Institution: University of Sheffield

Unit of Assessment: 11 - Computer Science and Informatics

Title of case study: ASR: Commercial, societal and cultural benefits of new advanced Speech Recognition Technology

1. Summary of the impact

One of the world-leading systems for large-vocabulary Automatic Speech Recognition (ASR) has been developed by a team led from the University of Sheffield. This system, which won the international evaluation campaigns for rich speech transcription organised by the US National Institute for Standards and Technology (NIST) in 2007 and 2009, has led directly to the creation of one spin-out, been largely instrumental in the launch of a second, has had significant impact on the development and growth of three existing companies, and has made highly advanced technology available free for the first time to a broad range of individual and organisational users, with applications including language learning, speech-to-speech translation and access to education for those with reading and writing difficulties.

2. Underpinning research

Until recently, the impact of ASR technology has been limited by the difficulty of ‘domain adaptation’: to move to a new vocabulary, a new environment, a new speaker or a new task. The underpinning research behind this case study has addressed this issue.

The research dates back to 2004, with the start of an EU FP6 Integrated Project AMI (Augmented Multiparty Interaction), which was succeeded by AMIDA (the DA stood for Distant Access). These projects addressed the automatic understanding of multimodal data generated in meetings. Professor Thomas Hain at the University of Sheffield led the automatic speech recognition team (between 8 and 15 researchers from Sheffield, the Univ. of Edinburgh, the Swiss IDIAP Research Institute, the International Computer Science Institute (ICSI), at Berkeley, Brno University of Technology, and the University of Twente). The AMI team improved the state of the art in all aspects of meeting-speech recognition.

The Sheffield work focused on several key areas of machine learning in speech recognition: front-end feature extraction, automatic system optimisation using sampling techniques, automatic language model data collection and adaptation and methods for improving ‘far field’ performance (i.e. using distant microphones). An example of our optimisation work is [R3] where we show that significant performance gains can be obtained from automatic optimisation of decoder parameters using a gradient descent method. This contributed significantly to the development of a real-time decoder for meetings, **JUICER** [R5, R2], based on weighted finite state transducers (WFSTs). The decoder includes novel forms of WFST creation making it possible to gauge dependence on underlying acoustic and language models. Overall decoder comparison shows that Juicer pruning is highly efficient and can even allow data type tuning.

The training of statistical models in these systems illustrates the complexity and scale of the task [R1]. Models are trained on a variety of sources: more than 2300 hours of audio and 2 billion words of text. Acoustic model training (at the time of publication) required almost a complete CPU decade, the recognition process operated in more than 20 stages, leading to more than 3000 automatically generated individual computing grid processing steps. To facilitate this complexity Sheffield developed the resource optimisation toolkit (ROTK), which brings a novel graph/metadata approach to system construction, allowing systematic optimisation of system structures and deployment. ROTK now forms the backbone of the webASR service [R4] and is licensed by spinout Koemei.

Further significant developments were made in language model data acquisition and lexicon generation. The wide range of meeting topics requires adaptive and learnable structures for language model and lexicon. In 2012 we extended our earlier previously published work on automatic language model data acquisition [R6] using search models and showed how wordlists can be augmented and improved in semi-unsupervised and fully unsupervised fashion. In 2008 we implemented an efficient approach to lexicon and dictionary creation.

We have also contributed to new methods for multi-channel segmentation of meeting audio. We have further proposed a novel solution for identifying echo speech signals in teleconferencing. The method can be used to remove distorted versions of the original speech played through loudspeakers at conference participant locations. We show that our method is very robust and can eliminate echo with a very high degree of accuracy.

Primary drivers for impact in ASR the international evaluation campaigns organised by NIST (the US National Institute for Standards and Technology). Both industry and academic groups participate in those competitions. Sheffield-led systems won the competitions for rich speech transcription in 2007 and 2009 [text removed for publication]. Because of their highly adaptive capabilities and excellent generalisation to different domains these systems are generating impact in a variety of real-world application areas.

The product of the underpinning research was a configurable software system for automatic speech recognition. The system has several variants which are easily targeted at particular application areas, as detailed below. It is one of the most advanced systems worldwide and continues to be developed, for instance in the EPSRC programme grant in Natural Speech Technology (2011-2016), and the EU FP7 project 'Documeet' (2012-2014).

3. References to the research *(*denotes outputs which best demonstrate underpinning research quality*).

R1. *T. Hain, L. Burget, J. Dines, P. N. Garner, F. Grezl, A. el Hannani, M. Huijbregts, M. Karafiat, M. Lincoln & V. Wan (2011). *Transcribing meetings with the AMIDA systems*. In IEEE Transactions on Audio, Speech and Language Processing, 20, 486-498. doi: 10.1109/TASL.2011.2163395

R2. *P. N. Garner, J. Dines, T. Hain, A. el Hannani, M. Karafiat, D. Korchagin, M. Lincoln, V. Wan & L. Zhang (2009). *Real-Time ASR from Meetings*. In Proc. Interspeech'09, pp. 2119-2122

R3. A. el Hannani & T. Hain (2010). *Automatic Optimization of Speech Decoder Parameters*. InIEEE Signal Processing Letters , pp. 95-98

R4. T. Hain, A. el Hannani, S. Wrigley & V. Wan (2008). *Automatic speech recognition for scientific purposes - webASR*. In Proc. Interspeech'08, pp. 504-507

R5. *D. Moore, J. Dines, M. Doss, J. Vepa, O. Cheng, T. Hain (2006). Juicer: A weighted finite-state transducer speech decoder. In Machine Learning for Multimodal Interaction, Springer, pp 285-296.

R6. V. Wan & T. Hain (2006). *Strategies for Language Model Web-data Collection*. In Proc. ICASSP'06

4. Details of the impact

The key output of the research, a configurable software system for automatic voice recognition, was made widely available through the launch in 2009 of **webASR** (www.webASR.org), the world's first free web platform for offline speech recognition. It takes any speech file and returns a transcription, providing a service to its approximately 300 registered users that was not previously available. Of these users, about 20% are from industry, 30% from academia and the rest individuals. Up to the end of 2012 users had uploaded 4000 audio files, amounting to 240 hours of speech. In addition to the online platform, the system can also be accessed through an Application Programming Interface (API) allowing companies and organisations such as dev-audio, Sonocent and the BBC to integrate the software into their own systems.

Commercial Impact

In order to maximise their commercial impact, the AMI and AMIDA projects incorporated a range of 'mini projects' with industrial partners which served as pilot studies for possible applications of the systems and technology developed during the research project. The success of these pilots, as well as the free availability of the ASR software through webASR's API, have led to the technology being adopted by a range of commercial partners worldwide.

Creation of start-ups

Koemei [S1] was founded in 2010 in Martigny, Switzerland with close involvement from Hain for the sole purpose of commercialising the AMI speech recognition system, which served as a starting point for what has now evolved into the Koemei recognition engine. The company provides a cloud service (approx.. 3000 registered users), which leverages its automatic speech recognition technology, with features including fully automatic and crowd sourced transcription and captioning, media indexing and search as well as a suite of tools focused on online education for universities, lecturers and students. In addition to its online service, Koemei provides professional services in the form of customisation of its technology and integration with third party services (e.g. media hosting providers, online video players, learning management systems). Its clients include IMD business school, the University of Geneva, UC Berkley, Columbia University, City University New York, Kaltura, UDemy, and Al Jazeera. The company currently has 6 full time employees and has received 500k CHF investment from a mix of private and institutional investors as well as 1m CHF in subsidies from Swiss and EU project funding agencies. Its revenue to date totals around 600k CHF and it was ranked 63rd (2011) and then 29th (2012) of Top 100 Swiss Start-ups. It has been recognised with numerous accolades including: a presentation at the DEMO Conference Santa Clara 2012; one of 15 contestants selected for the TechCrunch Disrupt New York 2012 start-up battlefield; a place in the top 5 favourite companies at 500 Start-ups Demo Day 2013 [S2]; and was one of 36 companies selected by the World Economic Forum as Technology Pioneers 2014, a title that aims to recognise "*companies, normally in a start-up phase, from around the world that are involved in the design, development and deployment of new technologies, and are set to have a significant impact on business and society. Technology Pioneers must demonstrate visionary leadership and show signs of being long-standing market leaders – their technology must be proven. Each year the World Economic Forum chooses a select number of Technology Pioneers from hundreds of applicants"* [S3].

Hain's research was also largely instrumental in the creation of **Quorate Technology**, a spin-out founded in October 2012 and based in Edinburgh with 5 full- and one part-time employees. Quorate works with meeting capture and recognition technology and uses the knowledge of these areas developed in the AMI and AMIDA projects by Edinburgh, Sheffield and IDIAP directly in its core product. The company is targeting the police, justice, defence, broadcasting and enterprise sectors and has ongoing relationships with a number of large systems integrators who are evaluating the use of their technology with their customer data analysis systems. They have contracts with a major European Defence Contractor to supply demonstration systems in a number of their business areas. Quorate is supported by a Scottish Enterprise SMART grant, and Scottish Enterprise Edge Award. According to Quorate's CEO, "*The research outcomes of the AMI and AMIDA projects in the field of speech recognition were essential steps on the path to Quorate being formed and commercially available products being developed...The research conducted by UoS in this area had significant impact on the efficiency of delivery of services by Quorate*" [S4].

Impact on existing businesses

dev-audio was founded in 2008 and is now a profitable export business with clients around the world. Its product, microphone array recording hardware (the Microcone) is produced in volume and distributed by the Apple Store Online in Europe and Amazon in the USA. The company fully integrated their Microcone Recorder App with webASR in an AMIDA mini-project, creating a commercial prototype system for end-to-end meeting capture and recognition that was used in demonstrations and proved key in convincing clients and investors. They also used webASR to test and iterate their technology affordably and easily on state of the art systems. "*The integration with webASR throughout various phases of our business growth has allowed us to stay ahead of the game and differentiate our solution in the marketplace. This proved critical in garnering early financial and client support during the start-up phase. More recently, webASR has enabled us to objectively quantify the performance of Microcone's voice separation technology - a key step in developing the industry partnerships that'll accelerate our market growth*". Founder, devaudio [S5].

Klewel is a company that commercialises a comprehensive webcasting solution, complete with an audio-visual recording station linked to a web platform, allowing customers (including major companies and organisations such as Nestlé and UNICEF) to automatically reference, edit and publish content in total simplicity. Klewel used the AMI ADR platform as part of a mini-project with AMIDA in 2009 to explore speech-to-text features on Klewel lecture recordings. "*We recognised the technology as extremely promising as a new and innovative feature to develop our business"* says Klewel's Managing Director. *"It would offer our company and our end customers ...more visibility and more efficient knowledge management. For this reason, our company decided to launch a joint development project in 2013, one of whose key aims is to integrate ASR into Klewel search. We expect in the near future to include the technology in all of our products*."

Inferret Japan, a company employing 5 staff and commercialising speech recognition, natural language processing and information retrieval systems, including voice-driven smartphone search applications, English language learning for Japanese children, and a cloud service helping non-experts to integrate speech recognition into their own technology. It also directly develops and customizes speech recognition technology for several large clients including Hitachi Systems and the Eiken Foundation. Their work has benefited significantly from Hain's research, especially the Juicer package. Inferret's CEO explains: "*The availability of Juicer, its associated tools and research publications gave us an important competitive technology advantage, reduced our overall* *development costs and accelerated our time to market*" [S7].

Cisco, market-leading provider of telepresence systems, used the Sheffield research within the Media Experience and Analytics Business Unit (MXBU) of its Emerging Technology Group to help sharpen the use cases during product development, as part of the statistical model infrastructure [S6]. [text removed for publication]

Impact on Practitioners

Practitioners in several fields have benefited from the free availability of the software via webASR. Researchers within the speech community have benefitted in two ways: by integrating advanced components to improve their own existing speech recognition systems; and (for the many research groups worldwide who do not have the capability to run complex speech recognition systems of their own) by using webASR to obtain transcripts of high volumes of speech that would otherwise have taken too long or been too costly to be available. Professionals and individuals outside the speech research community have also benefitted, such as psychologists and lawyers transcribing their interviews, or lecturers wishing to make their lectures more widely available.

Societal impact

Organisations like the **Liberated Learning Consortium** (www.liberatedlearning.com) are working to increase access to higher education for those with reading and writing difficulties through the development of transcription platforms. LLC's International Manager welcomes webASR as "*a key milestone in this research area*" and explains its importance: "*The lack of captioning/transcription stems from a combination of exorbitant costs associated with traditional outsourcing, a lack of accessibility awareness, and dearth of available tools/techniques. Hosted transcription systems such as webASR introduce a promising and increasingly robust solution that can help address accessibility challenges. I applaud the development of webASR, especially its open availability to anyone who wishes to transcribe audio or video content".*

The technology also has important applications in the field of language learning, and has been adopted by **ITSlanguage**, a company bringing new technology into schools in the Netherlands. Its Founder explains the benefits: "*Exercises will be more communicative and interactive by using speech technology, challenging students to practice more and learning more by adaptive learning and comprehensible feedback. In addition, we aim to save valuable teachers' time by automatic assessment*". The first implementation of the technology is underway, with software and technology being made production ready, and pilots being prepared for November 2013. It will be integrated into the online test- and assessment system QUAYN, part of the most widely used system in education in Belgium and the Netherlands (>200 schools and >5 publishers). The Founder adds: "*The availability of ... Juicer allows us to integrate recognition capabilities at a much earlier stage than originally planned. Schools have limited budgets and we were looking for best quality speech technology that most schools can afford. Purchasing costly technology licences would make it impossible to produce an affordable solution, bringing this kind of innovation in education to a halt***"**.

The technology also plays an important cultural role in increasing understanding through its incorporation into the UK-English engine of the platform that supports an iPhone app for speech-to-speech translation ("voicetra 4u"). To end July 2013, the app had been downloaded 42,263 times and used to translate 169,204 utterances. The app was rolled out during the Olympics 2012 in London and over a third of these utterances were translated during the course of the games [S8].

5. Sources to corroborate the impact

S1. Letter from VP Technology, Koemei, confirming performance and use of ASR technology.

S2. Website for 500 Start-ups Demo Day 2013 showing Koemei selected in top 5: http://thenextweb.com/insider/2013/07/25/500-startups-accelerator-sixth-batch-demo-day-favorites/

S3. World Economic Forum report nominating Koemei as technology pioneer: http://www.weforum.org/community/technology-pioneers

S4. Letter from CEO of Quorate, confirming role played by our research in founding the company

S5. Letter from dev-audio founder confirming role of ASR technology in business development

S6. [text removed for publication]

S7. Letter from CEO of Inferret Japan confirming impact on their business

S8. Download and usage figures for the translation app on file.