JUST-IN-TIME PREPARED CAPTIONING FOR LIVE TRANSMISSIONS

M.N. Simpson[1], J. Barrett[1], P.J. Bell[2], S. Renals[2]

[1]Ericsson, UK and [2]University of Edinburgh, UK

ABSTRACT

Latency remains one of the most significant factors (1) in the audience's perception of quality in live-originated TV captions for the Deaf and Hard of Hearing.

Once all prepared script material has been shared between the programme production team and the captioners, pre-recorded video content remains a significant challenge – particularly 'packages' for transmission as part of a news broadcast. These video clips are usually published just prior to or even during their intended programme – providing little opportunity for thorough preparation.

This paper presents an automated solution based on cutting-edge developments in Automatic Speech Recognition research, the benefits of context-tuned models, and the practical application of Machine Learning across large corpora of data – namely many hours of accurately captioned broadcast news programmes. The challenges in facilitating the collaboration between academic partners, broadcasters and technology suppliers are explored, as are the technical approaches used to create the recognition and punctuation models, the necessary testing and refinement required to transform raw automated transcription into broadcast captions and methodologies for introducing the technology into a live production environment.

INTRODUCTION

Over the last 30 years the volumes of 'SDH' captioning (Subtitles for the Deaf and Hard of Hearing), both live and pre-recorded, have increased considerably across the globe. As coverage approaches 100% in certain markets, the attention of the regulatory bodies has shifted from volume to quality (2). A closed caption is fundamentally a short section of timed text, and three simple measures of quality can be viewed as:-

- its fidelity to the original spoken word
- the textual accuracy of the transcript
- and the timeliness with which it's presented

For a pre-recorded programme it is possible to ensure that fully accurate, perfectly timed verbatim captions can be presented; with the majority of productions there is sufficient time between the completion of editing and publication for the captioner to create and review the caption file to the required standard.

Live programming presents a considerable challenge; the caption data will need to be streamed in real-time from the captioner to the point of insertion, and synchronisation methods such as timecode cannot be relied upon. Whilst a sizeable proportion of a live programme may be spontaneous or ad-libbed, possibly 100% for sporting events, much live programming relies on a running order, sections of pre-scripted material, and prepared and pre-edited video ‘packages’. Much effort has been made by live programme makers to ensure that teams producing captions have access to running orders and scripts in advance of broadcast – most of the major broadcasters in the UK, France and Australia provide access to such data for their captioners. This data can be used directly (with suitable correction to phonetic spelling and grammar) or to prepare the captioning platform for any names and places likely to be mentioned on air, but it is less common for access to be made available to video content pre-broadcast. Where this access is granted, the dynamic nature of live production, especially in news, can mean that there is little time between completion of the clip and its inclusion within the broadcast.

Therefore there is a considerable challenge to transcribe and prepare captions for this prepared content within the time available. Techniques based on live origination can be used to accelerate the production process (such as respeaking – a speaker-dependent ASR technique used to transcribe the spoken word into captions via dictation), but these take production effort (which may be focussed elsewhere) and time. Ideally it should be possible to utilise an Automatic Speech Recognition (ASR) engine to generate the transcripts, but these typically produce transcripts that require more editing time to reach broadcast quality than manual origination would have taken. Previous experimentation as part of the EU-Bridge Speech Technology consortium indicated that ASR engines trained on very specific domains, such as Broadcast Weather, could achieve very high levels of accuracy – typically over 90% (3). It was recognised that the broader News domain would impact the overall accuracy of such an engine, but it would prove usable if a target accuracy of at least 95% was achieved, or 90% accuracy with fewer than 10% of errors being omissions, for the clips processed, with system confidence scores being used to pre-filter the poor quality transcripts. Scores lower than these would require sufficient manual review and editing that it would render any time-saving by using ASR worthless; for this project scores were calculated using the approaches described in the section below.

This paper will illustrate how such an automated ASR engine was devised and created with research partners, showing the benefits of context-tuned models and the practical application of Machine Learning across large corpora of data – namely many hours of accurately captioned broadcast news programmes. The challenges of introducing such a system into the live production environment will be discussed, as well as the testing required to ensure that each iteration of the engine reaches the target accuracy levels. The paper concludes with a summarisation of the success so far and some considerations for future applications.

THE VALUE OF BROADCAST DATA CORPORA

ASR models are based on machine learning using transcribed corpora of spoken audio (4). Broadcasters providing SDH captions unwittingly create large corpora of data ideal for the training of such models, and various research projects have taken advantage of these data sources to improve speech recognition techniques, notably EU-Bridge (http://www.eu-bridge.eu/), SAVAS (http://www.fp7-savas.eu) and NST (http://www.natural-speech-technology.org).

Between 2014 and 2015 Ericsson and the University of Edinburgh (UEdin) ran an 'Engineering and Physical Sciences Research Council (EPSRC) Impact Acceleration' project to build upon broadcast speech transcription research carried out in EU-Bridge and NST, in order to produce a speech recognition engine that could be introduced into live production. The project used a large quantity of pre-recorded BBC News video packages and brief interviews, along with their matching captions as the primary corpus for training and tuning the proposed engine.

The Ericsson live captioning platform is used to deliver captions for BBC News output in the UK, and the platform's timed text database was used as the master source of text for the project; each news programme is stored separately within the system, and a copy of the BBC's ENPS newsroom data is also stored temporarily within the system to provide running orders and scripts. The web interface / API to the news production system was utilised to provide access to download an audio file for each clip.

A major challenge within any such project is to gain the approval of the content owners for the use of their content for R&D purposes. For this project Ericsson worked with the relevant stakeholders and content owners within the BBC to facilitate the sharing of data with development partners; this is best achieved via a legal document describing the terms of the release of the data, specifying restrictions on use, the duration of access and so forth. It is of particular benefit for the R&D body to be able to use a small number of clips for external presentation within academic presentations, but largely the processes needed to prepare the corpora for training will strip out any video, reduce the audio signal to a number of acoustic 'features' and effectively render it unusable for any purpose other speech research; this obfuscation of the data is extremely 'lossy' and reassures the content owner that their materials cannot be re-constructed and used for purposes beyond their control.

A couple of simple applications were created by the Ericsson team; the first harvested the clips from each story within a given programme running order (these are cross-referenced by unique ID for playout automation purposes) and transcoded them to mono 16KHz '.wav' files appropriate for model training. The second extracted the caption data from the captioning system for the matching programme, split the data story by story and exported the stories matching the clips as individual files in a format suitable for speech training.

The data collection processes were executed by the captioners themselves, immediately post broadcast, as it was discovered that the individual newsrooms housekept any transmitted clips fairly soon after they were shown within the programme. The collated data was forwarded to the research team at Edinburgh at periodic intervals; using these processes a total data corpus of 600 hours was created, in addition to the 3000 hours of Sky News data and 130 hours of BBC Weather data inherited from the EU-Bridge project, and 2000 hours of multi-genre BBC data used in the NST project (5).

ASR AND MACHINE LEARNING TECHNIQUES UTILISED

The process of building an ASR engine requires training two separate statistical models using machine learning techniques. The *acoustic model* (AM) is used to model *phonemes*, the distinct sounds of speech which combine to form words. English has around 40 of these sounds. Since the early 1980s, the dominant method of acoustic modelling has been the *hidden Markov model* (HMM). This allows models for individual phonemes to be efficiently learned from long audio recordings without the need for hand-labelling of the phonemes. Once the models have been trained, the HMM allows new recordings to be efficiently decoded into phonemes and thence words. This process requires a pronunciation dictionary mapping all possible words into their constituent phonemes. This can depend on the accent of the speaker. We used Combilex, a dictionary constructed at the University of Edinburgh which is specifically designed to handle UK English accents (6).

The second statistical model required for ASR is the *language model* (LM), which gives a probability to each possible sentence of English, indicating how likely it is to be spoken. This model is trained on millions of words of text data. The LM allows the ASR engine to select the most plausible option between similar sounding sentences: for example “david beckham played for manchester united” will receive a higher probability than “david peck ham plaid foreman chester you night hid”. The two models are combined in the *decoder*, the software that is used by the engine to search for the most likely sentence given a portion of input audio.

Live TV broadcasts are very different in their spoken content to other types of speech recordings such as read speech. In particular, there are likely to be frequent changes of speaker, a wide range of accents, and – outside the studio – noisy background conditions. This means that AMs trained on other speech databases often work poorly for TV broadcasts. Fortunately, for broadcast media, the availability of large quantities of captioned recordings allows accurate models to be trained on more closely matched data, such as that described in the previous section. However, a major difficulty of training the AM on broadcast media is aligning the captions to the correct portion of the audio, due to the lack of accurate time stamps (particularly on live-originated material) and the fact that the captions, whilst accurate in their underlying meaning, often differ from a word-for-word transcription. We have investigated several different approaches to this problem, using “lightly supervised” methods, where the captions are used as a general guide for AMs trained on other data, allowing the accurate phoneme timings to be identified where possible (7). Material that cannot be aligned is filtered out at this stage.

The further challenge of dealing with noisy background conditions can be solved by using state-of-the-art deep learning techniques (4). Deep neural networks (DNNs) pass the audio through many layers of neurons and are trained to predict phoneme probabilities. DNNs have been shown to be hugely more robust to noise than previous methods, and have played a key role in enhancing the performance of ASR to usable levels across many tasks. In this project, we used DNNs of six layers deep, with 2000 neurons per layer (8). We also experimented with alternative deep architectures, including convolutional neural networks inspired by image processing. For inputs, the audio is processed into 40 log mel-scaled filterbank bins at 10 ms time points, and the networks use 11 of these frames for every prediction. We found that it was very important to normalize the audio for every sentence to adjust for changing environment and speaker conditions.

The LM we used is a 4-gram model, meaning that the probability of a word is modelled along with the previous three words of context. Our model used over 5 million such 4-word groupings. As the ASR decoder effectively performs a search over all word in its vocabulary, it is necessary to limit the total size of the vocabulary to avoid this process becoming too slow. We elected to use a vocabulary of 150,000 distinct words, which is enough to cover almost all English words along with a large number of proper names. The LM is pruned to ensure acceptable limits on memory usage and the speed of decoding. Decoding uses the Kaldi toolkit (http://kaldi-asr.org).

DOMAIN SPECIFIC REQUIREMENTS

Previous joint Broadcast / Academic research projects had revealed that there was considerable divergence between the evaluation of ASR output within the academic community and those creating SDH captions. For Broadcast, and to ensure that certain regulatory requirements or targets are met, the goal is to produce a verbatim transcript of the output. These transcripts must be fully punctuated, need to be presented as grammatically structured language (this can be hard for certain conversational styles of speech) and the maximum speaking rate must not become excessive. Additionally it is necessary to disambiguate complex sections with many simultaneous speakers, and disfluencies (such as “huh or “erm”) are usually removed unless pertinent to the understanding of the content (i.e. an interviewee stumbling to answer a question).

Academic assessments of ASR output aim to measure the model’s fidelity to the noises uttered by the speaker; if they have specific verbal tics, or meander in their use of words then such systems should include all these utterances. Conversely punctuation and ‘true’ sentence casing that are required to make captions comprehensible are not typically part of the academic review scheme. As a result, mechanisms for handling disfluencies in an appropriate way, punctuation models, methods to handle sentence casing and where possible identify and correctly capitalise named entities are required in order to turn standard ASR engine output into usable text for captioning. To do this, we used an approach inspired by methods in statistical machine translation (SMT). Using a corpus of text with punctuation and capitalisation, we automatically stripped off these features to create bare text. Using Edinburgh’s Moses (http://www.statmt.org/moses/) statistical machine translation system toolkit, we trained an SMT system to “translate” between the stripped text and the original. These models can then be applied to the ASR output to generate text that is more suitable for captioning (9).

TESTING

We analysed 145 pieces of ASR-generated text from news clips that came from many of the BBC's regions, with the content covering a wide variety of subjects, containing regional accents, audio that was very clear, for example featuring journalists reading from scripts in a studio, and audio that was recorded at the side of busy roads while passers-by were interviewed. Clip lengths ranged from the very short - containing as few as 22 words - to longer pieces such as the weather, or in-depth reports that contained as many as 318 words. The text was marked for accuracy using a simple calculation:

total words – incorrect words = accurate words

accurate words\total words *100 = percentage accurate words

The number of errors produced was marked as above, while the number of missing words was recorded separately. The purpose of the project was to evaluate text for its suitability to feed into the captioning workflow at a point prior to transmission, the goal ultimately being to reduce the amount of operational time spent on preparing for news captioning. It was assumed that the text produced by the ASR engine would never be 100% accurate and therefore could not go to air without human intervention. It was then determined that a combination of the percentage accuracy of the text and the percentage of missing words provided a reasonable guide as to how suitable text is for quick correction, i.e. proof-reading, without reference to the media by a captioner.

This calculation differs from Ericsson's standard method of marking live captioning for accuracy, which is the NER model:

> *This calculates accuracy as the total number of words (N) minus edition errors (E) and recognition (R) errors, divided by the total number of words, or*
>
> (N – E – R / N) x 100/1

The NER model is designed to measure the accuracy of caption text as transmitted while providing a measure of the viewer experience, and therefore was not suitable for use with this project.

The percentage accuracy of pieces marked ranged from 40.91% to 100%, while the percentage of words missing ranged from 0% to 68.52%. The pieces were used by captioners in a mock-up of the live production workflow and it was determined that the optimal textual accuracy was 95%, with a minimum threshold for usable text being 90% accurate with no more than 10% of the content missing. Pieces that met these criteria could potentially be proof-read swiftly by captioners before transmission, as opposed to the existing workflow which has the captioner transcribing the audio manually.

Out of the 145 pieces reviewed, 89 were above 90% accuracy, and of those 52 were above 95% accuracy:

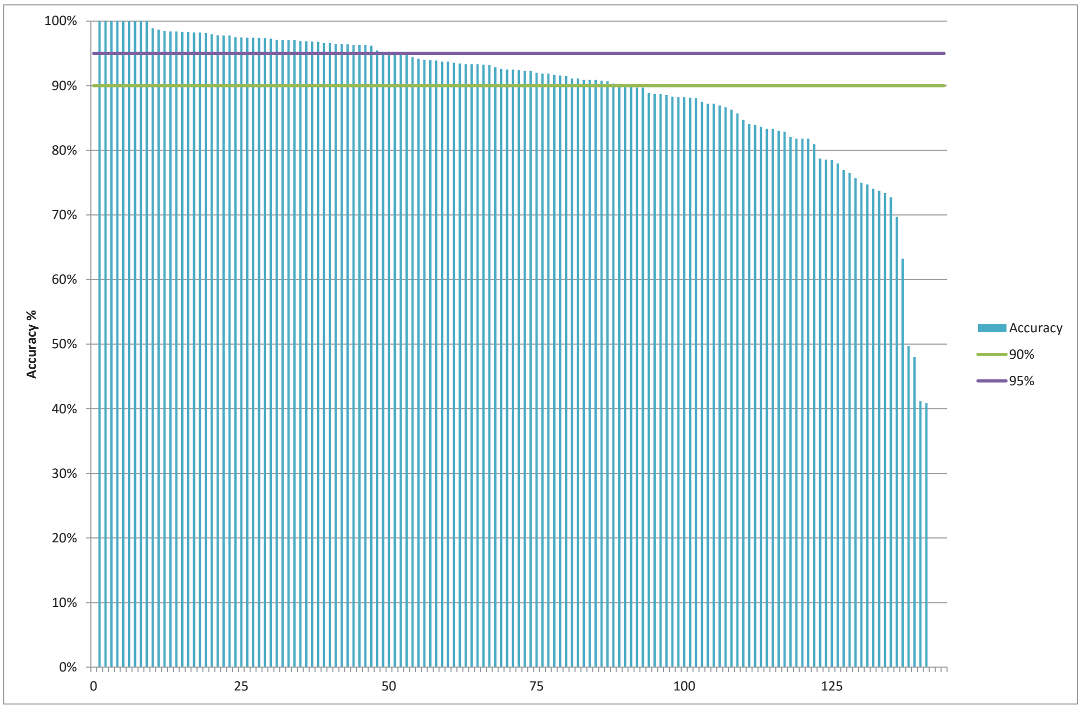

Figure 1 – Test Accuracy

PLATFORM IMPLEMENTATION

The final stage of the project is implementation of the completed model within a suitable software infrastructure, such that it can be integrated into the production workflow. The University of Edinburgh, along with is spin-off company Quorate Technologies, has considerable experience of taking a speech engine from a laboratory proof of concept to a fully functional platform, having previously built upon the results of the AMI. This is the approach the project will follow.

For implementation into a broadcast workflow, the API of the componentised form of the engine can be utilised to access the engine. The live captioning process for News output is largely governed by the creation of running orders within the relevant broadcaster's Newsroom Computer System; prior to transmission these running orders are shared with automation platforms and third parties (commonly via the Media Object Server or MOS communications protocol) and at this point the running order can be analysed for pre-recorded content. Typically this content is indicated in specific automation fields, using some form of unique asset or material ID. Ericsson intends to utilise the presence of these IDs to indicate the completion of a given clip – it can be assumed from the standard newsroom / studio workflow that the presence of a valid ID is an extremely likely indicator that the clip has been published within the video production system. These IDs (and any changes / additions after the initial reception of the running order) can be used to extract the clips from the production system using a simple interface, and transcoded on receipt into an audio format suitable for ASR processing. It should be noted that compression is best to be minimised or avoided for the best ASR results – as speech models do not respond in the same way as the human ear to lossy compression, and much valid ASR data can be lost; additionally any lossy compression in the upstream chain should be avoided where possible.

The ASR API can then be called using this audio file, along with any sub-domain specific metadata that may be necessary for optimising the model at runtime (i.e. 'Weather clip'). The API will return an XML based transcript of the audio (Ericsson works with EBU-TT as a timed text interchange file), containing both the transcript and word level timing, and the platform's recognition confidence score for the clip. The live captioning platform can then triage the incoming transcripts according to the confidence score; those with scores of 95% or greater accuracy will be presented in the user interface as requiring little additional QC or correction, those with 90-95% accuracy but fewer than 10% of the words likely to be wrong will be presented as require moderate correction, and the remaining transcripts will be abandoned (or stored for feedback to the speech recognition team for further refinement of the engine). In this way the captioners are able to utilise only the transcripts they are confident they will be able to correct prior to transmission, and will not waste valuable preparation time fixing long sections of poor recognition.

CONCLUSIONS

It is clear from the experimentation carried out to date that significant productivity and quality improvements can be achieved in live captioning solutions by the application of domain-specific Automatic Speech Recognition. The changeable nature of live broadcast News creates a challenging environment, and it is clear that there is a natural tension in between Productivity, Quality and available Time – the least-effort approach (originating all output in real time) is likely to fail on word accuracy and latency, whilst time / resources and available media will not favour a very manual approach trying to prepare all pre-recorded media pre-transmission.

A reliable ASR platform, with importantly a reliable ASR confidence scoring mechanism, are most likely to offer the best balance of accurate automated text generation and avoidance of unnecessary correction. If these can be obtained, then it is more likely that ASR can be implemented without frustrating the viewer, the captioner or those in charge of the captioning budget.

ACKNOWLEDGEMENT

This work was supported by EPSRC Programme Grant EP/I031022/1, Natural Speech Technology (NST), and an EPSRC Impact Acceleration Award.

REFERENCES

1 Armstrong M., 2013. The Development of a Methodology to Evaluate the Perceived Quality of Live TV Subtitles, BBC R&D White Paper WHP259 Oct 2013. pp 6 to 8.

2 Ofcom, 2013 - 2015. The quality of live subtitling and Ofcom. May 2013. pp 26 to 34. Measuring live subtitling quality - Results from the fourth sampling exercise. November 2015. Pp29 to 31

3 Driesen J., Renals S., 2013, Lightly Supervised Automatic Subtitling of Weather Forecasts, 2013 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2013), Olomouc, Czech Republic. Dec 2013.

4 Hinton G et al, 2012. Deep Neural Networks for Acoustic Modeling in Speech Recognition, IEEE Signal Processing Magazine, November 2012, 29(6):82–97.

5 Bell, P et al, 2015. The MGB Challenge: Evaluating Multi-genre Broadcast Media Recognition, 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2015), Scottsdale AZ, USA. Dec 2015.

6 Richmond, K, Clark, R and Fitt, S, 2009. Robust LTS rules with the Combilex speech technology lexicon. In *Proc. Interspeech*, pages 1295-1298, Brighton, UK, September 2009.

7 Bell, P. and Renals, S., 2015, A system for automatic alignment of broadcast media captions using weighted finite state transducers, IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2015), Phoenix, AZ

8 Bell, P, Yamamoto, H, Swietojanski, P, Wu, Y, McInnes, F, Hori C, and Renals, S (2013). A lecture transcription system combining neural network acoustic and language models. In *Proc. Interspeech*, Lyon, France, August 2013.

9 Driesen, J, Birch, A, Grimsey, S, Safarfashandi, S, Gauthier, J, Simpson, M, and Renals, S. (2014). Automated Production of True-cased Punctuated Subtitles for Weather and News Broadcasts. In *Proceedings of Interspeech*, Singapore, September 2014.