Institution: University of Edinburgh

Unit of Assessment: 11

Title of case study: Neural network speech recognition algorithms lead to accurate speech-to-text transcription deployed in diverse products and services

Period when the underpinning research was undertaken: 2004 – 2020

Details of staff conducting the underpinning research from the submitting unit:

Name(s):	Role(s) (e.g. job title):	Period(s) employed by submitting HEI:
Peter Bell Steve Renals	Reader Professor	2010 – present 2003 – present

Period when the claimed impact occurred: August 2013 – 2020

Is this case study continued from a case study submitted in 2014? No

1. Summary of the impact

The University of Edinburgh (UoE) has developed novel neural network algorithms in automatic speech recognition, which have enabled speech recognition capabilities not previously available. The research was commercialised by two companies: AI start-up Emotech, and UoE spin-out Quorate. Between them, they have used the research to generate products with a wide array of applications, from an AI assistance robot, to a tool that lets editors of the UK's Hansard automatically convert spoken proceedings in the Houses of Parliament into text. Through Quorate, the research has enabled subtitling provider Red Bee Media to improve the efficiency of their services, widening access to streaming and television for millions of viewers with hearing loss, while Emotech is currently collaborating with Huawei to distribute a language-learning platform across rural China.

2. Underpinning research

Since 2004, a research team at the University of Edinburgh's (UoE) Centre for Speech Technology Research (CSTR) has investigated automatic speech transcription, across a wide diversity of languages, domains and acoustic conditions. This work has been supported by a number of large-scale projects including the EU projects AMI (01/01/2004 – 31/12/2006) and AMIDA (01/10/2006 – 31/12/2009), and the EPSRC programme grant Natural Speech Technology (NST; 01/05/2011 – 31/07/2016); these projects were all coordinated by Professor Steve Renals (Chair of Speech Technology).

The AMI and AMIDA projects investigated the recognition and understanding of multiparty dialogue in meetings recorded using multiple distant microphones [3.1]. This work extended the state-of-the-art to challenging acoustic conditions including overlapping talkers, multiple acoustic sources, noise, and reverberation. The work on distant speech recognition was continued by Renals in the NST Programme Grant, with his PhD student Pawel Swietojanski and postdoc Arnab Ghoshal. They developed the first convolutional neural network architecture for recognising speech recorded using multiple distant microphones [3.2], that has been highly influential in the field, and was later taken up by Emotech – a collaboration beginning at an NST User Group meeting in 2016.

The work of Renals' team in NST pioneered novel neural network approaches to speech recognition, resulting in techniques surpassing the state of the art. Dr Peter Bell was the leading

researcher on the development of transcription systems applicable to a wide variety of complex domains including broadcast TV and TED Talks. The transcription system presented in [3.3] was applied to the benchmark TED Talks task, with error rates being reduced by 16% in comparison to the previous state-of-the-art systems developed by MIT, Karlsruhe, RWTH Aachen, and NICT.

NST also focused on transferring speech recognition systems to new languages, using deep learning to share phonetic representations across languages. Ghoshal, Swietojanski, and Renals pioneered these multilingual knowledge transfer approaches [3.4], and this work now forms the basis of Emotech's multilingual speech recognition technology.

Current commercial speech recognition systems use supervised trained based hand-transcribed recordings that are specific to the domain of use. This makes adaptation to new domains very expensive since transcribed training data needs to be collected. Bell and Renals developed approaches to training speech recognition systems in the absence of verbatim labelled training data (referred to as lightly-supervised training, enabling systems to be trained from TV subtitles or parliamentary proceedings), first used on the TED Talks system [3.3]. A particular contribution of this work was the development of an efficient and effective alignment algorithm [3.5] capable of aligning audio to partially incorrect and incomplete transcriptions having an error rate of 30% or more compared to the true reference. This algorithm also allows far-from-verbatim transcripts (such as TV subtitles or the official record of parliamentary sessions) to be aligned to the audio. These advances have been taken up by Quorate Technology and Red Bee Media, and used for both alignment and automatic speech transcription.

Typical speech recognition systems output a stream of words, without punctuation or capitalisation. Punctuation is very important -- both for readability by human users and for downstream automatic tasks such as machine translation. Klejch, Bell, and Renals developed an automatic approach to punctuating speech recognition output using recurrent neural networks [3.6], an approach that has been taken up by Quorate Technology.

The research on light supervision and punctuation was also supported by the EU SUMMA project (01/02/2016 – 31/01/2019; coordinated by Renals) and the US IARPA Material SCRIPTS project (25/09/2017 – 31/05/2021; PIs Renals and Bell).

3. References to the research

3.1. Renals, S., Hain, T., & Bourlard, H. (2007). Recognition and understanding of meetings: The AMI and AMIDA projects. In *IEEE Workshop on Automatic Speech Recognition & Understanding* (pp. 238-247). https://doi.org/10.1109/ASRU.2007.4430116 **(Invited talk; 156 citations).**

3.2. Swietojanski, P., Ghoshal, A., & Renals, S. (2014). Convolutional Neural Networks for Distant Speech Recognition. *IEEE Signal Processing Letters, 21(9)*, 1120-1124. https://doi.org/10.1109/LSP.2014.2325781 **(194 citations**)

3.3. Bell, P., Yamamoto, H., Swietojanski, P., Wu, Y., McInnes, F., Hori, C., & Renals, S. (2013). A lecture transcription system combining neural network acoustic and language models. In *Interspeech.* https://www.isca-speech.org/archive/interspeech_2013/i13_3087.html **(28 citations)**

3.4. Ghoshal, A., Swietojanski, P., & Renals, S. (2013). Multilingual training of deep neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing,* 7319-7323. https://doi.org/10.1109/ICASSP.2013.6639084 **(197 citations)**

3.5. Bell, P., & Renals, S. (2015). A system for automatic alignment of broadcast media captions using weighted finite-state transducers. In *IEEE Workshop on Automatic Speech Recognition and Understanding* (pp. 675-680). https://doi.org/10.1109/ASRU.2015.7404861 **(7 citations)**

3.6. Klejch, O., Bell, P., & Renals, S. (2017). Sequence-to-Sequence Models for Punctuated Transcription Combing Lexical and Acoustic Features. In *IEEE International Conference on*

Acoustics, Speech and Signal Processing (pp. 5700-5704). https://doi.org/10.1109/ICASSP.2017.7953248 **(32 citations)**

Citations based on Google Scholar, 2020-12-03.

Key research grants

- European Commission: AMI (506811, GBP1,757,637); AMIDA (033812, GBP1,258,536); SUMMA (688139, GBP1,332,742)
- EPSRC: Natural Speech Technology (EP/I031022/1, GBP7,650,991)
- ODNI IARPA: SCRIPTS (FA8650-17-C-9117, GBP1,326,444)

4. Details of the impact

In the REF2021 impact period, the research led to the creation of one start-up company in 2014 (Emotech), and the increased growth of another, already in existence since 2012 (Quorate). Between them, their clients include government agencies and commercial suppliers, and they have succeeded in using the University of Edinburgh (UoE) technology to pioneer new ways of generating parliamentary records, to increase access to television programming for people with hearing loss, and to enable education facilities in rural parts of the globe.

From its founding in 2014, AI company Emotech has used the Edinburgh research outlined above to develop new products and generate collaborations with global partners. During the REF impact period, Emotech created over 50 jobs across the UK and China, and attracted a strategic partnership with Huawei [5.1, paras. 3, 10].

Renals' PhD student Swietojanski (co-author on 3.2-3.4) took up post with Emotech in 2016, and worked on the development of the company's core product: Automatic Speech Recognition (ASR), based on the UoE research on distant speech recognition. This formed the basis for Emotech's first invention, an AI robot assistant named Olly, which went on to win four CES Innovation awards in Las Vegas in 2017; the first time an AI machine had achieved such recognition [5.1, para. 1].

ASR also became the foundation for the company's second major product, a Virtual Education Platform (VEP). The platform is a tool for learning to speak English, and is the first of its kind, using AI to mimic natural human interaction [5.2]. The UoE research on multilingual knowledge transfer has specifically helped minimize the need for training data in building the tool, where previous models would have been "economically prohibitive to obtain the considerable amount of data required for conventional model training" [5.1, para. 7].

Emotech's VEP led to a collaboration with Huawei and is now being rolled out across rural areas of China. Huawei's Education Informatisation programme already provides access to education for children and adult learners in more than 1800 institutions across 1000 cities; the VEP is now a component of that programme [5.1, para.10].

Huawei publicly cited Emotech's "advanced technology in voice and multi-modal AI" as being key to the collaboration, and the platform has been hailed by UNESCO as "very important technology" and the "key for human beings to learn new things and skills faster" [5.2, paras. 8-9]. As well as China, the platform has been sold for use in the Middle East, South America and South Africa [5.1, para. 10].

Spinout company Quorate was created as a result of research published in [3.1], in 2012 [5.3, paras. 2, 14]. However, with only an initial [text removed for publication] start-up investment in 2012-2013, Quorate has developed during the REF impact period to be a company able to grow entirely on its portfolio of customer-driven income. In 2020, the company's revenue was [text removed for publication] [5.3, antepenultimate para.].

Quorate built on its initial foundations, using the latest UoE research to create its main product, QSpeech, a speech recognition tool with multiple practical applications, that has had significant success in both subtitling and in editorial work, in commercial and public service arenas [5.3].

QSpeech has been sold on an extended trial basis to Hansard, the UK's parliamentary record, for the purposes of recording and automatically transcribing proceedings in the Houses of Parliament, and generating text transcripts to accompany audio recordings from the debating chamber and committee meetings [5.3 paras. 5-6; 5.4, p. 339]. These proceedings are made publicly available, with millions of unique page views on debates such as those prompted by public e-petitions [5.4, p. 338]. Between 1 September and 23 October 2020, 1,202 hours of recordings were processed by QSpeech to assist the work of the UK Parliament's 90 professional Hansard reporters [5.4, p. 339]. The Deputy Editor of the Official Report (Hansard) confirms:

> *The level of accuracy attained is impressive at around 90% on average … [it helps reporters] concentrate on clearly presenting what has been said, and on understanding the argument, rather than on typing.* [5.4, p. 339]

Quorate's technology is also in use in parliaments in New Zealand, Australia, Canada, Guyana, Jersey and the Isle of Man [5.3, para. 6].

QSpeech has brought enormous benefits to the editors who create Hansard. It has streamlined the process of transcription considerably, automatically transcribing large amounts of raw text, which can then be edited for consistency, clarity and accuracy. Thanks to QSpeech, the Official Report is now produced more efficiently: typically four times quicker than compared to a fully manual process [5.3, para. 9]. Editors have been able to direct their focus towards the editorial process, refining finer details of grammar, punctuation and syntax. Hansard staff have been effusive in their praise of the software, calling it "brilliant", "absolutely invaluable", "a big help psychologically", and "a great leap forward for House Reporters" [5.5, paras. 26-33]. Hansard also reports that, "some reporters who have repetitive strain injury have found the live recogniser a great help" [5.5, para. 27], and praises the technology's potential for mitigating RSI [5.5, para. 8].

QSpeech continued to evolve in line with ongoing UoE research. In 2015-16 an algorithm was added to it, enabling the feature of word-based alignment of a transcript to its accompanying audio file, along with the ability to timestamp the accompanying text [5.3, para. 15]. Hansard states that this has enabled them to provide subtitling for hours of Parliamentary video streaming, which is useful "from an accessibility point of view" [5.5, para. 12], and also allows anyone using the service to search for a word in the text file and automatically retrieve its corresponding audio-visual clip [5.5, para. 19].

Another major commercial and public service beneficiary of the research has been subtitling provider Red Bee Media. Working both with Quorate and directly with UoE, Red Bee Media has created a "robust alignment solution" based on the research that has radically improved its production processes [5.6, para. 3].

Red Bee currently subtitles over 200,000 hours of television and streaming per year, for companies including the UK's BBC and Channel 4, Australia's SBS and the US's ABC. In the UK, broadcasters commonly use services such as Red Bee to help them meet the subtitling requirements set by Ofcom "in a cost-effective manner" [5.6, para. 1].

Red Bee's new technology, and the resultant improvements to the production process, has made meeting these targets more economical and more efficient. The company states that the technology has "allowed us…to significantly reduce the cost of our workflows", impacting two

key areas of production; caption alignment and transcription. Red Bee estimates that It has currently improved its efficiency in the field of caption alignment by 20% (directly attributable to the UoE research), and in transcription by 15% (attributable to technology developed by the company, spring-boarded by the research) [5.6, paras. 3-4].

These cost savings have proved attractive to Red Bee clients, and have enabled Red Bee to offer their subtitling services at price points "not previously feasible" [5.6, para. 4]. This in turn has enabled a larger quantity of unregulated live streaming to be subtitled, including on YouTube and Facebook, increasing access for the 11,000,000 people with hearing loss in the UK alone, as well as international viewers [5.6, paras. 1, 4].

The sustained success and growth of Quorate and Emotech during the impact period has not only brought commercial benefits to those companies, but has also, through their clients, lead to societal changes with broad and significant impact. The research has fundamentally changed the way in which the UK's parliament proceedings are transcribed and its public record delivered. In addition, it has enabled novel and progressive forms of education in the field of language learning, and increased access to broadcast media for many millions of people across the globe with hearing loss.

5. Sources to corroborate the impact

5.1 Letter of corroboration from Emotech
5.2 Leps. (2019, September 20). Emotech, a well-known British AI company, joins Huawei in launching a multi-modal English teaching assistant program. Retrieved November 17, 2020, from https://www.rayradar.com/2019/09/20/emotech-a-well-known-british-ai-company-joins-huawei-in-launching-a-multi-modal-english-teaching-assistant-program/
5.3 Letter of corroboration from Quorate
5.4 Commonwealth Parliamentary Association. (2020). The Parliamentarian 2020: Issue Four - Social Media and Democracy in the Commonwealth. Retrieved December 7, 2020, from http://www.cpahq.org/cpahq/Main/Our_services/Publications/CPA_Publications/The_Parliamentarian/Latest_Issue/Main/Publications/The_Parliamentarian/Current%20Issue.aspx?hkey=7c22a05d-39c5-4d67-9eda-782e444212b8
5.5 Letter of corroboration from Hansard
5.6 Letter of corroboration from Red Bee Media