in relation to the audio while maintaining the best possible synchronization between the audio and the text. For this reason, the comparison is made by considering the accuracy, latency and speed of the captions. We compare the captions obtained by the respeaking (the method used by a Spanish broadcasting channel called Canal Extremadura) of 4 news programs with the captions obtained by the system developed by our group, which is able to generate the automatic captioning of the same programs using only the original audio as input. To do this, the broadcaster provided us with the original audio and the captions generated by respeaking in STL format (Spruce subtitle file).

## 2.3 Developed system

Figure 1 shows a general scheme of the system developed by our team, which generates captions automatically using only the audio of television programming. To generate the captions, our system requires the broadcasting company to transmit the audio to our servers; at the same time Speech to Text services are used to generate the text; this text is automatically edited using formatting and substitution rules; it is packaged in the specified captioning format for correct broadcast on television; and it is sent back to the broadcasting company, who transmits the captions to the television. On the other hand, the system that generates captions by speaking requires an additional step. In this case, the broadcaster sends the audio to the speaker; this person is responsible for dictating with a microphone and in real time what they are listening to; meanwhile, a Speech to Text service is used to generate the text; this text is edited in real time by the same speaker or by another person, who ensures that the format of the subtitle meets the specifications required for transmission on television, then send the subtitles to the broadcaster.

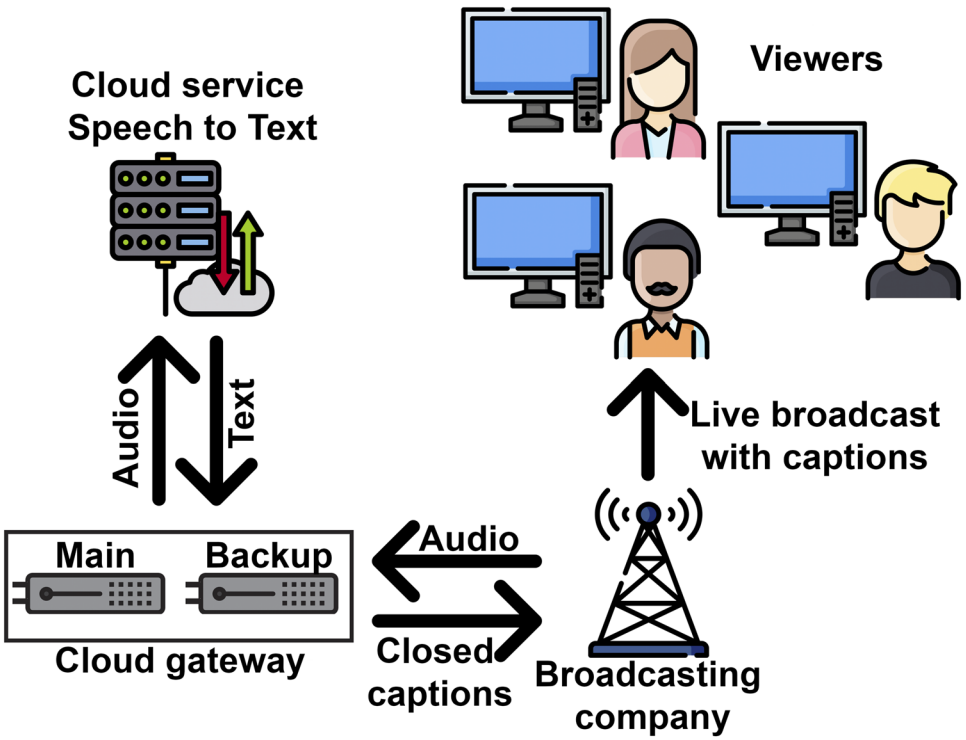


**Fig. 1** General scheme of the caption generator system

The system works in an asynchronous and sequential mode of operation in which each block is responsible for performing its own task and propagating the results to the next block. The developed system generates captions considering the UNE 153,010:2012 standard [17]. The system sets the limits for the captioning speed and latency of captions as indicated in the aforementioned Spanish standard to be considered acceptable in Spain: a maximum speed of 3 wps (equivalent to 15 cps indicated in the Spanish standard and 180 wpm indicated in the British standard, considering that in Spanish the average number of characters per word is 5) and a maximum latency of 8 s.

Captions are delivered considering the NewFor protocol, with a limit of 30 characters per line (keeping some bytes free for the inclusion of styles in the speaker recognition). In addition, captions are delivered in pop-on style (lines of subtitles that appear on screen and remain visible for one or several seconds before disappearing) even though the roll-up style is usually used in live programming because captions are textual and synchronized (the text accumulates progressively to form 2 or 3 lines, and the top line disappears when a new bottom line is created) [22]. This decision has been made because we will be captioning Spanish programs, and Spanish captioning regulations advise against using the roll-up style as this technique becomes the focus viewers' gazes too much [17]. Captioning standards and styles depend on the region in which the task is performed. For example, unlike Spain, in the United States, the roll-up style of captioning is usually used.

The automatic captioning system is responsible for generating the text, structuring the captions and sending these captions to be played in live broadcast programming. Before sending a caption block, the system performs some checks on the timing and content of the captions to ensure correct reading by viewers. Some configurable parameters for this task are the following:

- Number of lines per caption block (default 2);
- Maximum number of characters per line;
- Maximum number of characters in the caption block;
- Real reading duration of the caption considering the maximum speed set (maximum speed of 3 wps);
- Minimum and maximum duration of captions on a screen (setting lower and upper limits in case a very low or high reading duration occurs when calculating the real reading duration considering the set captioning speed);
- Maximum time that the system can wait to fill a caption line before it is sent, etc.

In addition, the system considers some rules that allow better reading of the captions. Some of these rules are described below:

- Articles must always be on the same line as the next word that accompanies it;
- A line break will be generated when some punctuation marks are present, such as a period (.);

- A line break will be generated when a change in speaker is detected, among other rules.

In addition, two word lists are implemented in the system so that fewer errors occur when generating captions (improving the accuracy). These lists are editable so that frequently used words and phrases can be added, deleted or edited. The first is a list of own words that includes information on how each of the words of interest is pronounced and is used by the system to better recognize the words that have been previously indicated in the list. The second list of words is only used for substitution. This list is used in cases where the recognizer makes a mistake when detecting proper names of places, people, political parties, etc. It is also used to change the way in which a phrase or word is presented. For example, the recognizer can detect "Felipe sexto"; and using this list, the system can generate the text with a change of format, showing "Felipe VI". All this is done to have a semiautomatic equivalence of the edits made almost always manually when captions are generated by respeaking.

Figure 2 shows a summary of the verification process from the moment a caption block is sent until the next block is sent. The consideration of the rules is not explicitly indicated in the diagram shown, but they are considered at all times throughout the process of structuring the caption block.

The reading duration is calculated considering the number of characters of the caption block previously sent and the predefined captioning speed; however, this reading duration has two limits: a lower limit and an upper limit. The lower limit exists so that the captions do not disappear quickly from the screen, and the upper limit exists so that captions do not remain too long on the screen, delaying the rest of the captions. These limits are considered with the purpose of presenting harmony and synchrony between what the speaker says and the captions. However, if the recognizer presents partial results, evaluations are made in order to ensure that the captions present a legible structure. For example, if the last thing that has reached the recognizer is a number, we wait additional time to ensure that the complete number is recognized. In addition, the number of lines that have accumulated in the recognizer and their content are considered, allowing us to wait an additional time if the captions are considered to have very little content. All this is summarized in Fig. 2.

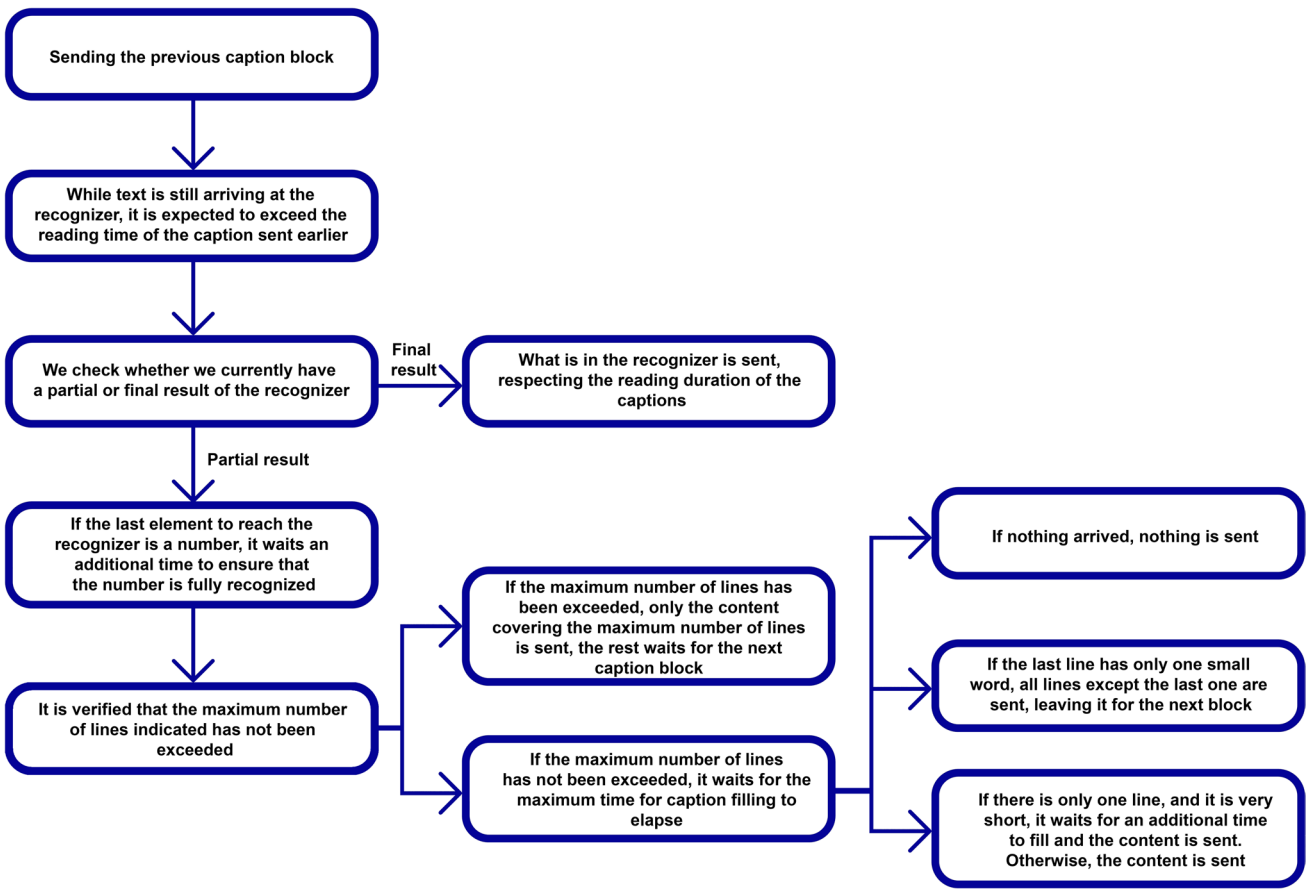


**Fig. 2** Caption block verifications prior to submission

## 3 Methods

A total of four live news programs were analyzed; the captions generated by means of two ASR methods, a method that uses respeaking audio as input and a method that uses only the original audio as input (without human intervention), were compared. To conduct this study, Canal Extremadura provided us with its captions in STL files generated by respeaking (method used thus far by this TV station to generate closed captions for its live programming). In addition, the TV station provided the original audio so that our system could generate automatic captions using this audio. The captions obtained by both methods were compared in terms of accuracy, speed and latency since these are the three factors commonly used to measure the quality of captions. In the evaluation and comparison of the quality of subtitles, we do not consider the semantics of the subtitled content, so for the purposes of this study, the reading of subtitles is not related to the understanding of the content. In the same way, words are taken as a value gathered by the physiology of the human eye, and not by a human reading and understanding.

In the present study, the captioning accuracy was measured using the WER. To calculate the WER, this model considers the erroneous substitutions of one word for another, the number of words that are pronounced in the audio and omitted in the captioning, and the number of word insertions in the captioning (not said in the audio) with respect to the total number of words actually pronounced in the audio. This metric was chosen because the present analysis was based on a quantitative comparison of the results obtained by both captioning methods when transcribing the captions as accurately as possible in relation to the audio.

The speed at which the captions appear, and their latency were also analyzed. This was performed through a temporal analysis of the generated captions. This is able to consider aspects slightly more focused on the viewer, such as the speed at which a text can be correctly read and the synchronization between text and voice (considering the captioning speed and latency limits described in previous works and in national and international broadcasting standards). To conduct these comparisons, three files were used (per program): a first file called the reference text, which contains the captions manually corrected by the authors of this study so that they correspond to the literal content transmitted in the different programs analyzed; a second file containing the captions generated with the respeaking method (the system used by the television station that broadcast the programs to be evaluated); and a third file containing the captions generated by the system developed by us. The second and third files are called hypothesized texts and are compared with the reference text to analyze the differences between them. In this task, we analyze the number of missing words, the number of wrong words, the latency of the captioning and the speed of the captions.

## 4 Results

In office meeting environments, errors made by humans when transcribing audio are estimated to be approximately 5% in terms of the WER while the WER for ASR in the cloud is approximately 20%. Considering the trend of improving ASR accuracy, some authors indicate that this accuracy limit should be set at a better level, proposing 15% for the WER in the cloud [23]. For the purposes of this information, the following ranges are considered for this study:

Excellent accuracy: $\mathrm{WER} \leq 5\%$
Good accuracy: $5\% < \mathrm{WER} \leq 15\%$
Acceptable accuracy: $15\% < \mathrm{WER} \leq 20\%$
Poor accuracy: $\mathrm{WER} > 20\%$

### 4.1 Precision

Considering the limits described above, the accuracy of the captioning generated without human intervention obtained good or excellent accuracy in terms of the WER (6.13%, 7.29%, 7.26% and 3.76%, respectively, for each of the programs) while the captions generated by the method currently used by this Spanish TV station obtained poor accuracy (32.24%, 41.79%, 44.14% and 40.06%, respectively). Figure 3 shows the WER in each of the live transmissions for both captioning methods used. In addition, the WER limits are highlighted with dark lines to facilitate the reading of the diagram.

In the analysis performed for these 4 news programs, 66.5% of the errors included in the captioning generated without human intervention were due to word substitution (the system recognizes a word incorrectly), 20.18% were insertion errors of words that were not pronounced and 13.3% were missing content errors (words that were pronounced and not recognized). Among the words that were omitted and inserted erroneously, most were connectors. Furthermore, the words that were replaced by erroneous words usually corresponded to names of cities, people, and political parties, among others. These types of substitution errors are solved by training the system and improving the quality of the generated captioning over time.

Regarding the captioning generated by the respeaking method, 94.14% of the errors made were due to lack of

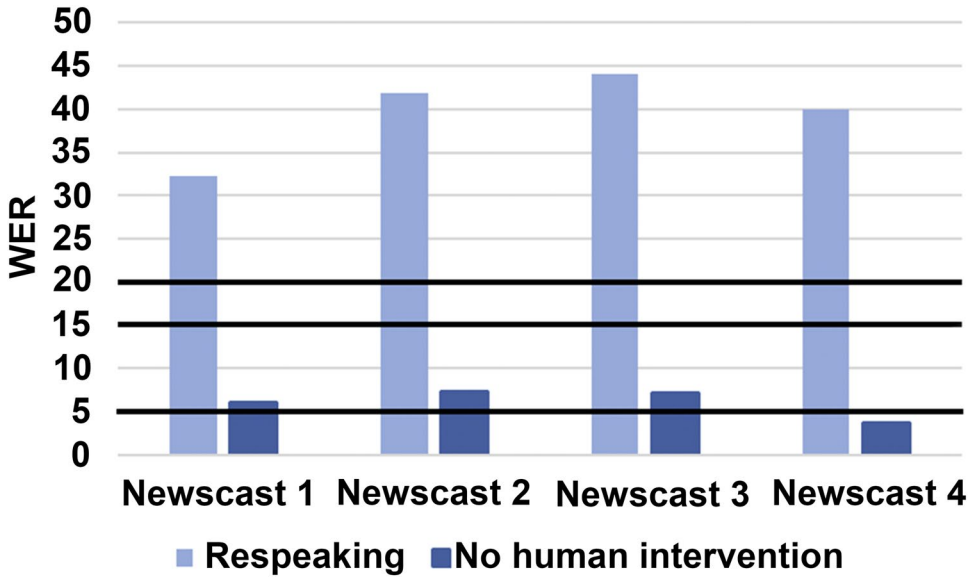


**Fig. 3** WER in each of the transmissions for both methods used

content (programming with missing captions), 4.77% were word substitution errors and 1.09% were word insertion errors. In other words, approximately 9 out of 10 errors found involved the omission of a word or phrase. In some cases, complete sentences were omitted, which can cause a viewer to miss important information in a conversation or news item. An example of this can be seen in Fig. 4, where a section of newscast 2 is detailed. The substitution errors are marked in red, and where the same content of this section begins and ends in the three files generated is marked in orange.

This figure shows that in the captioning generated by the respeaking method, a considerable amount of content has been lost. This occurs several times throughout the programming. Therefore, deaf viewers could lose relevant information to understand the programming, or those deaf viewers who are able to read lips could feel confused as the captions do not match what the speaker is discussing.

It was to be expected that our system would present a better WER than the respeaking system, as respeakers sometimes paraphrase what they hear, or omit small parts to represent the content in a simpler way and catch up with the speaker. However, in news programs, because of the speed of speech of the speakers and of some interviewees, many times the content omissions they make are large so as not to affect the latency of the captions too much. In this study it was observed that the respeaking method used by the television station presented a big omission of continuous content, being this omission of content the solution to improve the large latency that was present, without even being able to paraphrase the content. In future it would be interesting to extend this study using a method to evaluate the precision of the subtitles considering the semantics of the content.

### 4.2 Latency

The latency obtained in the captioning generated with our system is approximately 4 s (for the four newscasts analyzed). However, the captions generated by respeaking presented an average latency between 6.9 and 12.2 s. Figure 5 shows the behavior of the latency in the 4 captioning programs using both methods. By analyzing the respoken captions, we observed that one of the reasons why content reductions were made in the captioning (impairing accuracy) was due to the constant increase in latency throughout the programming; however, this is not a good practice if it is not done carefully since much content is lost in the programming, preventing deaf viewers from having access to all the information transmitted.

Figure 5 shows that the median latencies of the captions generated by our system are 4.4, 4.2, 4.75 and 4.5 s for each program, respectively. The minimum is between 1.9 and 2 s; the maximum is between 6.8 and 7.7 s; the standard deviations are 0.24, 1.01, 1.02 and 0.8, respectively. However, the captions generated by respeaking had a median per program of 13.3, 6.95, 12.1 and 11.9 s, respectively. The minimums are between 1.2 and 2.4; the maximums are between 13.8 and 26.7 s; and the standard deviations per program are 5.5, 3.05, 4.94 and 5.67 s, respectively.

Considering the Spanish UNE standards, the latency of captions in live broadcasted programs should not exceed 8 s while the best practices indicated by UK Ofcom establish that the latency should be a maximum of 3 s. Therefore, we can indicate that our system complies with the Spanish subtitling regulations, even though it does not reach the requirements of the UK regulations. However, as far as we know, the latter is very difficult to achieve in

| Transcripción de referencia | Subtítulo generado con audio original | Subtítulo generado con audio rehablado |
|---|---|---|
| Abrimos en este punto<br>de las portadas de la prensa | Abrimos en este punto de las<br>portadas de la prensa para | Abrimos en este punto<br>la prensa para ver |
| para leer que titulan<br>a esta hora a los medios | leer que titulan a esta hora<br>a los medios de comunicación. | qué titulan a esta hora los medios<br>de comunicación. |
| de comunicación, | Comenzamos por Canal<br>Extremadura a punto | Pasando de nuevo por Extremadura, |
| comenzamos por Canal<br>Extremadura punto es, | de 7 fallecidos por Coby | el pintor ecuatoriano ha pintado<br>en las paredes y los olores y |
| 7 fallecidos por Covid-19<br>y descenso en el número | 19 y descenso en el número<br>de ingresos. El diario | colores de su ciudad natal. |
| de ingresos, | y destaca a esta hora cierre<br>perimetral de Los Santos | |
| el diario Hoy destaca<br>a esta hora, | de Maimona y prorroga<br>en Hervás al Huéscar | |
| cierre perimetral de Los Santos<br>de Maimona y prorroga en Hervás, | y Arroyomolinos y El Periódico<br>Extremadura. | |
| Alcuéscar y Arroyomolinos<br>y el Periódico Extremadura, | Titular a esta hora de la | |
| titula a esta hora de la noche,<br>Extremadura lamenta 7 fallecidos | noche, Extremadura a la venta,<br>7 fallecidos más y | |
| más y 447 positivos<br>de coronavirus, | 447 positivos de coronavirus.<br>La prensa nacional el diario | |
| la prensa nacional, el diario<br>El País, España impondrá pruebas | El País España impondrá<br>pruebas PCR en origen | |
| PCR en origen a quienes<br>llegan de países de riesgo | a quienes llegan de países<br>de riesgo desde el próximo | |
| desde el próximo 23 de noviembre<br>y cerramos con el diario | 23 de noviembre y cerramos con<br>el diario El Mundo. | |
| El Mundo, | El Gobierno recula ante<br>la presión de las comunidades | |
| El Gobierno regula<br>ante la presión | autónomas y exigirá PCR<br>a todos los turistas que | |
| de las comunidades autónomas<br>y exigirá PCR a todos | lleguen de países de riesgo. | |
| los turistas que lleguen<br>de países de riesgo. | Pasando de nuevo por<br>Extremadura. El pintor | |
| Pasando de nuevo<br>por Extremadura, | ecuatoriano Joaquín Bórquez<br>ha colgado en las paredes del | |
| el pintor ecuatoriano Joaquín<br>Bórquez ha colgado | Palacio de la Isla de Cáceres<br>en forma de pinturas, los | |
| en las paredes del Palacio<br>de la Isla de Cáceres en forma | olores, los colores y los<br>sabores de su ciudad natal | |
| de pinturas, los olores,<br>los colores y los sabores | de Guayaquil. | |
| de su ciudad natal de Guayaquil, | | |

**Fig. 4** Part of the generated files (exact transcription, captioning generated without human intervention and captions generated by respeaking)

live programs without scripts, and the studios that have obtained latencies similar to or lower than those obtained by our system are systems that include scripts of the programming or part of the programming. Furthermore, captioning by respeaking only obtained results acceptable by Spanish standards in one of the programs, with a median of 6.95 s (newscast 2), while in the rest it presented higher latencies, preventing the correct understanding of the transmissions.

### 4.3 Captioning speed

Regarding the captioning speed, the system generated captions with a mean speed of 15.9 cps (approximately 190 wpm); medians per program of 16.44, 16.76, 17.5 and 16.93 cps, respectively; and standard deviations of 6.08, 6.09, 5.93, and 5.82 cps, respectively. The minimum captioning speeds were between 0.5 and 1 cps, and the maximum speeds were between 28.76 and 34.72 cps (Fig. 6). The average speed of the captions is higher than

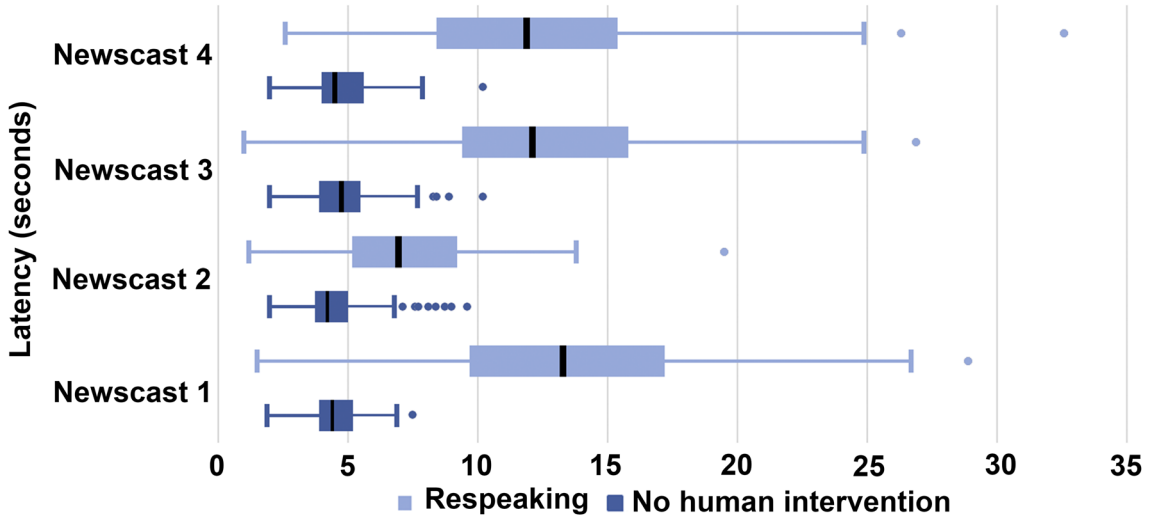


**Fig. 5** Latency in the captions of the transmissions for both methods used. The boxes represent inner quartiles, the thick black lines represent medians, the colored dots represent outliers and the thick colored lines represent maximum and/or minimum points

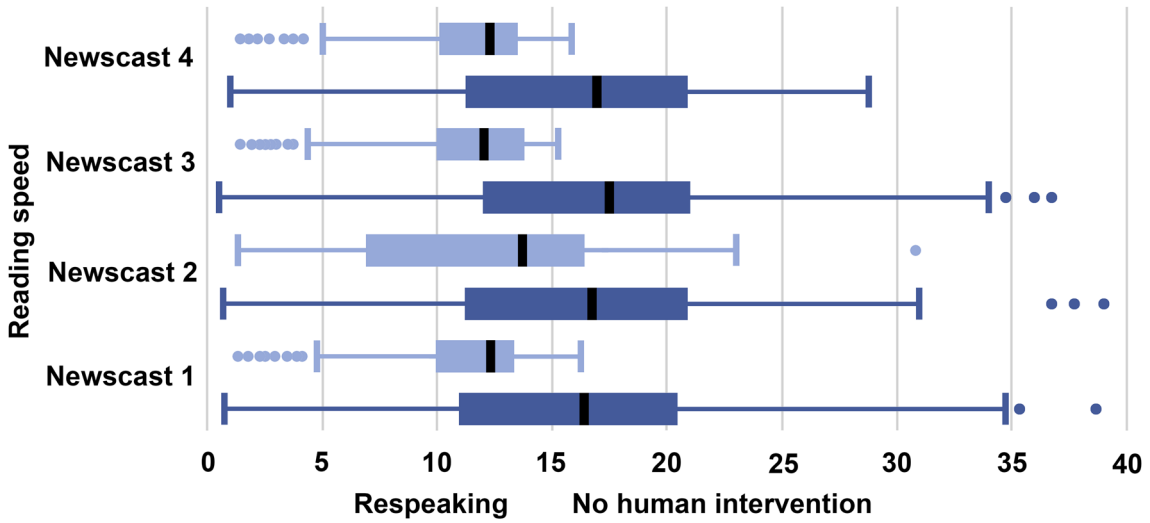


**Fig. 6** Speeds of the captions of the transmissions for both methods used. The boxes represent inner quartiles, the thick black lines represent medians, the colored dots represent outliers and the thick colored lines represent maximum and/or minimum points

the maximum captioning speed recommended for the captioning of television programs (15 cps or 180 wpm).

However, the captions generated by the respeaking presented an average captioning speed of 11.55 cps (approximately 133 wpm); medians per program of 12.31, 13.71, 12.03 and 12.28 cps, respectively; and standard deviations of 4.86, 5.39, 4.25 and 3.5 cps, respectively. In addition, the captions registered significant minimums between 1.33 and 4.15 and maximums between 15.28 and 23 cps. At this captioning speed, viewers will be able to read the captions correctly without any problem. Figure 6 shows the speeds obtained for the captions generated by both methods studied.

The results obtained were to be expected due to the accuracy and latency achieved by the system since the speech rate in news programs is high (an average of 2.57 words per second, which is equivalent to 154 wpm), and some programs even exceed average speeds of 3 words per second (180 wpm) [24]. Furthermore, the pop-on captioning style prevents us from displaying the text at the same time as the audio is recognized, making it necessary to accumulate a number of characters and/or wait for a specific time before sending the next caption block.

However, news programs, compared to entertainment programs (movies or series), do not require the same balance in eye focus between captions and video to understand

the information transmitted; therefore, it is possible that the viewer focuses longer on the captions to correctly read them without losing the content of the information transmitted. A previous study indicated that lower speed is not necessarily synonymous with good quality captions, even though the participants of the study preferred to watch a movie (programming that demands more attention to the video compared to news) at a high speed (20 cps, equivalent to 240 wpm), provided that good synchronization between captions and audio (latency) was maintained. They also indicated that the reduction in text content caused frustration (this is true for people who can compare what the speaker says with the caption text) [25].

Due to the results obtained in this study, we aim to start a process of updating the system in order to reduce the maximum speed points that have been registered in some parts of the programming. This will have to be studied in detail since in the case of modifying the system, these updates will have to be implemented without harming the other quality parameters of the captions (accuracy and latency) on a large scale. It will occur in this way, considering that all parameters are of great importance for the programming content to be correctly perceived by the viewer. Accuracy and latency cannot be disregarded in order to obtain a better captioning speed.

## 5 Conclusions

To the best of our knowledge, this is the first study that focuses on the development and analysis of a complete automatic captioning system without human intervention. A comparison is made of the quality of the captions obtained with our system to that of the system currently used by a Spanish television station. Generally, without considering the semantics of the content, better quality results are obtained with our proposed system. The system developed presented better quality results with respect to two of the three parameters evaluated (accuracy and latency). However, the respeaking system obtained better quality results from the perspective of captioning speed.

The respeaking method used by the television station presented a WER greater than 20% in all programming. Considering the scale used for the evaluation, this would imply poor accuracy, which is prejudicial for deaf people, since the information is not transmitted completely, and it could even be frustrating for those who are able to read the speakers' lips and realize that the captions do not match what the speakers are discussing. However, the captions generated by the proposed system presented promising results regarding accuracy (3 out of 4 schedules obtained $5\% < \text{WER} < 10\%$, and the last schedule obtained a $\text{WER} < 5\%$). It was expected that the system would have a higher accuracy, since the evaluation with WER does not consider the semantics of the content, however, the captioning system with respeaking not only omitted redundant content or due to paraphrasing (replacing content and maintaining the semantics of the sentence), often content was omitted only to solve latency problems, as indicated above, with the respeaking system 94.14% of the errors committed were due to lack of content.

The latency of the captions generated is also very promising, and previous studies have achieved similar results only when using hybrid captioning methods (using scripts and ASR). In addition, our system complies with Spanish regulations on the quality of captions in television programming. Currently, the system does not comply with the UK regulations; however, there is currently no unscripted captioning system that achieves the latencies desired by these regulations since doing so would greatly impair the other two quality parameters. However, the captioning generated by respeaking presented a latency higher than 8 s in three of the four captioned programs (maximum latency indicated by the Spanish captioning standards).

Regarding the captioning speed, the speed obtained by our system is slightly higher than the maximum speed indicated by the regulations for captioning television programs. However, at this speed, it is still possible to read the content, allowing accessibility to the information transmitted in the newscast. This also considers the NewFor protocol used for sending captions to the TV station, which limits the number of characters per line; and the pop-on style for sending captions, which forces us to wait to fill a block before sending the captions. Slowing down the speed and impairing the accuracy and/or latency of captions are not options to consider as these two factors are usually the most important for viewers (as long as the speed allows the captions to be read). Although the captions generated by respeaking have an average captioning speed below the maximum indicated by the captioning regulations, allowing a proper reading of the captions, these captions are inadequate since the content displayed and the latency of the captions make it very difficult to understand the programming.

This study constitutes a first step towards the further development of fully automatic subtitling systems without human interaction. The results obtained are promising, considering the regulations for the transmission of subtitles on television. With this paper we seek to encourage other researchers and ourselves to continue with the study of different methods to generate and correct subtitles without human interaction.

## 6 Limitations and future work

A limitation of our system, which may not be present in a captioning system in another country, is the style in which the captions must be delivered to the Spanish TV station.

The pop-on style, unlike the roll-up style (used in the United States and recommended for live transmissions), prevents us from sending captions constantly, which creates delays (increasing latency) and forces us to keep the captions on screen for less time (increasing the captioning speed); however, the results obtained for automatic subtitling without scripts or human intervention are the best recorded thus far. Furthermore, we document the quality results of the captions generated by our system by including an additional script synchronization module. These results will be published in a future manuscript.

For future work, we propose enriching the study by also performing subjective analysis of the accuracy of the captions considering metrics such as the NER, since, in this way, not only the subtitled word would be considered, but also the semantics of the sentence.

As mentioned in the results on captioning speed, we aim to continue testing system updates in order to lower the maximum speed points recorded in these tests (allowing the captions to be visible for a longer time on the screen). The reduction of this speed will be studied considering the results obtained in this study in order to obtain better captioning speeds without greatly affecting the other two quality parameters (accuracy and latency).

For future work, we propose including a field test to determine the preferences of deaf or hard-of-hearing viewers when they want to access the information transmitted in the newscasts of their territorial, national, or international area.

**Acknowledgements** The authors would like to thank the broadcasting channel Canal Extremadura for allowing us to carry out this study using the content of its programming, as well as the staff of the channel for giving us all their support, providing us with the original audio and captions generated for their programming.

**Author contributions** Conceptualization (AR-A and AG-C), methodology (AR-A, AG-C and FF-G), software (AR-A and FF-G), validation (AR-A, FF-G and RR-G), formal analysis (AR-A and RR-G), writing—original draft preparation (AG-C and RR-G), writing—review and editing (AR-A and FF-G), project administration (AG-C), funding acquisition (AG-C).

**Funding** Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. No funding was received for conducting this study.

**Data availability** The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

## Declarations

**Conflict of interest** The authors declare that they have no conflict of interest/competing interests.

**Ethical approval** The authors declare that the study involved no human intervention.

**Consent for publication** All authors read and approved the manuscript for publication.



## References

1. European Union of the Deaf: Accessibility of information and communication. https://www.eud.eu/about-us/eud-position-paper/accessibility-information-and-communication/ (2018). Accessed 19 Feb 2021
2. De los Reyes Lozano, Mejías-Climent, L., Martí Ferriol, J.L.: Retraso y velocidades de lectura en la subtitulación para personas sordas de los informativos. Sendebar **31**, 69–86 (2020). https://doi.org/10.30827/sendebar.v31i0.11836
3. Rodríguez, J., Díaz, M.V., Collazos, O., García-Crespo, Á.: GoC-C4All a pervasive technology to provide access to TV to the deafblind community. Assist. Technol. (2021). https://doi.org/10.1080/10400435.2020.1829176
4. García-Crespo, A., Montes-Chunga, M., Matheus-Chacin, C.A., Garcia-Encabo, I.: Increasing the autonomy of deafblind individuals through direct access to content broadcasted on digital terrestrial television. Assist. Technol. **32**, 268–276 (2020). https://doi.org/10.1080/10400435.2018.1543219
5. Ofcom: The quality of live subtitling. https://www.ofcom.org.uk/__data/assets/pdf_file/0017/45602/subtitling.pdf (2013). Accessed 2 Mar 2021
6. Ortega, A., Garcia, J.E., Miguel, A., Lleida, E.: Real-time live broadcast news subtitling system for Spanish. In: INTERSPEECH 2009, pp. 2095–2098. Brighton, United Kingdom (2009)
7. Gao, J., Zhao, Q., Li, T., Yan, Y.: Simultaneous synchronization of text and speech for broadcast news subtitling BT. In: Yu, W., He, H., Zhang, N. (eds.) Advances in Neural Networks—ISNN 2009, pp. 576–585. Springer, Berlin (2009)
8. Levin, K., Ponomareva, I., Bulusheva, A., et al.: Automated closed captioning for Russian live broadcasting. In: INTERSPEECH 2014, pp. 1438–1442. Singapore (2014)
9. Pražák, A., Loose, Z., Psutka, J.V., et al.: Live TV subtitling through respeaking with remote cutting-edge technology. Multimed. Tools Appl. **79**, 1203–1220 (2020). https://doi.org/10.1007/s11042-019-08235-3
10. Boulianne, G., Beaumont, J.-F., Boisvert, M., Brousseau, J.: Computer-assisted closed-captioning of live TV broadcasts in French. In: INTERSPEECH 2006, pp. 273–276. Pittsburgh, PA, USA (2006)
11. Imai, T., Homma, S., Kobayashi, A., et al.: Speech recognition with a seamlessly updated language model for real-time closed-captioning. In: Kobayashi, T., Hirose, K., Nakamura, S. (eds.) INTERSPEECH 2010, pp. 262–265. Makuhari, Chiba (2010)
12. Rufino Morales, M.: Subtitulación a través de la técnica del rehablado. Integración **96**, 53–61 (2020)

13. Ofcom: Subtitling—an issue of speed? https://www.ofcom.org.uk/__data/assets/pdf_file/0018/16119/subt.pdf (2005). Accessed 2 Mar 2021
14. Apone, T., Brooks, M., O'Connell, T.: Caption viewer survey: error ranking of real-time captions in live television news programs. In: National Center for Accessible Media. http://ncamftp.wgbh.org/ncam-old-site/file_download/CCM_survey_report_final_Dec_17_2010.pdf (2010). Accessed 26 Mar 2021
15. Romero-Fresco, P., Pérez, J.M.: Accuracy rate in live subtitling: the NER model BT. In: Piñero, R.B., Cintas, J.D. (eds.) Audiovisual Translation in a Global Context: Mapping an Ever-Changing Landscape, pp. 28–50. Palgrave Macmillan, London (2015)
16. Fresno, N., Sepielak, K., Krawczyk, M.: Football for all: the quality of the live closed captioning in the Super Bowl LII. Univers. Access. Inf. Soc. (2020). https://doi.org/10.1007/s10209-020-00734-7
17. UNE: Subtitling for deaf and hard-of-hearing people. https://www.une.org/encuentra-tu-norma/busca-tu-norma/norma?c=N0049426 (2012)
18. Ofcom: Ofcom's Guidelines on the provision of television access services. https://www.ofcom.org.uk/__data/assets/pdf_file/0025/212776/provision-of-tv-access-services-guidelines.pdf (2021). Accessed 2 Mar 2021
19. Ofcom: Measuring live subtitling quality. https://www.ofcom.org.uk/__data/assets/pdf_file/0011/41114/qos_4th_report.pdf (2015). Accessed 2 Mar 2021
20. Romero-Fresco, P.: Accessing communication: the quality of live subtitles in the UK. Lang. Commun. **49**, 56–69 (2016). https://doi.org/10.1016/j.langcom.2016.06.001
21. Meinedo, H., Viveiros, M., Neto, J.: Evaluation of a live broadcast news subtitling system for portuguese. In: INTERSPEECH 2008, pp. 508–511. Brisbane, Australia (2008)
22. DCMP Captioning Types, Methods, and Styles. https://dcmp.org/learn/38-captioning-types-methods-and-styles. Accessed 5 Aug 2021
23. Berke, L., Kafle, S., Huenerfauth, M.: Methods for evaluation of imperfect captioning tools by deaf or hard-of-hearing users at different reading literacy levels. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, pp. 1–12. Association for Computing Machinery, New York (2018)
24. Leetaru, K: How Fast Do People Speak On Television News? In: Forbes. https://www.forbes.com/sites/kalevleetaru/2019/08/07/how-fast-do-people-speak-on-television-news/?sh=4c7ab6a747fd (2019). Accessed 5 Aug 2021
25. Szarkowska, A., Gerber-Morón, O.: Viewers can keep up with fast subtitles: evidence from eye movements. PLoS ONE **13**, e0199331 (2018). https://doi.org/10.1371/journal.pone.0199331