Research & Development
White Paper
WHP 259

September 2013

The Development of a Methodology to Evaluate the Perceived Quality of Live TV Subtitles

Mike Armstrong

BRITISH BROADCASTING CORPORATION
The Development of a Methodology to Evaluate the Perceived Quality of Live TV Subtitles

Mike Armstrong

Abstract

This paper describes the development of a methodology to measure the perceived quality of television subtitles based on well established guidelines for measuring perceived audio quality. The aim was to estimate the relative impact of reducing the delay in the appearance of live subtitles as against an increase in the accuracy of the subtitles. The results show a clear difference in the impact of timing and accuracy for our audience depending on whether they are watching television with the sound turned on or off. For those watching with sound the timing is the most important whilst for those watching with the sound off the accuracy is the key factor.

This paper was originally published at IBC2013

Additional key words: access services, captions
The Development of a Methodology to Evaluate the Perceived Quality of Live TV Subtitles

Mike Armstrong

1 Introduction

The BBC attained its own target of subtitling 100% of its television programming in 2008 and, with the exception of the occasional technical problem has continued to meet this target in subsequent years, BBC (1) & (2). However, where subtitles are created live, there are still big challenges in making subtitles which are accurate, in that the words in the subtitles match the words spoken, and timely, arriving as soon as possible after the words are spoken. Whilst the accuracy of live subtitling has improved somewhat the impact on our audience of the subtitles arriving late, relative to the sound, was less well understood.

A number of potential mechanisms for reducing the subtitles delay had been identified and this work was carried out to investigate whether reducing the time between a word being spoken and the subtitle arriving (improved timing) could have a beneficial effect for the audience, and if so, how large that benefit would be when compared to a further improvement in subtitle accuracy. The aim was to test the impact of delay on the perceived quality of the subtitles across a range of programme genres whilst making the experimental conditions as close as possible to the normal TV viewing experience, thus maximising the ecological validity of the work. It was also felt important to use subtitles with the rhythm and errors of the kind that occur naturally in the subtitling, and the word-at-a-time presentation style. Furthermore the programme material needed to be items that the test participants had either not seen before or would not have remembered.

Further issues were clear from the outset. The delay of live subtitles can vary quite widely depending on whether the subtitler is working from a programme script, or re-speaking a live interview, and both often occur within a single programme item. This work would need to arrive at a suitable measure for subtitle delay and cover the possibility of subtitles arriving early. It was also decided that the participants should reflect the broad range of people who use subtitles to give as much insight into the issues as possible.
2 Background Research

This work set out to provide a quantitative measure of the impact of delay and errors on the perceived quality of the subtitles by means of a controlled experiment. Previous work on subtitle quality in the UK mostly focused on issues such as reading rate and presentation, or has been in the form of general satisfaction surveys.

The most relevant piece of research was published by Ofcom, the UK communications regulator, in their report, “Subtitling – An Issue of Speed”, Ofcom (3). In this work there was an element of user research where participants were asked to rate the quality of a series of programme clips subtitled at different reading rates. However, this question was simply added to disguise the interest in subtitling speed. The two main questions were about how much the participants had been able to read and how comfortable they were with the speed of the subtitles.

Issues of reading rates and subtitle presentation were also explored by Pablo Romero-Fresco of the University of Roehampton as part of the DTV4All project, DTV4All (4 & 5). The work on reading rates used comprehension tests to find out how much each person understood from watching the subtitles at different rates and the work on presentation used eye-tracking equipment to measure the time spent looking at the subtitles compared to the rest of the image. The DTV4All work also included an opinion survey about live subtitling conducted by the RNID (UK Royal National Institute for the Deaf) in 2009.

In the USA, where verbatim (rather than edited) subtitling is the norm, work has gone into studying the impact of different types of errors. The National Centre for Accessible Media has looked at the severity of different types of subtitling errors by means of a web survey and ranked the different types of error by the severity of their impact on the viewer, in order to try and improve the metrics for subtitling accuracy, Apone et al (6).

Having found no equivalent prior work on the perceived quality of subtitling the chosen approach was to borrow from other areas where quality metrics have been used reliably for many years. The methodologies selected were from the long-established guidelines for assessing audio quality, Rec. ITU-R BS.1116-1 (7) and Rec. ITU-R BS.1284-1 (8). These were adapted to meet the specific challenges of assessing the quality of subtitles.
3 Initial Prototyping

To provide a measurement of the relative impact of accuracy and timing on the quality of subtitles it is first necessary to show that there is a relationship between accuracy and quality and a relationship between timing and quality, preferably all in one experiment.

The problem with trying to measure subtitle quality is that any previous viewing of the item could influence the quality score because it would improve comprehension. This rules out showing the same item in two different conditions. Also, because subtitling is about conveying meaning, and the need for ecological validity, it was felt wise to use clips which were complete narrative elements, such as a single news story. For this reason the clips are longer than is normal for an audio test. Finally, because we were using the subtitles as-broadcast any variation in subtitle accuracy could only come from the variation in the measured accuracy of the clips shown.

In order to be able to calculate mean opinion scores a continuous rating scale was used ranging from 0 to 100. This was annotated with the 5 ITU quality scale, ‘bad’, ‘poor’, ‘fair’, ‘good’ and ‘excellent’ to give participants guidance.

The prototype tests were built using a Python script to control VLC video player (9) and provide a method of entering the quality scores. VLC has the advantages of being able to display DVB subtitles and vary the synchronisation between the audio & video and the subtitles. The clips were in the form of edited DVB transport stream recordings, exactly as broadcast.

Initial prototyping of the tests was carried out with colleagues from BBC R&D as test participants - none of whom were regular users of subtitles. They viewed a number of opening stories from BBC One news bulletins with the subtitle delay randomly reduced in steps of 0, 2, 4 or 6 seconds. The first item in each bulletin was used because it has a clean start with no subtitles from a previous story. The item included the newsreader’s introduction and the main item and ended with the last subtitle of that story. Some participants were shown clips with the sound off whilst for others the sound was set at a low level such that they relied on the subtitles for some words.

The initial results showed a noticeable trend for increased quality with reduced delay with the sound on, but not with the sound turned off, and the scores for the different clips seemed to relate more to the type of content in the story rather than the subtitle accuracy. There were also adverse comments about the length of some of the clips, and items where there were large, or disturbing errors in the subtitles.

The main insight from this initial prototype was that careful clip selection was going to be an important factor in creating a successful experiment. The clips used would all have to be similar items, less than 3 minutes in length, with no obvious flaws in the subtitles, but with a variation in the subtitle accuracy between items. It was clear that this similarity could only be achieved by sticking to a single programme genre and news is the main genre that is subtitled live on the BBC.
4 Test Material Selection
The BBC Six O’Clock News bulletin has a set of midway headlines followed by a music jingle. The item that then follows has a clean start to the subtitles and is usually a short feature item rather than a long story about politics or military conflict. Also being an early evening programme these tend to contain fewer upsetting images. The bulletins were examined over the period of April to November 2011 and 32 news stories were selected as containing suitable content. From these, 21 were selected as having reasonably consistent subtitles with a clean start and end, and no obvious flaws.

These 21 clips were then shown, in 3 batches, to 12 members of BBC R&D with the sound muted. There were no changes made to the timing, the aim was to check for any further problems with the clips and arrive at a quality score for each item. Participants also made notes of any issues they had with items or the procedure.

Each item was separately scored for accuracy using a modified word error rate count. Because live subtitling in the UK often omits whole sentences or phrases where they add little to the narrative, omissions were not counted as errors. Whilst other approaches are also used to measure errors and have a different level of sensitivity, there is reasonable correlation between most approaches and currently there is no single recognised standard, Apone et al (10). Plotting the mean quality score against the accuracy score showed a clear (and statistically significant for p<0.05) trend of increasing quality with improved accuracy across the clips, despite the differences in timing.

The median delay for the subtitles in each clip was estimated by playing the clips using VLC and adjusting the delay until around half the subtitles were early and half late. This median delay was estimated separately for the newsreader’s introduction and the news report to check that there was no major difference. Attempts at measuring the delay of each word within an item proved impracticable.

Of these 21 items 9 were eliminated from these results. The problems were mostly with large omissions or long pauses as these caused large fluctuations in the subtitle delay. Two were eliminated because the subtitles had an unusually large number of errors and two others because the subtitler had extensively re-worded the subtitles. This left 12 usable test items with accuracy levels covering the range 94.6 to 99.6%. One further item was used as a training/familiarisation clip.
5 Test procedure

A new set of tests was devised to increase the sensitivity of the procedure. The clips were divided into two sets of 6 pieces and a set of 6 timings was chosen to cover the range of interest and give additional information around synchronisation. These were -6, -4, -2, -1, 0 & 1 seconds with negative numbers indicating delay, so -6 seconds means the subtitles were 6 seconds behind the audio. The measured delay of each clip was used to calculate the timing offset required to achieve the required timings for each showing.

Each set of 6 clips were then shown 6 times to 6 different participants so that each item had been shown at each delay setting and each delay setting had been used for every item. The order in which the clips were shown to each participant was randomised.

This time, prototyping the test with R&D colleagues gave a clear trend for increased quality with reduced delay when the sound was present, but no obvious trend without sound. The trend for accuracy, by contrast, was clear without sound, but not with sound.

A full controlled experiment was then run with participants who were recruited through an external market research recruitment agency. All participants had to be regular users of subtitles, regular viewers of news programmes and able to read subtitles at a normal TV viewing distance. The experiment was carried out in the BBC R&D North Usability Lab. This was set up like a living room to try and reproduce a normal TV viewing experience. A total of 28 participants were shown a training item followed by two sets of six clips with a short break in between for questions. They were asked to set the TV to their normal sound level whilst viewing with subtitles during the training clip. Participants were asked about their reasons for using subtitles but not directly questioned about their hearing ability.

Around half normally used subtitles with the sound off, the rest viewed subtitles with the sound turned on. Those watching subtitles without sound tended to be people viewing TV in a gym, at work or at home with a sleeping child. Those watching with sound were mostly older people who needed help because of their hearing and others who used subtitles because of noisy children and the like.

Scores from 4 participants had to be rejected because the participants did not engage fully with the exercise, giving maximum scores to several items, usually the first 3 or 4 items. This left 24 valid sets of data, of which 12 viewed with sound and 12 without and fortuitously there were 2 complete sets of 6 items in each half giving a fully balanced test.
6 Results

Analysis of the data using linear regression using SPSS (11) showed a strong and statistically significant (p<0.05) relationship between improvement in perceived quality and increased subtitle accuracy for people watching subtitles without sound. For the range of accuracy tested a 1% improvement in accuracy (reduction in word error rate) gave an improvement of 6 points in the quality score, but with an uncertainty of nearly 3 points in both directions.

For people watching with sound on, the improvement in the quality score was not statistically significant – see Figure 1. There is overlap with the mute case so the difference is not statistically significant.

Figure 2 shows the shape of the plot of the mean quality scores for each clip against the measured subtitle accuracy.
By contrast the clearest trend for timing was for people watching with sound where there was a strong and statistically significant increase in the quality score with improved timing (reduced delay). For the range of timings tested, each 1 second reduction in the subtitle delay gave just over 5 points improvement in the quality score.

For those watching without sound the trend was much smaller but just reached statistical significance - see Figure 3 – and the difference between the two trends is also statistically significant.

Figure 3 – Mean Quality Score vs Timing

Figure 4 shows the shape of the plots of the mean quality scores for each of the subtitle timings tested.

Figure 4 – Plot of Mean Quality against timing for each timing tested
7 Discussion
The results from this experiment indicate that this methodology has been successful in measuring perceived subtitle quality. The striking degree of linearity in the trends for quality against timing – Figure 4 – shows good correlation between the perception of delay and the rating of quality over the range tested. The departure from the linear trend around the zero delay point is possibly due to the variation in timing within each clip which meant that some subtitles would be early and some late in most cases for target timings of -1, 0 and +1 seconds. The fact that the trend was so close to linear for the mean scores ($R^2 > 0.9$) was a cause for concern at first. However, after some further background reading it was realised that, for the time intervals tested here, the human sense of timing is very accurate, so the test is showing that the quality scale is remarkably linear, Sacks (12).

The trends for accuracy are less clear because each clip is different in content as well as accuracy. Further tests artificially manipulating the accuracy of the subtitles could be run to more closely estimate the impact of accuracy whilst holding the timing the same. This would require the re-authoring of the subtitles which was beyond the scope of this work.

Similarly, the presentation style, word-at-a-time rather than block, was not varied in this work. It would be instructive to repeat this work with the inclusion of block presentation of the re-timed subtitles to see what the impact of presentation style would be on the perceived subtitle quality. It would also be interesting to see how perceived quality varied with altered reading rates and how the results on a quality scale mapped back onto the eye-tracking work of the DTV4All project.

8 Conclusion
This work has developed a methodology which applies the quality metric, familiar from audio and video quality work, to the measurement of the perceived quality of subtitles. The work has shown the level of care that needs to be taken in selecting suitable test items in order to control for issues which would otherwise impact on the experiment and decrease its sensitivity to the variables under test. The results have shown a strong linear trend for the relationship between the timing of subtitles and their perceived quality for people viewing with the sound turned on and a smaller trend for people watching without sound. It has also shown a clear trend for the relationship between the accuracy of the subtitles and the perceived quality for people watching without sound. However, the trend for improved quality with improved accuracy is much less clear for people watching with sound.

The main conclusion from this work is that people watching subtitles with sound have a far greater need for synchronisation of the subtitles and are less worried by inaccuracies, whereas for people watching without sound the problems of inaccuracy dominate. Thus your perception of the importance of problems with subtitles will depend on whether you have access to the sound whilst watching television with subtitles. If you are using the subtitles to assist the soundtrack then you can use your comprehension of the soundtrack to correct the subtitle errors, but the delay in the arrival of the subtitles makes this difficult. However, if you cannot hear the soundtrack then you do not have the additional information provided by the soundtrack to overcome the errors in the subtitles, and you are less bothered by the lack of synchronisation with the pictures.

It is clear from these results that for the section of the population with hearing difficulties who use subtitles alongside sound to watch television, any reduction in the delay of subtitles can provide a useful improvement to the quality of their viewing experience.
REFERENCES


ACKNOWLEDGEMENTS

This work was commissioned by Nigel Megitt of BBC Technology, Distribution and Archives, Solution Design team who also supplied a great deal of support and enthusiasm along the way. Further valuable assistance and specialist advice was provided by Trevor Ware, Penelope Allen, Elizabeth Valentine-House, Maxine Glancy all from BBC R&D and many more colleagues working at BBC R&D’s North Lab helped by acting as participants in the prototyping stage of this work.