Ameliorating the quality issues in live subtitling

ANDREW LAMBOURNE[1]

[1] Leeds Beckett University, Leeds, UK
e-mail: a.d.lambourne@leedsbeckett.ac.uk

Abstract

Access services have evolved significantly over the past 30 years, and optional subtitles are widely available on mainstream channels in the UK and Europe. Live subtitles now routinely accompany news, sports and chat-shows. The production and consumption of live subtitles both impose heavy cognitive loads, not helped by the constraints of time, practical limitations and the inevitability of errors. Live subtitles of broadcast quality are normally created by real-time transcription of phonetic key strokes or re-spoken text produced by a human intermediary. If the subject matter and vocabulary of the content is not known in advance, transcription errors are very likely. Such errors are distracting or confusing – regulators deprecate them, the press mocks them, but producers have to risk them to deliver the service. Less obvious quality issues also arise to do with the timing lag and the style and position of subtitle text. Recent studies into audience perceptions of live subtitle quality are reviewed, and the results of a pilot study in classifying apparent errors according to likely cause are used to illustrate possible opportunities for mitigation. This suggests that aspects of re-speaking style may be adjusted to enhance accuracy, and that there may be opportunities for new approaches to underpin further quality improvements in future.

1 Introduction

Access to broadcast television by means of optional subtitle services for people with hearing impairment started in the UK in 1979. Initially carried by teletext, subtitle data is now delivered in the digital television multiplex. Service levels have grown considerably, with 100% subtitling on six BBC channels. Ofcom, the UK regulator, makes subtitling a licensing requirement.

Television subtitling is a well-understood process with mature guidelines, technology, commercial models and editorial skills. Digital broadcasts reliably carry subtitle data in one or more languages. Perhaps the only area debated in initial research which has not been explored yet in broadcast subtitles is the provision of dual-level services to cater explicitly for deaf children and less able readers. While technically feasible, cost constraints on service production have discouraged provision. The competitive nature of service contracts motivates providers to leverage automation opportunities arising from advances in speech and language tools, Artificial Intelligence, image analysis, and production workflows.

Quality and reliability have been closely monitored by lobbies representing people who depend on subtitles for their access to and enjoyment of broadcast television. Research by Romero Fresco and others at the University of Roehampton and elsewhere has developed ways to describe subtitle service quality and accuracy. Reports by the charity Action On Hearing Loss represent viewer attitudes to quality issues in the broader context, including most recently highlighting the lack of subtitle availability on widely used catch-up channels. Ofcom has consulted with service providers in an effort to improve subtitle quality and availability. Even after nearly 40 years, access services are still under scrutiny.

For real-time subtitling, the challenges are at their most acute. The subtitler acts as an intermediary between the live presenter and a computer transcription system – be it Stenographic (using a special phonetic keyboard and transcription software), or speech recognition (by re-speaking text and punctuation to a speaker-dependent speech-to-text system). Such intermediation imposes the simultaneous cognitive tasks of listening to a presenter, formulating a target representation, keying or speaking the target text rapidly and precisely, and reviewing the results in case a serious error demanding correction has occurred. The mental workload is high, reactive and unpredictable. Inevitably there will be transcription errors, and deciding whether an error is serious enough to warrant a correction demands a shift of concentration. This itself may cause another error, or the omission of succeeding text due to having fallen behind.

For all these reasons, if technology can be used to assist in the process of avoiding, or in detecting and correcting errors, then until such time as speaker-independent tools can deliver near-100% accuracy in text and punctuation, the quality of live subtitles will be further enhanced.

2 Related Work

Viewers of live closed captions (access subtitles) in the US have since the 1980s been exposed to near-verbatim Stenographic transcripts which scroll smoothly onto the screen. UK live subtitling users have seen a variety of speeds and styles: QWERTY summary subtitles at 60-80wpm (words per minute), dual QWERTY or Velotype production at 90-100wpm, near-verbatim Stenography at up to 180-220wpm, and respoken live subtitles at around 140-160wpm. Different editorial philosophies between ITV and BBC coupled with different technical approaches have exposed viewers to more or less heavily edited text, and row-scrolled or blocked text. Each approach attracted supporters and critics in its day; none met the same quality standards as subtitles prepared in advance for recorded material.

Live subtitles suffer from text errors and from a lag between spoken text and corresponding subtitle content. Work over many years between Romero Fresco at the University of Roehampton and Martínez led to the NER model - a standard measure of subtitle accuracy [1]. NER penalises omission of key facts, thus encouraging a fuller rather than a more edited approach for dense factual content. Though the model does not specifically take account of delay, nor of presentation style, an assessor would normally include a subjective comment on these aspects. Romero Fresco has also highlighted the readability challenges of scrolling subtitle text [2].

Armstrong at BBC R&D developed a scoring system to enable subjects to rate live subtitle quality which takes account of accuracy and delay [3]. His findings indicate that those who rely on the text because they are not using sound are less sensitive to lag but more sensitive to text errors. Those who use both sound and subtitles are more sensitive to lag but less to textual accuracy.

Sandford, also at BBC R&D, investigated tolerance of the speed of test-subtitled clips among regular subtitle users [4], but acknowledged that first-language signers were not included and deserve more specific research. The findings indicated that a subtitle rate matching the natural rate of the programme was tolerated even if it was higher than recommended. The clips of scrolling text were not actual live subtitles but synchronised simulations which avoided the cognitive load imposed by dealing with text which lags the sound being heard. It would be instructive to repeat this study using clips with typical subtitle lag in order to explore the impact of asynchrony on use of lip-reading or residual hearing to assist in reading. Eye-movement tracking could also be used to assess whether the viewer has time to look at the rest of the picture while reading fast verbatim subtitles.

A recent online survey of subtitle users was carried out by an EPQ student mentored by the author [5]. Participants included a mix of students at a school for profoundly deaf children, members of a deaf club for young people, and retired hard-of-hearing individuals. The informal exploration of feedback about subtitling offered a mix of multiple choice and free-text responses.

The cognitive confusion caused by out-of-synchrony text was highlighted as the most annoying problem, and led some hard-of-hearing viewers to mute the sound in order to cope better with delayed subtitles. Some viewers preferred minimal editing, others requested an option to select a simplified and lower-speed version. Textual accuracy was mentioned as a problem, as was the annoyance caused by subtitle text obscuring key visual events, and (conversely) subtitles being moved around the screen too often.

The issues which annoy subtitle users appear to be the same today as they were when the first detailed UK research into access subtitling was conducted for ITV by Baker at Southampton University [6]:

- The audiovisual translation process to augment or replace a TV soundtrack with subtitles should take account of readability, accuracy of text, accuracy of content, level of editing, text presentation style, position, timing, speed and duration: some of these factors are interrelated.
- The users of subtitles may have a range of hearing loss, may be in noisy environments, may be language learners or first-language signers, and will span the same age groups and tastes in programme content as non-users of subtitles.

The quality studies and applicable regulatory guidelines (for example those of Ofcom in the UK [7]), motivate subtitle producers to strive for as much quality as budget, time and practical constraint will permit, recognising that there is a need to serve the majority user profile rather than any individual sub-group.

As the power of receiving devices improves, however, the possibility exists to provide a degree of computer-assisted personalisation of the viewing experience, and this potential is explored in terms of subtitle position in the work by Brown et al at BBC R&D [8]. This theme also featured in the scope of the Hbb4All EU project as outlined in the summary presentation by Menendez [9].

In the specific case of real-time subtitles, the foregoing studies confirm audience sensitivity to the fundamentals already enumerated. We now briefly consider which of these could feasibly be controlled and improved.

Readability of the text presentation style is a delivery rather than a production issue, and could even be addressed by receivers offering an option of a scroll or block style. Position can be controlled during production but, with picture analysis software, may be able to be automated as described by Hu et al [10]. The accuracy of text and content, the level of editing (hence text speed and duration), and the timing lag are clearly production matters. Can these interrelated factors be controlled and improved to maximise the quality of live subtitles?

The level of editing may be defined by editorial policy or dictated by technical limitations constraining speed. To edit rapid, factually dense speech in real time and in a balanced way adds to the cognitive load of the re-speaker, though in practice a "rounding off" technique may have to be adopted to reduce the re-speaking rate to one where precise enunciation can be maintained and the transcription engine can keep up. This relationship between speed and accuracy is highlighted in studies at the University of Antwerp and Artesis University College supervised by Leijten, Remael & van Waes[11].

Timing lag is dictated by methodology and presentation method. The real-time task sequence includes listening, formulating, delivering re-spoken text or key strokes, transcription, and (in block mode) filling to the end of a block subtitle. The resultant lag can be reduced by initiating the above sequence without waiting for the broadcast signal to be encoded, thus giving the subtitler "advance audio". This method was pilot tested in 2015 as described by Ware and Simpson [12].

Accuracy of text and content remain as two areas where improvement may be possible without a step-change in methodology and/or technology. Accuracy studies have been conducted by Moores and Romero Fresco [13] on text errors in subtitled weather reports, and at Artesis (ibid) in Belgium using VRT subtitles. Errors were respectively classified according to parts of speech or Technical vs Human categories. The study described below explores classification based on the likely causes of error, derived using first-hand experience as a re-speaker trainer. The objective was to assess whether this classification could usefully guide amelioration priority.

3 Methodology

Professionally re-spoken subtitle data was logged during transmission for a range of genres including chat shows, political discussion and live sport. The study focused on classifying the errors which were apparent by reading subtitles without access to the soundtrack – as a viewer. These "apparent errors" were then grouped according to a judgement of "likely cause" drawing on experience of re-speaking and re-speaker training to relate symptom to proposed cause. Data of this kind is clearly not objective and depends on expert judgement, which may differ from one individual to another. Nevertheless the method and the resulting categorisations are offered as a useful starting point. The intention was to identify the most frequent apparent causes, and to use this to prioritise subsequent work into possible ameliorations.

The sample set comprised the subtitles for 23 broadcasts of duration between 30 minutes and 4 hours:

- 19 half-hour chat shows
- 3 hour-long shows: chat show, talent show, debate
- 1 football commentary spread over 4 hours

The subtitle texts comprised some 4-5,000 words each for the half hour chat shows and some 18,000 words for the football commentary. All the texts were re-spoken.

4 Experimental Results

The first notable observation was that "apparent errors" were few in number: a half-hour chat show contained between 1 and 40 such errors; the football commentary contained 127. It should be noted that an error such as a named entity error, where a proper name was rendered as an erroneous word group, was for this experiment counted as a single error, since causes not outcomes are of interest. As also noted, these are text errors apparent to a reader, ignoring potentially less obvious errors of edition, omission or fact. In many cases the error was obvious because a corrected form was sent by the subtitler, preceded by a marker. Although infrequent (affecting less than 1% of the content), text errors can mislead or confuse the viewer, and if corrected by the subtitler, have an associated cost through loss of concentration and loss of time.

The error categories which were developed during the experiment, along with a typical example of each, were:

Single-word phonetic blurring
(it is not the country *will* want for our children)
Single-word homophone
(he *through* the microphone into a lake)
Missing single word
(I am not happy with <*how*> much was lost)
Inserted single word
(it changes everything *to* for the better)
Multi-word phonetic blurring
(standing *over nation* [ovation])
Multi word homophone
(*into* or three years time)
Capitalisation error
(the funny thing is, I am *A* prude)

Pluralisation error
(the European *Championship's)*
Number-grammar error
(they changed to *4-14-1* [4-1-4-1])
Named entity error
(*Andrey and silver* [Adrien Silver])
Punctuation misinterpretation
(I don't like pressure *for stop* [.])

The categories labelled "phonetic blurring" were ones in which the error appeared most likely to have been caused by imprecise enunciation, as opposed to direct homophones where phonetic ambiguity is definite.

Results across all the sample texts showed the following distribution of error tallies for a total of 384 error cases:

Single-word phonetic blurring	53.40%
Multi-word phonetic blurring	18.80%
Single-word homophone	9.10%
Named entity error	4.40%
Multi word homophone	3.90%
Punctuation misinterpretation	2.90%
Missing single word	2.60%
Capitalisation error	1.80%
Number-grammar error	1.30%
Pluralisation error	1.00%
Inserted single word	0.80%

The majority (62%) of these errors affected only single words and appeared to be due to imprecise enunciation or direct homophones. In a further 23% of cases the phonetic imprecision had an effect beyond a single word.

For the set of programmes analysed in this sample, textual errors were in the great majority (85%) of cases judged to be associated with phonetic imprecision rather than, for example, missing vocabulary. Furthermore, the cases judged to be due to imprecision were more often due to the confusion of a word with one sounding similar, rather than a word and one sounding the same – ie a homophone.

These observations suggest that text error rates could be reduced by more precisely enunciated re-speaking. To understand whether the apparent imprecision was a characteristic of particular re-speakers, or depended on the instantaneous speed of text delivery at the moments when such errors occurred, would require more detailed study. But for now, the observation is that a focus on clearer enunciation could deliver benefit.

Consideration was given to whether the apparent errors could easily be detected automatically using available tools. Since speech recognition systems select words from a dictionary, misrecognition will produce correctly spelled results, even if the words are wrong. A spelling checker is therefore unlikely to add value in detecting such errors. A basic grammar checker such as that found in a standard word processor did not flag a sufficient number of the errors in the context of surrounding text to be useful.

Since the errors were identified by reading through the subtitle texts and flagging "nonsensical items", the next step in the experiment will be to further investigate how the human reader identifies the apparent mistakes, and then examine the consistency between different readers in identifying such errors. If consistency exists then it may be possible to identify features which can be detected and recognised by an automated tool, though the risk of false positives may make this difficult to achieve.

In any case, corrected forms for such errors, with audio context, could be used to retrain the speech recognition system and bias it towards the correct rather than the erroneous forms when speech is less distinct than ideal.

5 Conclusion and Proposals

While subtitle error rates can be assessed and discussed, categorising the apparent errors from the perspective of likely cause may, based on the results from this sample, indicate remedies which are reasonably achievable.

Identifying and classifying sufficient errors to produce a reliable pattern of their relative frequency involves a significant investment of time, but once the categories are clear this could be approached as a crowd-sourced activity.

Investigating in more detail the way in which humans identify the kinds of transcription errors which occur in live subtitles may inform the next steps in improving the capability of speech transcription systems. If obviously faulty text can still emerge from state-of-the-art speech recognition systems in the hands of trained professional users, it is worth asking why, how, and what could be done to avoid it.

References

[1] P. Romero-Fresco & J. Martínez, "Accuracy Rate in Live Subtitling: The NER Model" in *Audiovisual Translation in a Global Context. Mapping an Ever-changing Landscape* ed J. Díaz-Cintas & R. B. Piñero, pub Palgrave, London, 2015

[2] P. Romero-Fresco, *Subtitling through Speech Recognition: Respeaking*, pub Routledge, Manchester, 2011

[3] M. Armstrong, "The Development of a Methodology to Evaluate the Perceived Quality of Live TV Subtitles" *BBC R&D White Paper WHP 259*, Sep 2013

[4] J. Sandford, "The Impact of Subtitle Display Rate on Enjoyment Under Normal Television Viewing Conditions" *BBC R&D White Paper WHP 306*, Sep 2016 and associated blog http://www.bbc.co.uk/rd/blog/2015-09-how-fast-should-subtitles-be (accessible Sep 2017)

[5] I. Hawkins, "The cognitive challenges of live subtitling" *unpublished EPQ project*, Sep 2017

[6] R. Baker, "ORACLE subtitling for the deaf and hard of hearing" *Department of Electronics, University of Southampton*, Jan 1982

[7] "Ofcom's Code on Television Access Services" pub Ofcom, UK, A4.11 – A4.21, May 2013

[8] A. Brown, R. Jones, M. Crabb, J. Sandford, M. Brooks, M. Armstrong & C. Jay "Dynamic Subtitles: the User Experience" *BBC R&D White Paper WHP 305*, Aug 2015

[9] J. Menéndez & C. Martín, "The HBB4allProject: From the Accessibility Vision into Market Reality", presentation at *media4D, Saint-Denis* Jul 2014

[10] Y. Hu, J. Kautz, Y. Yu & W. Wang, "Speaker-following video subtitles" *ACM Transactions on Multimedia Computing, Communications and Applications*, vol 11, no 2, 32:1–32:17, Jan 2015

[11] M. Leijten, A. Remael & L. Van Waes (sup) "Live subtitling with speech recognition" *Pilot research project and training at the University of Antwerp and Artesis University College* 2008? from www.respeaking.net/programme/remael.ppt (accessible Sep 2017)

[12] T. Ware & M. Simpson, "Live subtitles re-timing proof of concept" *BBC R&D White Paper WHP 318*, Apr 2016

[13] Z. Moores & P. Romero-Fresco "The Language Of Respeaking – A Classification Of Errors" *Presentation given at the 5th International Symposium on Respeaking, Live Subtitling and Accessibility*, Rome, Jun 2015