Using generative-AI speech-to-text output to provide automated monitoring of television subtitles

Michael Armstrong
Associate staff - University of Dundee
& member of the UK Subtitling Audiences Network

13/09/2025

Speech-to-text and AI

- Unlike many other uses of machine learning and generative AI, speech-to-text is a well understood and well bounded problem and its performance can be objectively measured.

- However, problems of bias in the selection of training and test material still exist, along with the questions around the ownership of the training material used to build these systems.

The problem with subtitle quality

A survey by the *UK Subtitling Audiences Network has* highlighted problems the audience has with subtitle quality.

- Top of the list was **delay** with 2/3rds of the respondents selecting this as the worst problem - consistent with a 2012 RNID survey.

- Second was **subtitles do not accurately reflect what is said**. Deaf and hard-of-hearing people have been consistent in expressing their preference for verbatim subtitles.

Manual subtitle quality monitoring

Manual surveys are expensive and time consuming so only examine short samples and have focused on word errors.

- One notable example is an exercise in monitoring live subtitle quality run by Ofcom and the University of Roehampton in 2014/15. 10-minute clips of live subtitles, were tested at 6 month intervals over a period of two years.
- The extremely sparce sampling, questionable methodology and the high cost to the broadcasters, meant the exercise was ultimately counterproductive.

Previous work on automatic monitoring

In response to the Ofcom survey, at BBC R&D I commissioned a 6-month, trainee project which successfully demonstrate 24/7 monitoring of DSAT Teletext subtitles. It measured subtitle word-rate, position and mode, i.e. snake or block (live vs prerecorded).

In the USA, the Media Access Group at WGBH had run a 3-year research programme, ending in 2011, which compared the output of a speech-to-text engine to live subtitles to gauge word accuracy.

This project

This work combines the BBC R&D approach with the WGBH project's use of speech to text technology.

- It is a proof-of-concept which demonstrates the viability of automated subtitle quality monitoring.
- It uses *Whisper*, OpenAI's speech-to-text engine, which is currently the industry leader for accuracy. It was *"trained on 680,000 hours of multilingual and multitask supervised data collected from the web".*
- This work uses a modified version which gives more accurate timings called *whisper-timestamped.*

The workflow

- This project is written in *python 3* and runs under *Ubuntu* on domestic grade, desk-top PCs, off-line on the local machines.
- The source of test material is transport stream recordings from UK Freesat services, which carry Teletext subtitles.
- The recording are made using a USB DSAT receiver.
- The main audio track and Teletext subtitles are extracted using *ffmpeg* to a **.wav** file for the audio and subtitles as a **.srt** file.
- The audio is then passed to *Whisper* to produce a transcript.

Subtitles are not structured data

Subtitles describe how text should be displayed on a screen.

They contain non-speech elements and repetition which need to be removed to leave just the speech.

The **.srt** subtitle file is converted into a **.json** structured data format where repeats, as with snake subtitles, are removed and different components are stored separately.

This process is largely successful, but not 100% reliable.

Alignment

To measure timing, the transcript output by *Whisper* has to be aligned to the speech content of the television subtitles.

- This is straightforward with high quality, pre-prepared subtitles and a clear speech soundtrack.
- However, as the subtitle and audio quality decline, the difficulty of aligning the transcript to the subtitles increases.
- Techniques from natural language processing are used to improve the accuracy of the alignment.

Things that make alignment difficult - 1

- The timing in the subtitles and transcript may not match.
- The subtitles may omit many of the spoken words.
- Word errors in both the subtitles and the transcript.
- Spelling differences between the subtitles and the transcript.
- Compound words and contractions vs as separate words.
- The words in the subtitles can be in the wrong order.
- Long sections of subtitles can be repeated.

Things that make alignment difficult - 2

- The subtitles include non-speech utterances not transcribed.
- The transcript include non-speech utterances not subtitled.
- Speech content or singing which contains a lot of repetition.

Also, the software needs to cope with...

- Channels with no subtitle stream.
- Programmes with no subtitles.
- Programmes that do not contain speech.
- Programmes broadcast with the wrong subtitles.

Alignment 2

- To improve the chance of correct alignment the software first looks for long n-grams, that occur only once in both the subtitles and transcripts, starting from the longest and working downwards.
- The first pass takes sections of subtitles and transcript in overlapping 4 minute samples, at 2 minute intervals and looks for these matches. It starts with 250-grams and works downwards to 20-grams, which leaves unmatched gaps.
- The process is repeated to within the gaps, matching n-grams from a minimum length of 20, then progressively reducing the minimum to 3, reducing the size of the gaps each time.

Alignment 3

- A final, pass attempts to match the remaining words by checking for differences in spelling, numerals, compound words and contractions.
- Matches are not allowed, at this stage, for the most common 20 words to avoid false alignments.
- Each match is checked for sequence errors, indicating a either a false alignment, or that words in the subtitles are in a different order to the transcript.

Calibration

The accuracy of the system be judged by the output it gives on known high quality subtitles from pre-recorded, factual programmes with a clear speech content. In these cases:-

- The word count differences are less than ±1%.
- The timing measurements are within ±1 second.
- The word alignment is above 97%.

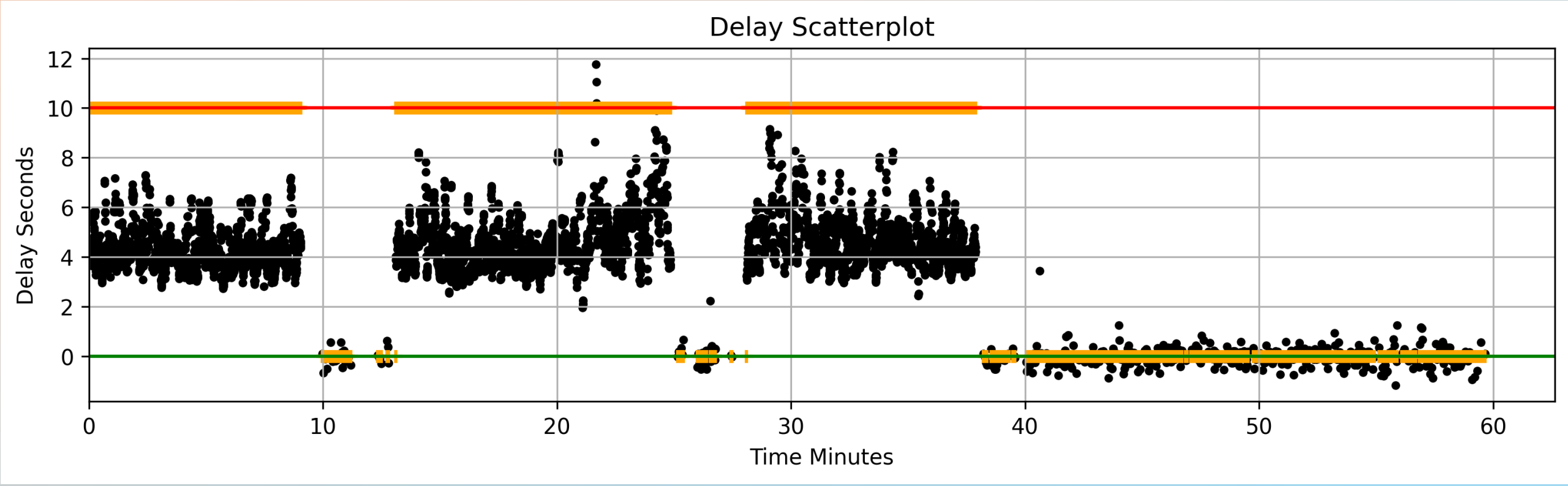
The results

The system can measure the timing of a high proportion of the subtitles in a recording and gives a reliable indication of whether subtitles are delayed or early.

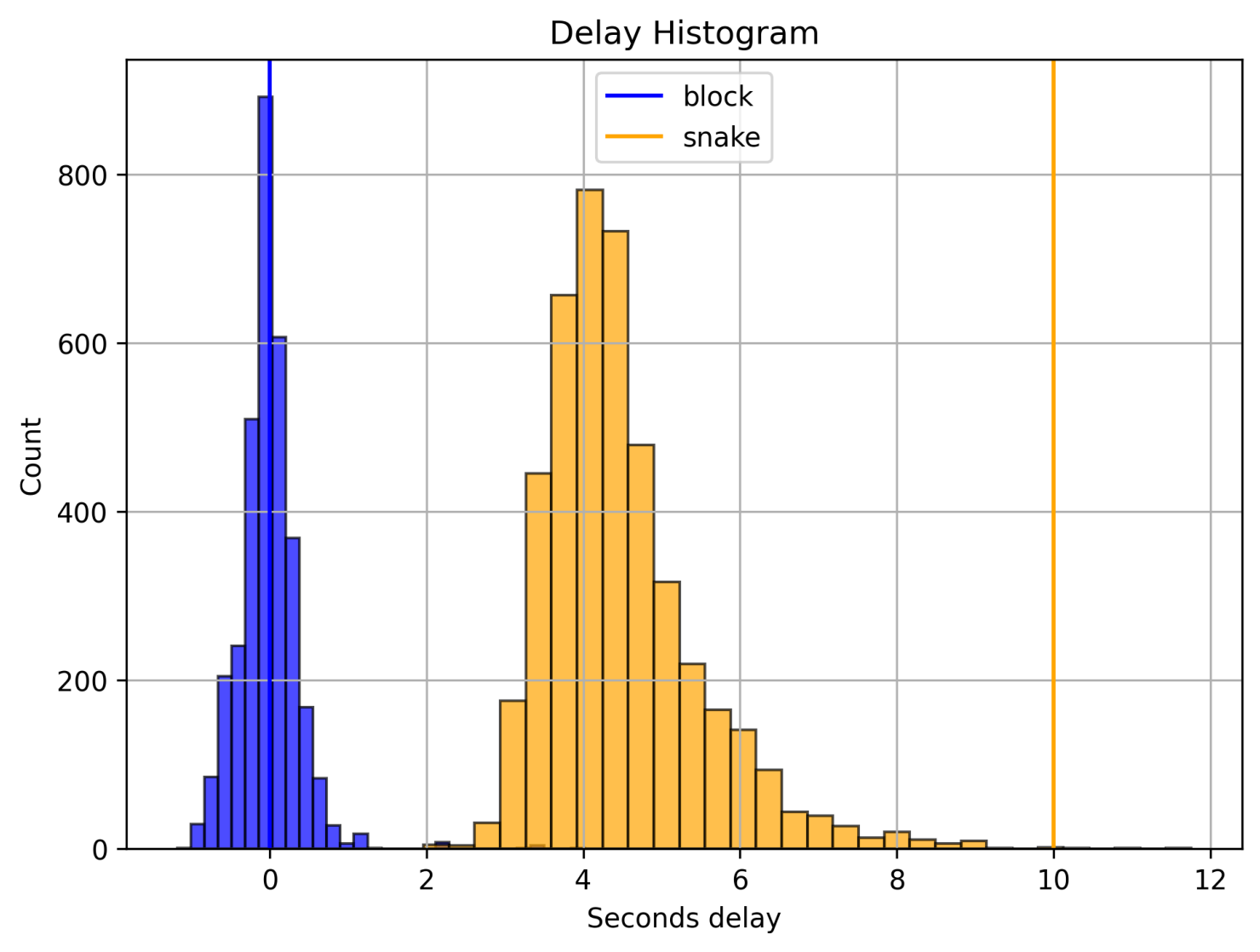
It also gives a good indication of the proportion of spoken words which have been omitted from the subtitles, provided the audio mix is of a reasonable quality.

The results are plotted against a time-line in a series of graphs.

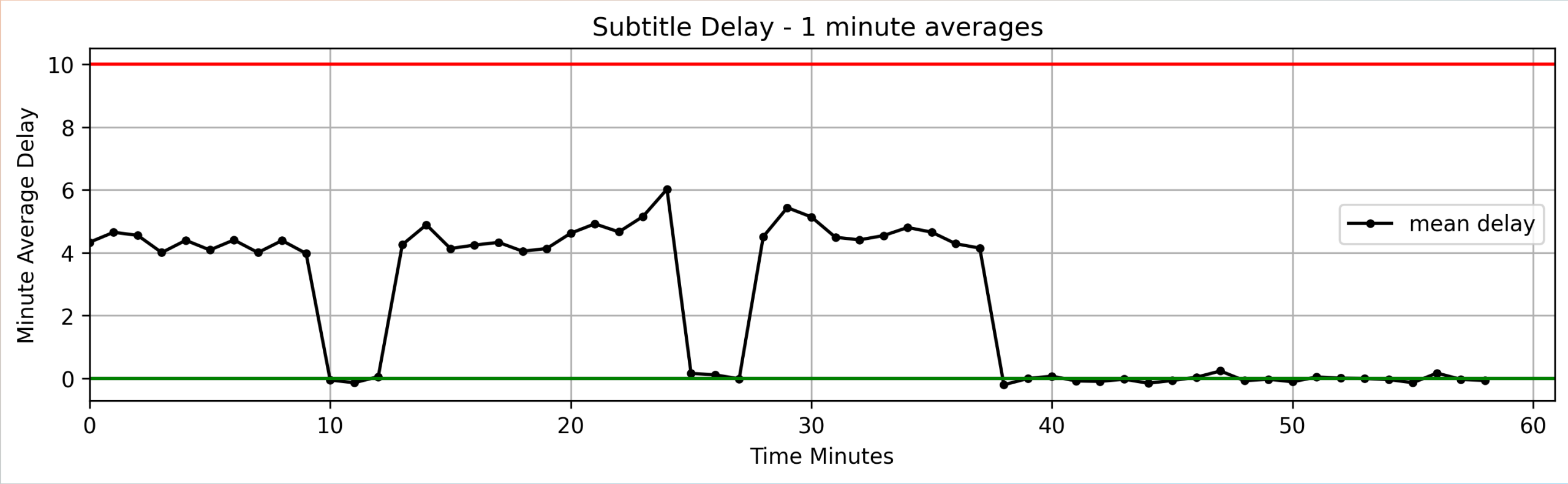
Timing - scatter plot



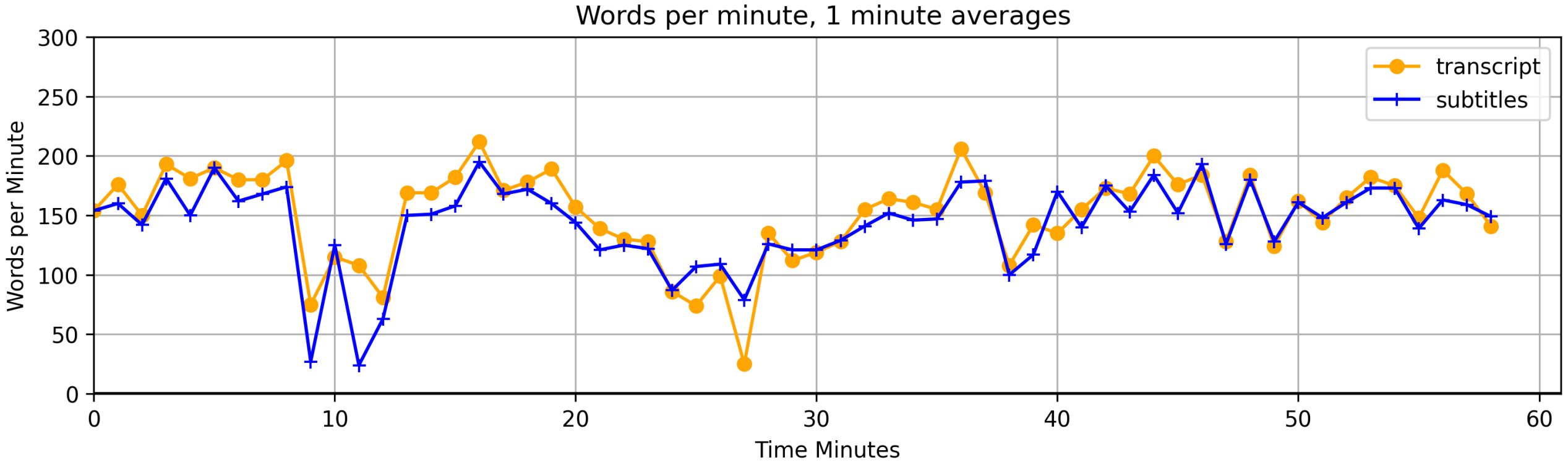
Timing - histogram



Timing - one minute-average



Words in each minute

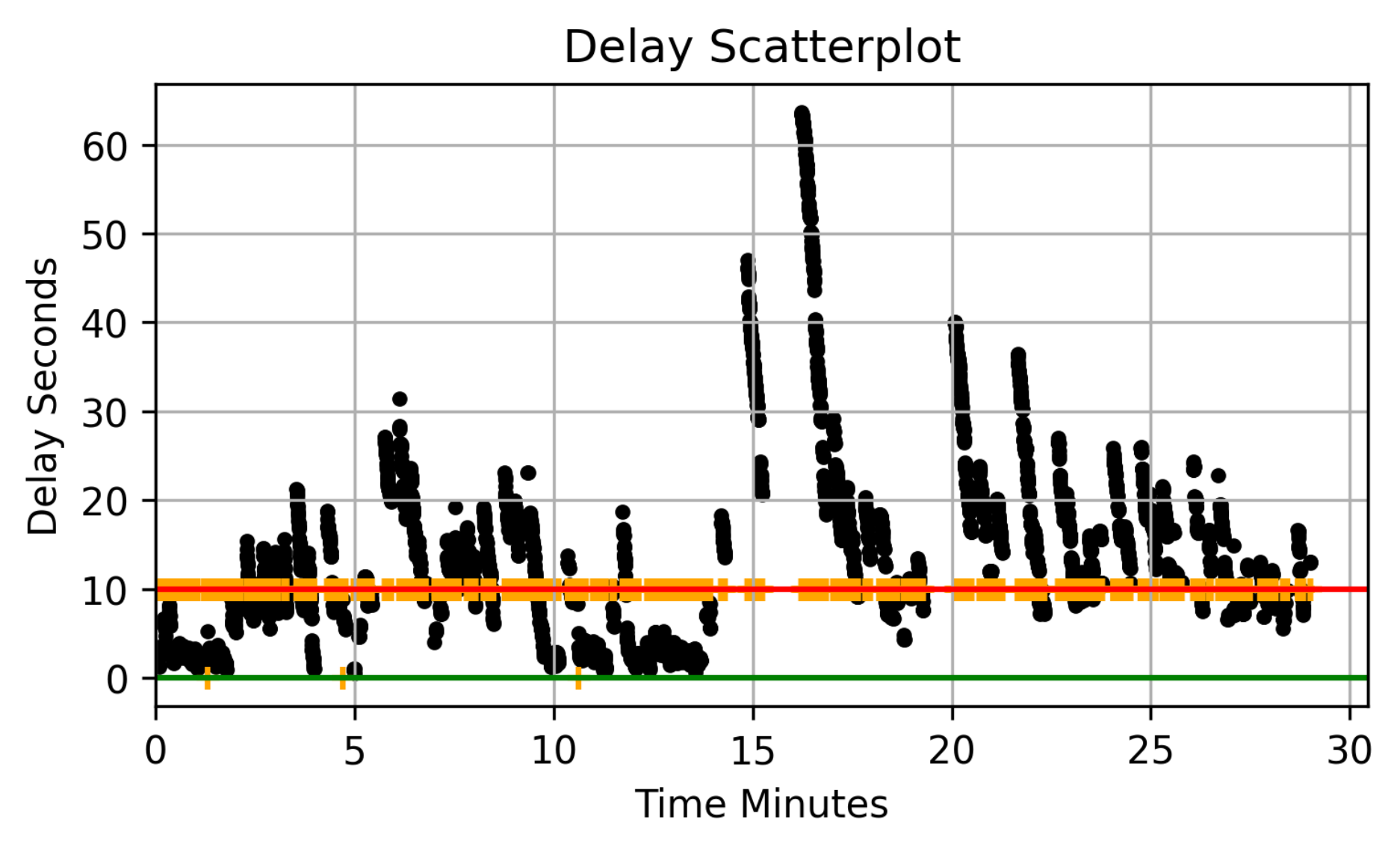


Findings - now for the bad news

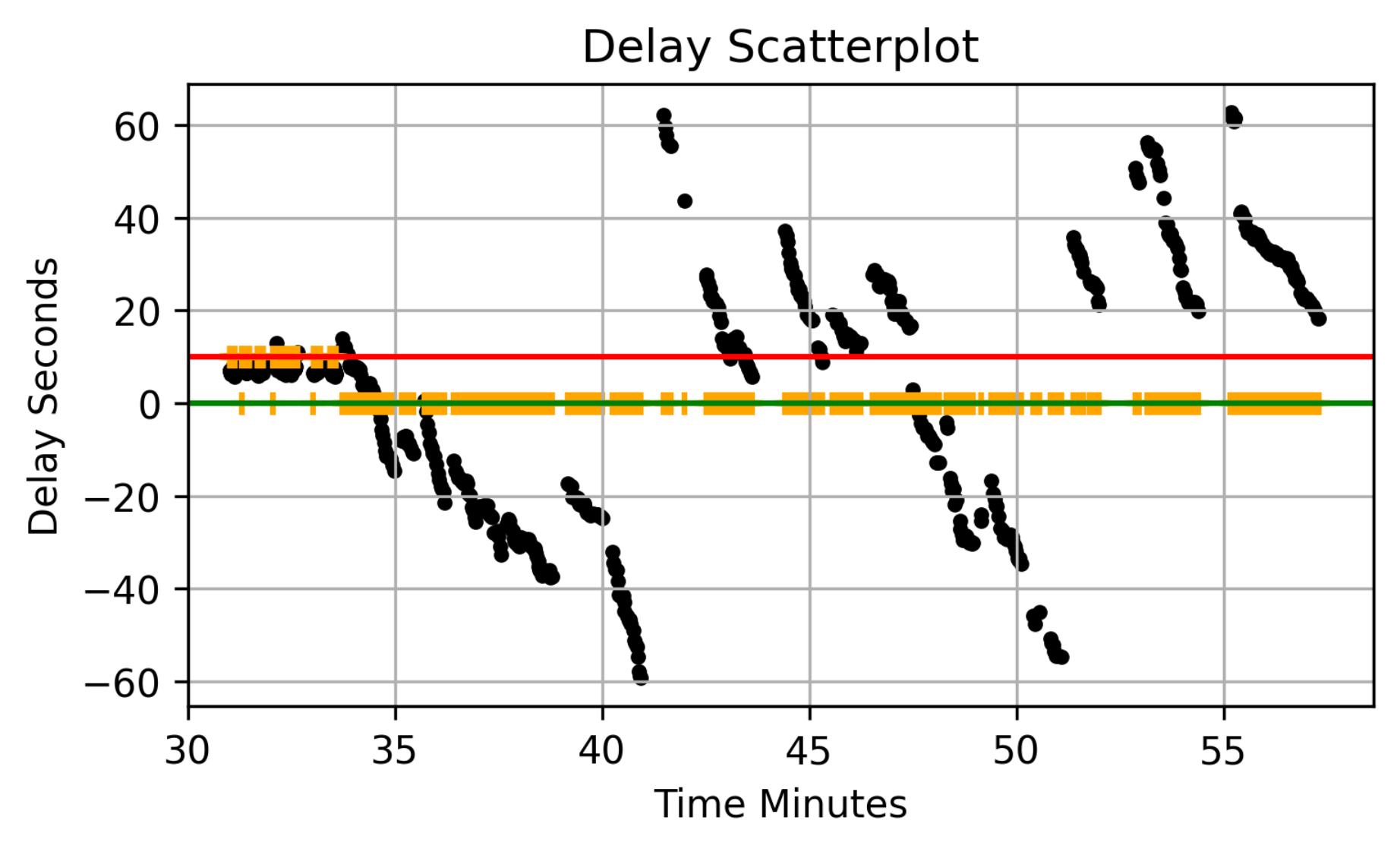
My 2013 user study of subtitle quality measured the impact of delay on the perceived quality of subtitles.

- Each 2 second increase in delay reduces the subtitle quality by approximately one ITU grade.
- Delays over 10 seconds are totally unacceptable.
- My system has shown that some channels regularly broadcast subtitles over 10 seconds late, sometimes many times that.
- It has also found examples of subtitles being broadcast early.
- And subtitles that omit anything up to half the spoken words.

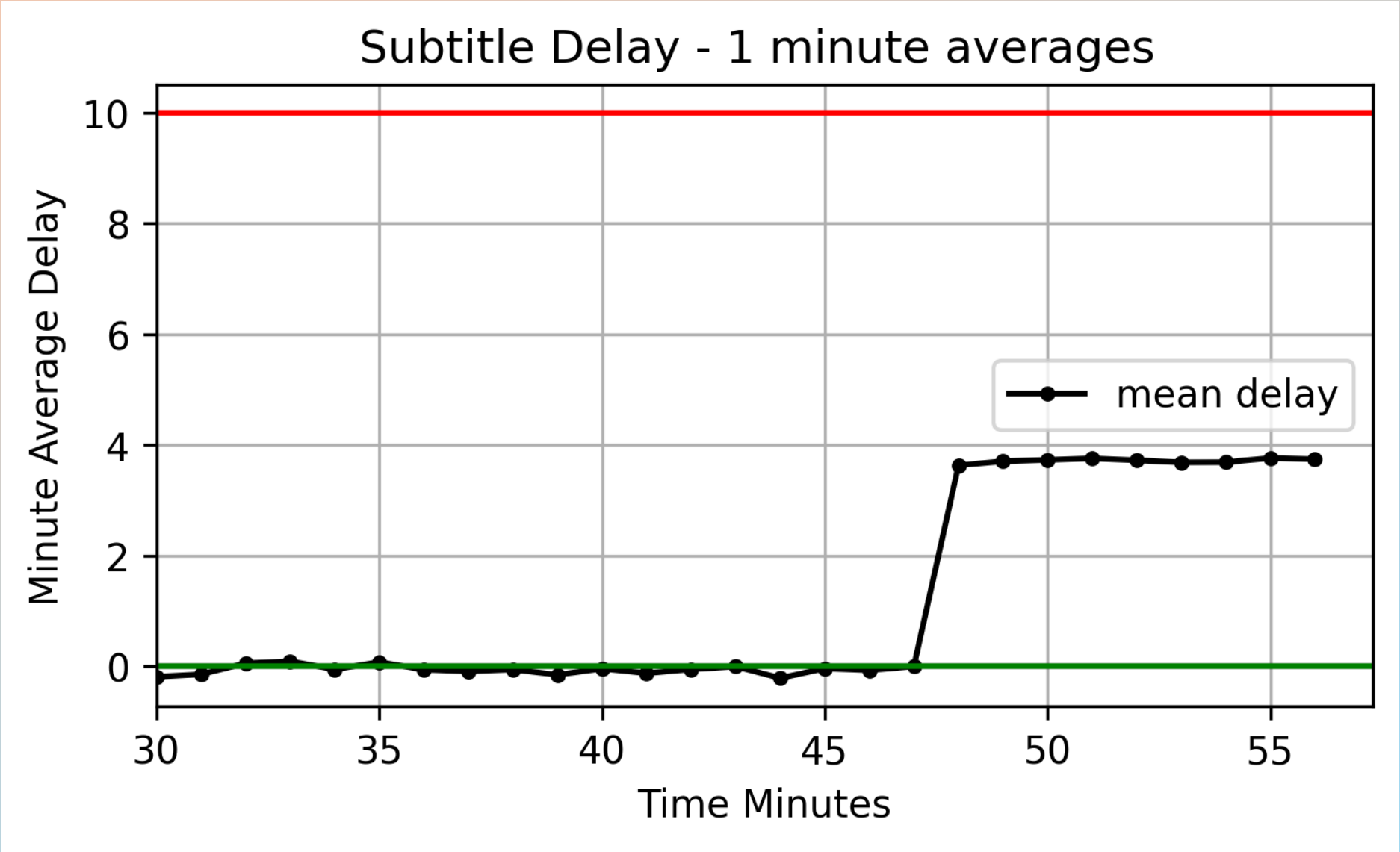
Problems with timing – live (snake)



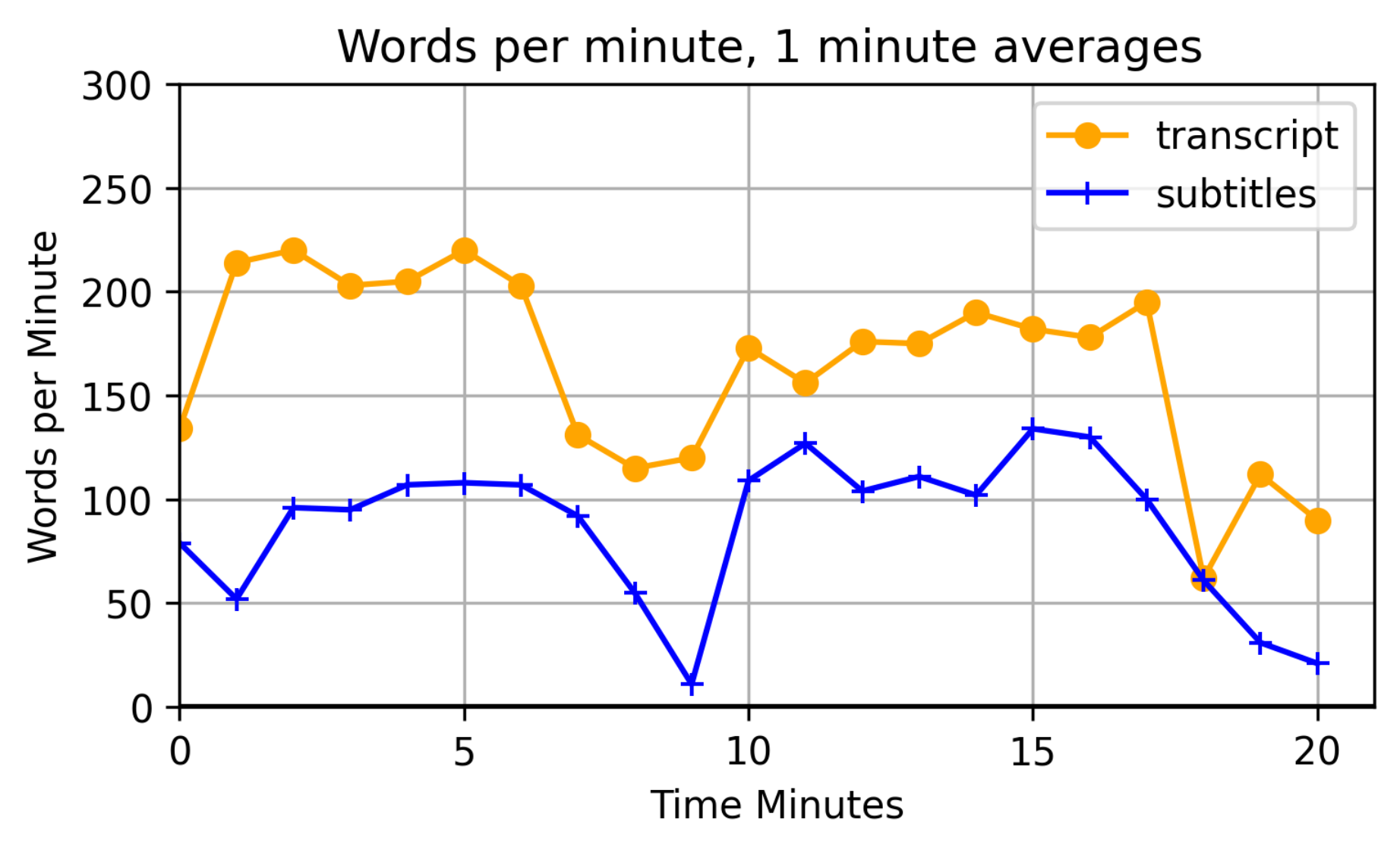
Problems with timing – live (block)



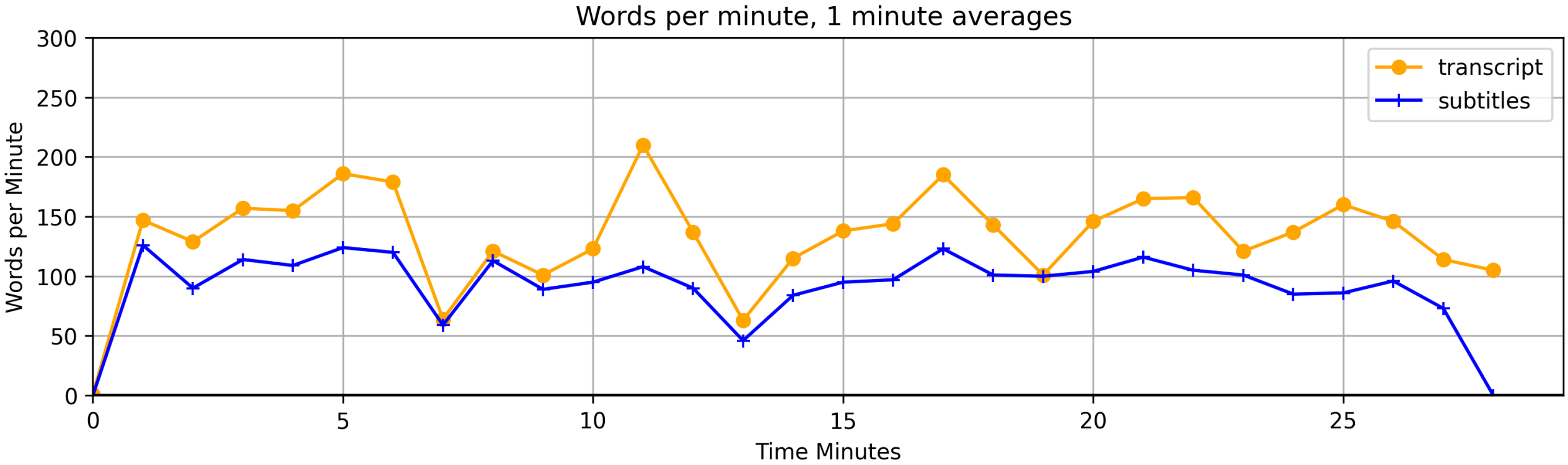
Problems with timing – pre-recorded



Problems with word loss - live



Problems with word loss – archive (1988)



Conclusions

The quality of live television subtitles remains a significant problem, and some pre-prepared programmes are being repeated with out-of-date subtitles.

This work demonstrates the viability of automated subtitle monitoring for delay and word loss. It cannot *currently* estimate the proportion of word errors, but this is work in progress...

A production version of this system could be used for quality control of programmes before broadcast and off-air monitoring to detect technical failures which could lead to improvements to subtitle quality.

Acknowledgements

I would like to thank **Dr Michael Crabb**, Head of Computing at the University of Dundee and **Dr Carol O'Sullivan**, Associate Professor in Translation Studies at the University of Bristol and co-founder of the UK Subtitling Audiences Network for their support and encouragement for this work.

Next steps...

Updates will be posted at www.subtitles.org.uk

Email: info@subtitles.org.uk

We are currently looking for funding and partners as further progress will depend on the level of financial support.