IBC2025

USING GENERATIVE-AI SPEECH-TO-TEXT OUTPUT TO PROVIDE AUTOMATED MONITORING OF TELEVISION SUBTITLES

Michael Armstrong

Associate Staff, University of Dundee, UK

ABSTRACT

This paper describes a proof-of-concept approach to monitoring timing errors and word loss in TV subtitles. It reviews previous attempts at subtitle monitoring and the problems caused to viewers by subtitle timing errors and word loss. It then introduces the use of speech to text technology and the conventions in subtitling where repetition, non-speech content and errors can make the task of aligning the speech-to-text transcript to subtitles more challenging. The paper describes the approach taken to remove non-speech content from the subtitles and transcript, along with the natural language processing techniques used to ensure a sufficiently accurate alignment between the two. It then gives examples of the ways in which the results are displayed and some sample results showing the scale of problems with subtitle quality. The paper concludes by reviewing the limits of this approach in terms of accuracy and points out the need for human oversight. Then it goes on to discuss where this approach could be used and other subtitle quality issues which could be monitored automatically.

INTRODUCTION

AI-based speech-to-text tools cannot currently provide broadcast-quality subtitles without human intervention. However, because speech-to-text tools produce different types of errors to those usually found in the production of television subtitling, they can be used to monitor some aspects of subtitle quality. This paper shows how problems with timing and word omission can be detected and quantified by using speech-to-text tools along with natural language processing and simple statistics.

This paper uses the word "subtitles" to refer to text which represents the spoken words along with additional information about the soundtrack in the same language as the speech as provided on UK TV services. These primarily aim to serve D/deaf and hard of hearing audiences but are used by many more people to enhance their viewing experience. This service is also known as "closed captions" outside of the UK. This paper does not cover subtitles for translation which is a more complex topic with different quality issues. The two display modes of subtitling are referred to as "block" where each subtitle appears as sets of words and are replaced by the next subtitle and "snake" where words are added to a subtitle one (or more) at a time and lines scroll up to make space for the next line.

BACKGROUND

The quality of live television subtitles has been a cause of audience complaints for many years. As part of a programme of research into subtitle quality at BBC R&D in 2012 [1] we carried out user research to quantify the effect of word errors and delay on perceived subtitle quality [2]. This led to changes to the systems for delivering live subtitles at the BBC, using pre-prepared subtitle blocks for scripted speech and pre-recorded packages along with reduced encoding delays for snake subtitles generated by respeaking [3].

Around this time Ofcom ran an exercise in monitoring live subtitle quality. This involved broadcasters and subtitling students manually assessing the quality of 10-minute clips of live subtitles once every 6 months over a period of two years [4,5,6,7,8]. As a response to the questionable methodology, in particular the extremely sparse sampling, we developed a prototype tool capable of continually monitoring some aspects of DSAT subtitles, such as the subtitle word rate, the subtitle format, and the position on screen. This provided 24/7 baseline data for both live and pre-prepared subtitles and helped identify some fault conditions [9]. While the tool was short-lived, it contrasted with Ofcom's approach and provided useful information about UK TV subtitling practice. Importantly it revealed that UK subtitles were being produced with peak word rates up to and over 220 words per minute, well above the nominal 180wpm limit. This led to further research into the impact of word rate on subtitle enjoyment and perceived speed [10].

In the USA, the Media Access Group at WGBH ran a 3-year research programme called *Automated Error Ranking of Real-time Captions in Live Television News Programs*, ending in 2011. This explored the idea of comparing the output of a speech-to-text engine to live subtitles to gauge their accuracy. This project inspired this work, in particular their observation that while the speech to text engine produces errors, they are likely to be different from subtitling errors produced by stenography and respeaking [11].

More recently, the EBU Quality Control project [12] includes several definitions of Quality Control (QC) tests for aspects of subtitle quality. The test of *Subtitle Alignment* is listed as "human-only review" [13]. This work demonstrates a way of automating this test.

QUALITY FAILINGS IN SUBTITLES

Word Errors

Almost all academic research on subtitle quality has viewed word errors as the main issue affecting quality. This is, in part, because the researchers focused on the making of subtitles rather than on the audience experience. Also, word errors are easy to visualise and illustrate in print. As a result they have been the topic of much press coverage [14], but, in practice, word errors are not the most important quality issue for the audience.

Timing

In interim data from a survey by the UK Subtitling Audiences Network, the participants rated subtitles being out of sync with the speech, as the most noticeable problem. This was selected by two thirds of respondents [15]. This reinforces previous work on the impact of subtitle delay which showed that it was a far more significant problem for most people than word errors. The impact is so serious that delays of over 10 seconds render the subtitles effectively useless for most people [2]. Subtitle delay is an ongoing problem with live programmes as well as late-delivered, pre-recorded programmes, which are subtitled live. Subtitles that appear early also negatively affect subtitle quality, especially if the subtitles end before the speech starts.

Word Loss

Deaf and hard-of hearing people are consistent in their preference for verbatim subtitles [16], and "subtitles do not accurately reflect what is said" was the second most noticeable problem in the UK Subtitling Audiences Network survey [15]. However, the first UK subtitling guidelines issued by the Independent Broadcasting Authority in 1981, stipulated a maximum word rate of only 120 words per minute (wpm) and gave extensive advice on editing subtitles [17,18]. This was to some extent based on research at the University of Southampton [19], but mostly followed the approach taken by the WGBH Captioning Centre in the USA [20, p85]. The recommended maximum word rate of subtitles in the UK gradually increased in response to audience feedback, and verbatim subtitles became the norm in the early 2010s. However, it wasn't until 2024 that Ofcom updated their guidelines to say, "*In general, subtitles should be synchronised with the audio, and reflect the speech verbatim, as closely as possible*." [21] As a result, subtitles produced before 2010 are often heavily edited. Problems still occur with live subtitles produced by respeaking. Most re-speakers cannot achieve rates over 180wpm and many only reach 160wpm. However, free-flowing speech in live programmes can often reach speeds over 200wpm and this work has discovered examples of speech up to 290wpm.

THE STRUCTURE OF SUBTITLES

Subtitles are not structured data; they specify an arrangement of text on screen. There is no one-to-one correspondence between the subtitles delivered by the broadcaster and the speech content. A subtitle may be repeated in the stream but visible only once to the viewer. Non-speech content is represented in subtitles in a wide variety of forms, including content warnings, copyright notices, sound effects and speaker identification along with other conventions that indicate sound effects. In order to accurately correlate the output of a speech-to-text engine with subtitles, repeats need to be removed along with non-speech elements. However, subtitling conventions vary between broadcasters and change over time, so the process is imperfect. Another issue is the difference between block and snake subtitles. Block subtitles are sent as one or more lines of text and are replaced by the subsequent subtitle, while snake subtitles are sent many times with words being added each time. Further problems are caused by "reverse snake" subtitles where the last word of the previous subtitle is removed to correct errors.

THE SPEECH-TO-TEXT ENGINE

Whisper from OpenAI is the leading speech-to-text engines for the English language [22]. It is simple to install under python and runs on the local machine. It was used by the BBC for its 2024, trial of subtitle for radio [23], part of a drive to use generative AI [24], though the output was manually edited before publication. Whisper is not without its problems [25]. It can produce nonsense or no output at all. It is noticeably poorer on female voices, especially those with accents. Spelling is variable between USA and UK variants. It also has a cold start problem, especially in the presence of music. This can be overcome, to some extent, by providing an initial prompt. The least worst approach to this problem seems to be to use the first line of the first subtitle as the prompt which tends to cause the engine to start with the following line. The engine also produces some sound effect labels. The main problem with Whisper is that the word timings drift, so a modified version called *whisper-timestamped* is used to overcome this [26]. However, where a loud sound effect runs into a spoken word, the start time for the word is given as the start of the sound effect, so if a word duration is over-long then it is probably in error, but the word count is generally reliable. BBC Kaldi speech-to-text engine has been used for previous subtitle matching projects [27] and could have been used in place of Wisper, had it been available.

THE WORKFLOW

All the software for this project is written in python 3 and runs under Ubuntu on desk-top PCs. The source of test material for this work is transport stream recordings from UK DSAT Freesat services. This is a pragmatic choice as these services still carry subtitles as Teletext alongside DVB subtitles, thus avoiding the need to use optical character recognition to recover text from DVB subtitle images. The test recordings used a USB DSAT receiver on a communal antenna feed, so packet errors were not uncommon. If an off-air transport stream recording has errors at the start, this can cause the demultiplexed audio and subtitle files to start at different times and packet loss during recordings can both lead to errors in the measurement of subtitle timings.

The main audio track and Teletext subtitles are extracted from the transport stream using *ffmpeg* to produce a **`.wav`** file for the audio. Timing issues are largely overcome by instructing *ffmpeg* to resync the audio to the presentation time stamps by replacing missing sections with silence. The subtitles are extracted by *ffmpeg* and rendered as a **`.srt`** file. While the **`.srt`** file lacks the colour and positional information, it carries sufficient information for the work described here. The subtitle file is parsed to separate out as much of the non-speech content as possible and identify subtitles flagged as music or sound effects. It is also at this point that snake subtitles are detected and each additional word is saved as a separate subtitle with the repeated words discarded. The result is saved as structured data in a **`.json`** file. The **`.wav`** file is then passed to the speech-to-text engine along with the first line of the first subtitle as the prompt. The transcript is saved as a separate **`.json`** file. The next stage takes the **`.json`** files for the subtitles and transcript and create two lists of words. Each item in the list is a spoken, or sung, word along with any timing information and a flag to indicate whether the subtitle word is part of a subtitle block or snake. Fully capitalised words are not included in the transcript list as they are likely to be sound effects. This approach can be extended to ingest other subtitle formats, converting them to the same structured data format as all subsequent processes use the **`.json`** representation of the subtitles.

Word Loss Estimation

Initial comparisons are made between the transcript and the subtitles. The length of the transcript word list gives an estimate of the number of words spoken and this is compared to the number of spoken or sung words in the subtitles to give an estimate of the number of words missing from the subtitles. Where the recording is of known-good, pre-prepared subtitles with a straightforward speech content, the two numbers agree to within ± 1%. This also serves to demonstrate the validity of the approach. With drama content and reality TV the disparity can be as high as 5% because there are many non-speech utterances. Differences are also caused by commercial breaks and programme trails which often lack subtitles. If the number of missing words in the subtitles exceeds 10% then this suggests significant problems with the subtitles. With archive programmes, subtitled in the early 2000s or before, and some live programmes, the subtitles may be missing up to 30% of the words spoken. If the number of words in the subtitles exceeds the number in the transcript this usually indicates problems with subtitles where whole sections of subtitles are repeated or snake subtitles which could not be parsed correctly. Another comparison made at this stage is to measure the word frequency of each word on the two lists. Large discrepancies in the words found in the transcript and subtitles are an indication that the programme was broadcast with the wrong subtitles.

The Alignment Challenge

Each word in the subtitles should correspond to a word in the transcript, and it should be possible, with good subtitles and an accurate transcript, to tag each word in the subtitle list with the corresponding word in the transcript and vice versa. However, the two can differ from each other in a number of ways: -

- There can be spelling differences between the subtitles and the transcript.
- Compound words and contractions may appear as separate words
- There can be errors in the words in both the subtitles and transcript.
- The subtitles may not contain all the spoken words.
- The words in the subtitles may be in a different order to the words spoken.
- Sections of subtitling may be repeated.
- The timing of the subtitles may not match the timing of the speech.
- The subtitles may contain non-speech utterances not transcribed.
- The transcript may contain non-speech utterances subtitled as a sound effect.
- There may be a lot of repetition in the speech content, especially in songs.

Additionally, the software needs to respond appropriately to situations where: -

- The recording does not contain a subtitle stream
- The subtitle stream is empty
- There is no speech in the programme
- The programme has been broadcast with the wrong subtitles.

The Alignment Process

Simple approaches to alignment, such as stepping through the subtitles looking for matches in the transcript can work on good quality subtitles and transcript. However, where there are problems with subtitles these approaches don't work, at best resulting in false matches and at worst failing completely. Initial attempts at alignment revealed subtitle delays of up to 50 seconds and subtitles omitting up to 30% of words. Techniques from natural language processing were then used to overcome this. This involves looking for long strings of words, called n-grams, which occur only once in both the subtitles and transcripts, starting from the longest strings and working downwards. The first pass takes sections of subtitles and transcript and subtitles in overlapping 240 second sections, at 120 second intervals and looks for these unique matches. In some cases, it is possible to find matching n-grams more than 200 words long, so the matching starts with 250-grams and works sequentially downwards to 20-grams. This results in sections with matches and these break up the transcript and subtitles into short sections of unmatched words. The process is then repeated to fill in the gaps, progressively matching n-grams from a minimum length of 20 down to a minimum of 3, reducing the size of the gaps each time.

A final, pass attempts to fill the remaining gaps by matching individual words. To increase the number of matches, words are checked for differences in spelling, numerals, compound words and contractions. Contractions are detected by the presence of an apostrophe and compound words by pairs of words match to a single word. Common numerals are matched by a look up table and Levenshtein distance is used to overcome most of the differences between US and UK spelling. Matches are not allowed for the most common 20 words, because these often cause false alignments. Each match is checked to see if it has caused a sequence error. Some of these sequence errors indicate false alignments between the subtitles and transcript, but others highlight where the subtitles contain the right words but in a different order to the speech.

SUBTITLE TIMING MEASUREMENT

Once the subtitles have been aligned with the transcript, the start time for each subtitle can be compared with the begin time for the first word of the subtitle in the transcript. Subtracting the transcript timing from the subtitle timing gives a measure of the subtitle delay. Negative values indicate that the subtitle appeared early. A scatter plot of delay against the start time of the subtitle is then created, and the subtitling mode, snake or block, is indicated by an overlying set of orange marks, zero for block and 10 for snake (figure 1). The average subtitle delay for each minute is then calculated and rendered as a line plot (figure 2). Note this example starts with a live programme, followed by a pre-recorded programme and both contain commercial breaks, clearly seen in subtitle timings for the live programme.

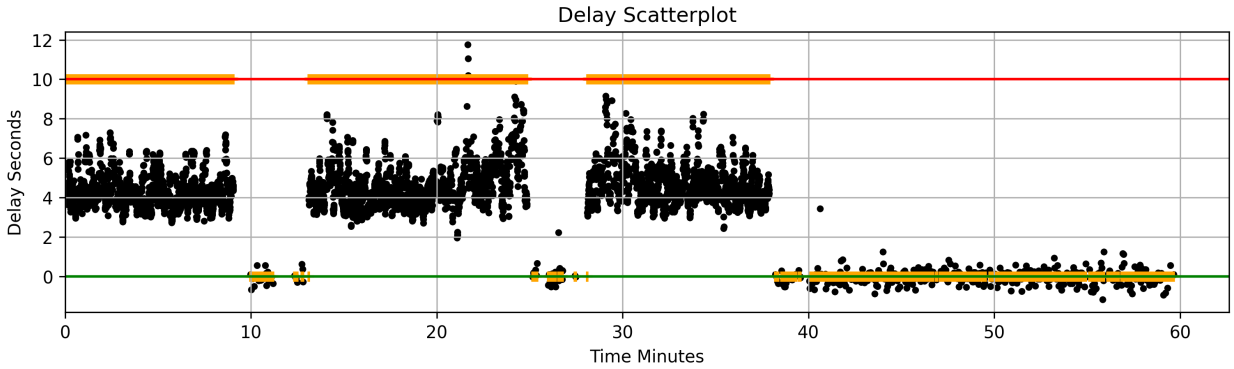

Figure 1 – Scatter plot of individual subtitle delay.

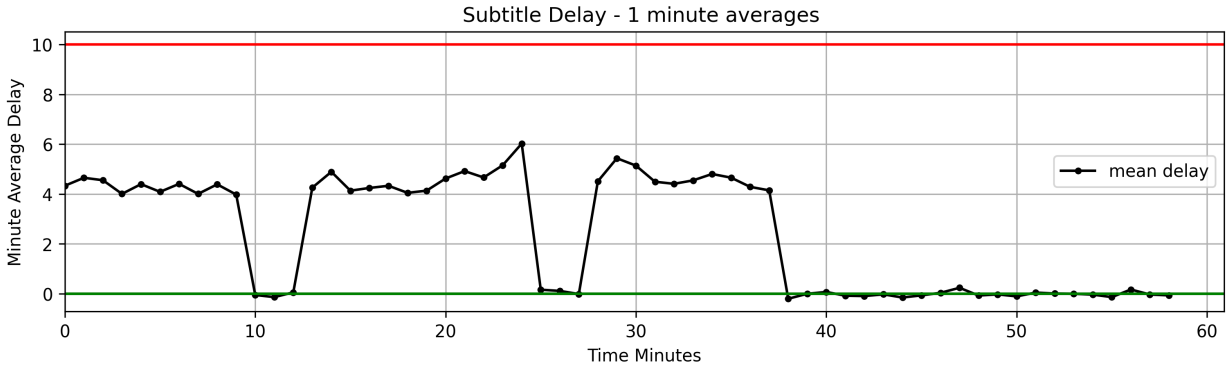

Figure 2 – Subtitle delay - average delay for each minute.

The mean subtitle delay for the recording is calculated along with the variance, skew and kurtosis. A high kurtosis value is likely to indicate errors in the alignment process resulting in outliers, while the variance gives an indication of the spread of the delay measurements. Note that the mean subtitle delay can be misleading, as for some material there can be a bimodal spread in timings, so by plotting a histogram of subtitle delays the nature of the distribution can be seen (figure 3). A further refinement is made by separately plotting histograms for block and snake subtitles. Because there is a snake subtitle for each individual word, the distribution is weighted by the number of words in each subtitles to reflect this (figure 4).

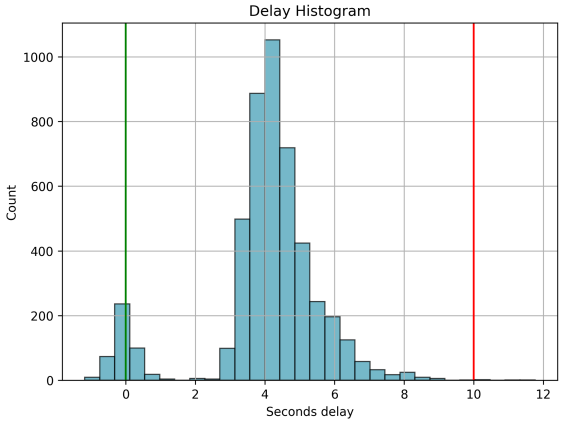

Figure 3 – Overall Delay Histogram

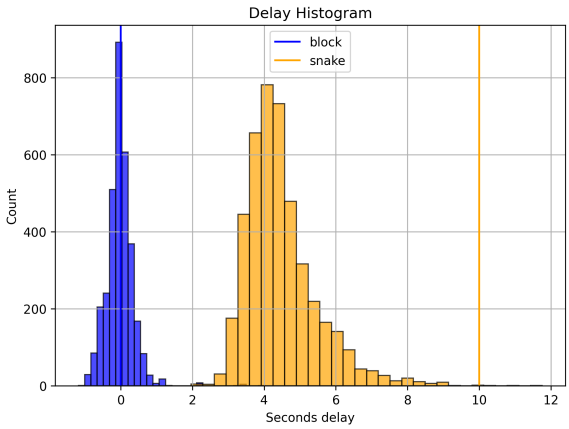

Figure 4 – Split Delay Histogram

WORD RATE DIFFERENCES

One further plot is produced showing the number of words in each minute for the subtitles and the transcript. This can reveal changing patterns in the number of words missing from the subtitles (figure 5). If the number of words in the subtitles exceeds the number in the transcript this can indicate a range of different issues. It is sometimes indicative of people talking over each other which causes the speech-to-text to fail to recognise words. It can also occur with high levels of background noise or if the subtitles contain non-speech utterances, common in children's programming. However, it can also indicate that there is a problem with the subtitles repeating strings of words in a fault condition, or that the snake subtitles are not scrolling up correctly, so the de-repeating stage for snake subtitles has failed. Usually, the number of words in the transcript will exceed the number of words in the subtitles, especially on live subtitles and archive content.

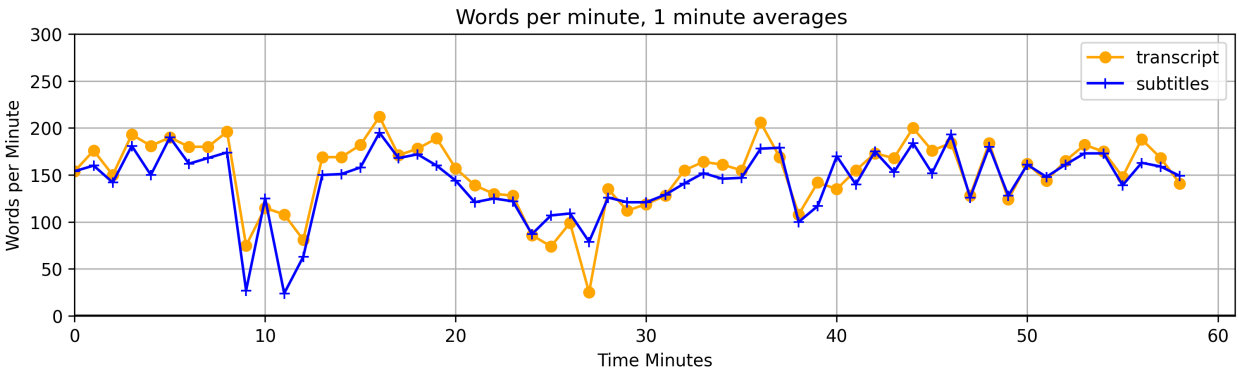

Figure 5 – Words per minute comparison between transcript and subtitles

PERFORMANCE AND LIMITATIONS

Running the processing software on a domestic "games" spec PC the processes of demultiplexing and speech-to-text take around 5 minutes for a one-hour recording, while the alignment and measurement can take anything from 30 seconds for high quality subtitles up to 4 minutes for content which is live and archive content with heavily edited subtitles. Whisper-timestamped takes advantage of the games-PC's graphics card for processing so for the speech-to-text the CPU load is negligible, while the alignment runs as a single thread and only utilises one CPU, leaving room for optimisation.

The accuracy of the measurements obtained were judged by running the software against high value content with clear narration, where it can be safely assumed that the subtitler has been given sufficient time to prepare high quality subtitles. The results demonstrated high levels of accuracy in the measurement of both word rates and timing. The word rates match to within ±1%, while the subtitle timings show a spread of around ±1 second. Visual inspection of the subtitles confirmed that there was indeed such a variation in the subtitle timings, even in such high value content. The software also detected the occasional word sequence reversal in this type of material. The software was able to align around 98% of the words in the subtitles with the words in the transcript.

As the subtitles depart from the speech content the process of alignment becomes more complex, especially where there is a great deal of repetition, or where the soundtrack is more challenging, with high levels of background noise or singing. Here the results become indicative of the character of the subtitles, while still usefully showing where human review is required. Also, while the initial intention of this work was to establish that information about subtitle delay and word omission can be automatically obtained using current speech-to-text technology, during testing it become clear that the results can sometimes affected by additional problems with broadcast subtitles, especially, but not exclusively, live subtitles. Further work is required to characterise these additional problems and flag them up for the user.

RESULTS

While most of the recordings featuring pre-recorded subtitles were within 2% of the transcript in terms of number of words and within ±2 seconds in terms of timing, the recordings containing live subtitles were far more variable. There is not room here to illustrate all the issues, but the following four examples illustrate just how far the subtitles have been found to depart from the speech content. In the first example (figure 6), the plot shows a section of 20 words over 60 seconds behind the speech with the delay gradually reducing to around 15 seconds over a period of a minute. This was verified by inspection of the recording and was caused by the subtitles freezing for 40 seconds before continuing from where they had left off. The one-minute average delay was over 30 seconds for a period of 3 minutes. In the second example (figure 7), pre-prepared live subtitles were played out too fast, overtaking the speech. At the worst point they are nearly a minute ahead of the speech before a 2 minute section of the subtitles is repeated, resulting in the subtitles becoming around minute late.

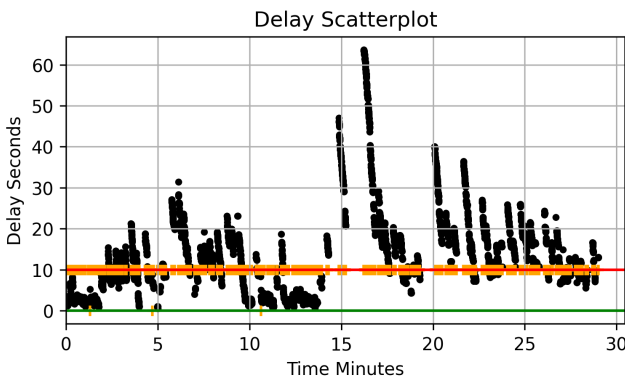

Figure 6 – Subtitles over 60 seconds late

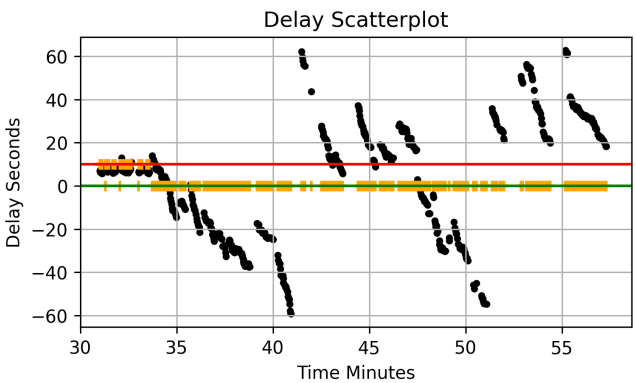

Figure 7 – Subtitles 60 seconds early & late

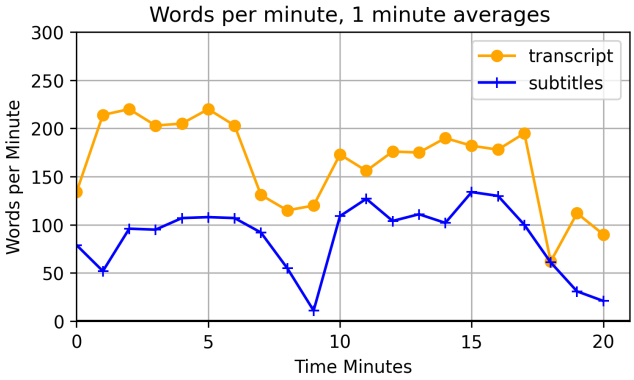

Figure 8 – Subtitles with around 50% word loss

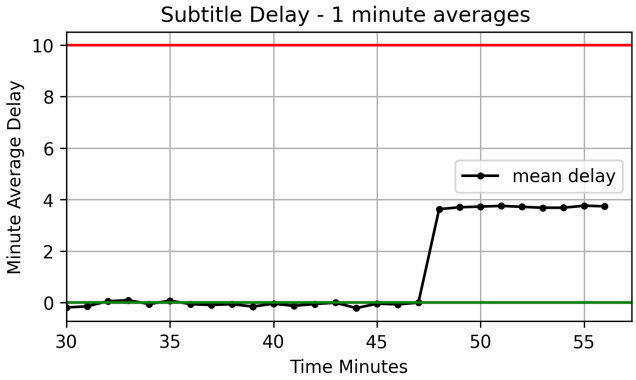

Figure 9 – Edited video with original subtitles

In the third example (figure 8), durring the first 6 minutes of a live discussion/interview, the subtitles lack around half the spoken words, although the subtitles did improve towards the end of the programme. The final plot (figure 9) is of a pre-recorded programme where the audio-visual content had been edited but the subtitles had not been updated. Not only is the new version missing around 4 seconds of video, causing the last section of subtitles to be late, but on inspection, it was found that some sections had been revoiced, so the subtitles did not match the content at the point where the edit took place.

These first three examples demonstrate the need for automated subtitle monitoring to provide feedback to the broadcasters and subtitle providers so issues like these can be identified and improvements made to subtitling systems. The fourth example demonstrates the need for automated quality control, so these errors can be detected and corrected before broadcast.

DISCUSSION

This work was conceived as an exercise in demonstrating automated subtitle quality monitoring. The expectations were that most subtitles would be no more than 10 seconds late and that word loss would be minimal, except for archived content. However, it became clear that the range of timing errors extended well beyond acceptable limits and the combination of word loss and errors made the task of measuring timing challenging. Also, with subtitle timing errors of anything up to a minute, some measurements, like the word rate in any one minute, can become out of sync, requiring some human interpretation. The main thing that has become clear is that while the accuracy of the measurements tend to decline as the quality of the subtitles decline, the results are still a good indication of quality failures that require attention.

This work has highlighted a number of quality problems in UK TV subtitles, most of which would appear to be caused by technical limitations, technical faults and process failures. The use of pre-prepared subtitles for live playout has been seen as a way of improving subtitle quality, both in terms of accuracy and timing [28]. However, while it undoubtably reduces word errors, in practice it is causing its own problems with timing. The main issue being live subtitles appearing early, in one example even overlaying the preceding commercial break. Where subtitles freeze and resume leading to excessive delays, it would appear to be caused by some kind of buffering issue in the subtitling system, or downstream in the broadcast chain, which suggests some form of technical fault.

Further work

The techniques outlined here could be developed into both systems for 24/7 monitoring of television subtitle quality to detect problems during broadcast and systems for subtitle quality control to check subtitles before broadcast and/or uploading onto a streaming service, giving time for faults to be corrected.

The software, as described here, is a proof of concept. Any system in long-term use would need to be more robust and maintainable. It would need to create reports for the user with the main indicators such as mean delay and word loss, alongside data on subtitle anomalies and key charts for each sample. Additional information about colour and position could be obtained by directly decoding the Teletext subtitles or by decoding the DVB subtitles and using character recognition to recover the text. The latter would enable the off-air monitoring of UK DTT services. Other subtitle formats could be added to the system as required. EPG data could be used to segment the transport streams into individual programmes. It would also be possible to use face recognition and text detection to determine whether any subtitles are obscuring faces and text in the TV image, another fairly common problem with live subtitles.

Because this approach enables the use of samples of much greater lengths than previously it could be used to create baseline data on subtitle quality in UK television broadcasting. It would enable a review of some key aspects of subtitle quality across a wide range of channels and broadcasters and provide useful information about the effectiveness of different approaches to live subtitle production. Given access to archive content it could also be used to examine historic trends in UK television subtitling practice and provide improved understanding of the issues for practitioners, researchers and audiences.

CONCLUSIONS

This work has demonstrated the use of generative-AI speech-to-text to automate aspects of subtitle monitoring and quality control and highlighted examples of quality failure. While such a system cannot identify individual word errors and omissions and the measurements cannot be 100% correct, this approach can usefully quantify subtitle timing on a minute-by-minute basis along with highlighting word loss in subtitles. Where the measurements fall outside the broadcaster's guidelines, the content can then be manually reviewed to verify the measurements and appropriate action taken.

Previous reviews of subtitle quality have been restricted to 10-minute samples of television programming by the cost of manual measurement. While not a replacement for human inspection, automation can effectively highlight the parts of a programme that require further attention and shows the potential for 24/7 monitoring, something that would otherwise be too expensive. This paper has avoided any mention of the broadcasters and programmes involved so as not to distract from the main message of the paper.

While individual word errors are beyond the scope of this work, it has demonstrated the detection and measurement of word omission and timing. Examples of large errors in subtitle timing and significant word loss have been detected amongst the recordings made and a few examples of these have been shown. These examples point to the need for automated subtitle quality monitoring to detect and rectify problems and thus improve the audience's experience.

ACKNOWLEDGEMENTS

The work has been carried out as an independent research project with the support and encouragement from Dr Michael Crabb, Head of Computing at the University of Dundee and Dr Carol O'Sullivan from the University of Bristol and the UK Subtitling Audiences Network[1].

REFERENCES

[1] Subtitle Quality: Measuring and improving subtitle quality. BBC R&D Project Page. January 2012. https://www.bbc.co.uk/rd/projects/live-subtitle-quality

[2] Armstrong, Michael. "The development of a methodology to evaluate the perceived quality of live TV subtitles." In IBC2013 Conference, pp. 11-1. London UK: IET on behalf of IBC, 2013. https://digital-library.theiet.org/doi/abs/10.1049/ibc.2013.0044

[3] Ware, Trevor (BBC), Matt Simpson (Ericsson). Live subtitles re-timing proof of concept. BBC R&D White Paper WHP 318 https://www.bbc.co.uk/rd/publications/whitepaper318

[4] Ofcom. Measuring live subtitling quality: Results from the first sampling exercise. April 2014. https://web.archive.org/web/20210419114907/https://www.ofcom.org.uk/research-and-data/tv-radio-and-on-demand/tv-research/live-subtitling/sampling-results

[5] Ofcom. Measuring live subtitling quality: Results from the second sampling exercise. November 2014. https://web.archive.org/web/20210419114907/https://www.ofcom.org.uk/research-and-data/tv-radio-and-on-demand/tv-research/live-subtitling/sampling-results-2

[6] Ofcom. Measuring live subtitling quality: Results from the third sampling exercise. May 2015. https://web.archive.org/web/20210704063409/https://www.ofcom.org.uk/research-and-data/tv-radio-and-on-demand/tv-research/live-subtitling/sampling-results-3

[7] Ofcom. Measuring live subtitling quality: Results from the fourth sampling exercise. November 2015. https://web.archive.org/web/20210419114907/https://www.ofcom.org.uk/research-and-data/tv-radio-and-on-demand/tv-research/live-subtitling/sampling_results_4

[8] Romero-Fresco, Pablo. "Accessing communication: The quality of live subtitles in the UK." Language & Communication 49 (2016): 56-69. https://www.sciencedirect.com/science/article/abs/pii/S0271530916300398

[9] Bason, Samuel & Michael Armstrong. TVX2014 Short Paper - Subtitle Monitoring. https://www.bbc.co.uk/rd/blog/2014-10-tvx2014-subtitle-monitoring

[10] Sandford, James. "The impact of subtitle display rate on enjoyment under normal television viewing conditions." In *IBC 2015*. IET, 2015. https://www.bbc.co.uk/rd/publications/whitepaper306

[11] Apone, T., B. Botkin, M. Brooks, and L. Goldberg. "Research into Automated Error Ranking of Real-time Captions in Live Television News Programs." *The Carl and Ruth Shapiro Family National Center for Accessible Media at WGBH (NCAM)* (2011). http://ncamftp.wgbh.org/ncam-old-site/file_download/CC_Metrics_research_paper_final.pdf

[12] Quality Control: Helps to optimise the use of automated quality control systems. EBU, 2017- 2025. https://tech.ebu.ch/qc

[1] https://uksubtitlingaudiences.wordpress.com/

[13] 0110B - Subtitles Alignment v1.0. Frans De Jong, 2014. https://qc.ebu.io/items/0110B/versions/1-0-0/

[14] Rane, Aditi. BBC Look North's subtitle mix-up has Peter Levy and fans in hysterics. Grimsby Live, April 2022. https://www.grimsbytelegraph.co.uk/news/celebs-tv/bbc-look-norths-subtitle-mix-6928975

[15] UK Subtitling Audience Survey. UK Subtitling Audiences Network. Awaiting publication.

[16] Szarkowska, Agnieszka, and Olivia Gerber-Morón. "Two or three lines: A mixed-methods study on subtitle processing and preferences." Perspectives 27, no. 1 (2019): 144-164. https://discovery.ucl.ac.uk/id/eprint/10054421/3/Szarkowska_Two%20or%20three%20lines_a%20mixed-methods%20study%20on%20subtitle%20processing%20and%20preferences_plain_FINAL.pdf

[17] Baker, Robert G. Guidelines for the Subtitling of Television Programmes. Independent Broadcasting Authority, 1981.

[18] Baker, Lambourne, Rowson. Handbook for Television Subtitlers, Independent Broadcasting Authority, 1982.

[19] Baker, Robert G., and Alan F. Newell. "Teletext Subtitles for the Deaf: Problems in Linguistics and Psychology." In Proceedings of the International Broadcasting Convention, pp. 97-100. 1980.

[20] Downey, Gregory J. Closed captioning: Subtitling, stenography, and the digital convergence of text with television. JHU Press, 2008.

[21] Ofcom's Guidelines on Providing Television and On-Demand Access Services. Ofcom July 2024 https://www.ofcom.org.uk/tv-radio-and-on-demand/accessibility/tv-access-services/

[22] Kuhn, Korbinian, Verena Kersken, Benedikt Reuter, Niklas Egger, and Gottfried Zimmermann. "Measuring the accuracy of automatic speech recognition solutions." ACM Transactions on Accessible Computing 16, no. 4 (2024): 1-23. https://dl.acm.org/doi/pdf/10.1145/3636513

[23] Dimri, Aniruddh . Using Gen AI to add subtitles on BBC Sounds. BBC Media Centre, August 2024. https://www.bbc.co.uk/mediacentre/2024/using-gen-ai-to-add-subtitles-on-bbc-sounds

[24] Talfan Davies, Rhodri. An update on Generative AI (Gen AI) at the BBC. BBC Media Centre. January 2025. https://www.bbc.com/mediacentre/2025/articles/update-generative-ai-at-the-bbc

[25] Teka Hadgu, Asmelash & Timnit Gebru. "Replacing Federal Workers with Chatbots Would Be a Dystopian Nightmare." Scientific American, April 2025. https://www.scientificamerican.com/article/replacing-federal-workers-with-chatbots-would-be-a-dystopian-nightmare/

[26] https://github.com/linto-ai/whisper-timestamped

[27] Armstrong, Michael. "Automatic recovery and verification of subtitles for large collections of video clips." SMPTE Motion Imaging Journal 126, no. 8 (2017): 1-7. https://www.bbc.co.uk/rd/publications/whitepaper323

[28] Armstrong, Michael, Andy Brown, Michael Crabb, Chris J. Hughes, Rhianne Jones, and James Sandford. "Understanding the diverse needs of subtitle users in a rapidly evolving media landscape." SMPTE Motion Imaging Journal 125, no. 9 (2016): 33-41. https://downloads.bbc.co.uk/rd/pubs/whp/whp-pdf-files/WHP307.pdf