Institution: University of Sheffield		
Unit of Assessment: B-11 Computer Science and Informatics		
Title of case study: Commercial impact in the rapidly growing market for automatic voice recognition services		
Period when the underpinning research was undertaken: 2000–2020		
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:**
Hain, T.	Professor of Speech and Audio Technology	2000–present
Saz Torralba, O.	Research Associate	2012–2016
Deena, S.	Research Associate	2014–2019
Period when the claimed impact occurred: 2018–present		
Is this case study continued from a case study submitted in 2014? N		

1. Summary of the impact (indicative maximum 100 words)

Work by Professor Thomas Hain and colleagues at the University of Sheffield underpins the automatic speech recognition (ASR) systems of VoiceBase, a major US provider of voice data analytics services to companies operating large call centres. Thanks to state-of-the-art ASR tools provided by Sheffield, VoiceBase is able to accurately transcribe and analyse over a billion minutes of calls per year, offering their customers the ability to mine their entire volume of calls for a wealth of nuanced information (e.g. relating to fraud prevention and performance management). The competitive edge provided by Hain's technology has allowed VoiceBase to grow their business significantly by adding at least $100m to the value of the company and attracting clients such as Uber, Home Depot, NASDAQ, Delta Dental, and GrubHub.

2. Underpinning research (indicative maximum 500 words)

ASR for conversational speech is a challenging research problem, particularly in the context of adverse acoustic conditions such as over the telephone or in multi-party meetings. Commercial ASR can be achieved only by systems that combine numerous state-of-the art components, many of which are machine learning models that require massive volumes of speech data to be processed with computational efficiency. Professor Thomas Hain's team at the University of Sheffield has addressed these issues since 2000, developing research in three key areas:

Robust methods for conversational speech. Hain's work on ASR of telephone conversations was first developed in the context of international evaluation campaigns organised by the US National Institute for Standards and Technology (NIST) (2000-2004). This research yielded highly adaptive system architectures for ASR that encompassed both lexical and acoustic variations. Building on this, research was conducted for the EU projects Augmented Multi-party Interaction (AMI) (2004-2006) and Augmented Multi-party Interaction with Distance Access (AMIDA) (2006-2009), which focused on the automatic understanding of multimodal data generated in meetings **[R1]**. Sheffield addressed several key areas of machine learning in ASR: improvements in 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). Sheffield-led systems won NIST competitions for rich speech transcription in 2007 and 2009.

The work on the recognition of conversations in meetings (2004-2009) led to advances in the robustness of model estimation and system structure design. Model robustness, especially for novel deep learning architectures, was achieved by enhanced confidence-based methods to filter and select training data. The research community adopted so-called sequence training of deep neural networks for improved robustness using a state-level supervision method developed by Sheffield. A combination of such methods was shown to allow efficient and highly accurate filtering of training data **[R2]**.

Adaptation to background conditions. Starting in 2011 with the EPSRC-funded Natural Speech Technology (NST) programme grant (2011-2016), Hain's team developed novel methods for adaptation to complex and non-stationary background conditions **[R3]**. Unsupervised switching of background models was developed in the context of classical Gaussian mixture models and expanded to neural networks. A novel technique was developed – acoustic latent Dirichlet allocation (aLDA) – which allowed acoustic models to be informed by subtle acoustic variations. Adaptation of language content was shown to be highly effective in ASR of multi-party meetings using acoustic and text LDA **[R4]**.

Adaptation was further facilitated by metadata information pertaining to the speaker and domain. The team developed a novel method for diarisation of multi-party recordings that achieved outstanding performance even with overlapping speakers. A novel extension involved so-called h-vectors, which improved the characterisation of speakers for adaptation purposes **[R5]**. Novel models for sentiment and emotion recognition were introduced to enable ASR sensitivity to unusual conditions (e.g. a speaker becoming angry during a phone call) **[R6]**.

Scalable tools. The complexity of ASR systems in both training and recognition is high. To alleviate this problem, the Sheffield team developed a resource optimisation toolkit (ROTK), which brings a novel graph/metadata approach to system construction, allowing optimisation of system structures and their deployment. The group has also developed a scalable and fully automatic framework for integrated training of acoustic and language models on a very large scale (>10,000 hours of acoustic data).

3. References to the research (indicative maximum of six references)

Sheffield staff and students in **bold**.

R1. **Hain, T.**, Burget, L., Dines, J., Garner, P. N., Grezl, F., **El Hannani, A.**, Huijbregts, M., Karafiat, M., Lincoln, M., & **Wan, V.** (2012). Transcribing Meetings With the AMIDA Systems. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(2), 486–498. https://doi.org/10.1109/tasl.2011.2163395. Cited by 124.

R2. **Saz, O., Deena, S., Doulaty, M., Hasan, M., Khaliq, B., Milner, R., Ng, R. W. M.**, Olcoz, J., & **Hain, T.** (2018). Lightly supervised alignment of subtitles on multi-genre broadcasts. *Multimedia Tools and Applications*, *77*(23), 30533–30550. https://doi.org/10.1007/s11042-018-6050-1. Cited by 2.

R3. **Saz, O.**, & **Hain, T.** (2017). Acoustic adaptation to dynamic background conditions with asynchronous transformations. *Computer Speech & Language*, *41*, 180–194. https://doi.org/10.1016/j.csl.2016.06.008. Cited by 1.

R4. **Deena, S., Hasan, M., Doulaty, M., Saz, O.**, & **Hain, T.** (2019). Recurrent Neural Network Language Model Adaptation for Multi-Genre Broadcast Speech Recognition and Alignment. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *27*(3), 572–582. https://doi.org/10.1109/taslp.2018.2888814. Cited by 10.

R5. **Shi, Y., Huang, Q.,** & **Hain, T.** (2020). H-Vectors: Utterance-Level Speaker Embedding Using a Hierarchical Attention Model. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Barcelona, Spain*, 7579-7583. https://doi.org/10.1109/icassp40776.2020.9054448. Cited by 8.

R6. **Jalal, M. A., Milner, R., Hain, T.,** & **Moore, R. K.** (2020). Removing Bias with Residual Mixture of Multi-View Attention for Speech Emotion Recognition. *Proceedings of the 21st Annual Conference of the International Speech Communication Association (Interspeech), Shanghai, China*, 4084-4088. https://doi.org/10.21437/interspeech.2020-3005. Cited by 0.

4. Details of the impact (indicative maximum 750 words)

In May 2018, Sheffield launched the Centre for Speech & Language Technology, a joint research centre with US voice data analysis company VoiceBase (with Hain as Director). VoiceBase provides speech recognition and speech analytics services through web-based software to companies whose business involves large call centres. Sheffield technology underpins the ability of VoiceBase to transcribe 1 billion minutes of calls every year, processing up to 60,000 live audio streams simultaneously **[S1]**.

When VoiceBase was launched in 2010, it relied on temporary solutions and licensed third party software to support its services. However, with rapid progress in artificial intelligence (AI) and speech recognition technology, the company needed to develop a world-class speech engine of its own, supported by an internationally recognised research base, in order to compete in a rapidly growing market. VoiceBase's collaboration with Sheffield has impacted the company in three key areas:

Company-wide infrastructure for safe speech system updates. The collaboration between Sheffield and VoiceBase began with Sheffield's creation of a framework for the development of new speech models, which was immediately implemented across the whole of the company (60-70 employees across the US, Russia, India and the UK). Drawing on his expertise in this field and the ROTK developed at Sheffield, Hain developed a system that allowed staff in any location to incorporate novel research and technology from a multitude of different platforms and apply them to diverse aspects of the company's speech engine consistently and safely, without jeopardising the system's operation **[S1]**.

Upgraded speech engine. A speech engine comprises some 30 different strands of fast-moving technology, each one of which is the subject of multiple research projects worldwide. Keeping pace with the leading edge across all of these developments and identifying the key breakthroughs and advances requires an exceptional understanding of the field. Hain's research makes him one of only a small handful of people to have such understanding, allowing him to direct the development of VoiceBase's speech engine, deciding which elements to replace, and identifying the most advanced solution to use **[R1-R6] [S1]**.

New speech models. The Sheffield VoiceBase Centre for Speech & Language Technology has developed a set of new (probabilistic) speech models that were rolled out to all VoiceBase clients in 2020. These models are based on hundreds of thousands of hours of archived recordings and their transcriptions, together with processes that optimise how audio source data

are cleaned, processed, aligned, and separated to improve the accuracy of future transcriptions. Sheffield's expertise in deep learning architectures, emotion detection, and sentiment analysis were key to the development of these new models **[R4-R6] [S1]**.

The technology developed allows VoiceBase to create highly accurate transcriptions of all their customers' calls, including non-verbal elements (such as pauses, pace, and voice pitch of speakers) and emotion (such as empathy, anger, and excitement). AI is then applied to allow tailored, bespoke analysis of these transcripts to generate the performance and predictive data required by clients for a range of purposes, including the following:

- **Fraud prevention**: keywords, pauses, and pitch/tone analysis allow clients to monitor all of their call centre operatives automatically for fraudulent or illegal activity, and take swift action to avoid millions of dollars in fines from the US Federal Communications Commission and its equivalents **[S2]**.
- **Performance monitoring and management**: enabling the ability to identify when staff talk over callers, respond empathically to anger or distress, and provide the information requested **[R6]**.
- **Maximising sales and allocating resources**: giving a reliable percentage score in the opening minute of a call to indicate the likelihood that it will lead to a sale or appointment, allowing the company to decide whether and how to follow up.
- **Protection of sensitive information:** automatically identifying and deleting sensitive information (e.g. credit card details) from call recordings and transcripts.
- **Market analysis**: automatically identifying common demands and complaints to improve offerings.
- **Meeting analysis**: minutes, action points, keywords, identification of participants, individuals' input, concerns, and sentiment, etc. **[R1]**.

Since the collaboration began, the company has recorded strong growth (100% in 2018 and 100% in 2019, with the same expected in 2020) and is expanding into Europe and Australia. Voicebase's CEO attributes a large part of this success to Sheffield research. "*Everything we do as a company relies on Thomas* [Hain]. *Since the end of Q2 2020, 95% of our speech recognition technology has been based on Sheffield research. Based on acquisitions by Apple, Google, Facebook over the last ten years, the new engine and research capability provided by our collaboration with Sheffield is worth at least USD100M+ if looking at company valuations and merger and acquisition historical comparables*." **[S1]**.

VoiceBase attributes their ability to compete with the likes of Microsoft, Google, Amazon, and IBM to attract multi-billion-dollar companies such as HomeDepot, NASDAQ, Uber, Dish Network, GrubHub, Delta Dental, Centurylink, and Twilio as clients to the increased transcription accuracy enabled by Sheffield and the global reputation of the research team **[S1]**. The company is also partnering with call centre hardware providers such as Poly (70% share of the global market, 90% of the US market) and GN Netcom/Jabra to launch products directly connected to VoiceBase Speech analytics **[S1]**. The Senior VP for Strategic Alliances and Partnerships of Poly said, "*I was blown away by the analytic capabilities of your* [VoiceBase] *platform and how you're disrupting the economics of what it means to do sentiment analysis*." **[S3]**.

The COO of Proactive Dealer Solutions (PDS), which assists car dealerships by maximising call centre sales, gives an example of the impact of Sheffield technology on VoiceBase clients: "*Within a matter of minutes using the predictive indicators that we built with VoiceBase […] we are able to score the call on 30 different metrics […] The average dealership loses 20% of their inbound calls and only appoints about 18%. Thanks to the tool we built, our clients across the board are seeing a fail rate of less than 10% and we are now appointing at 55%. The average dealer that works with us and really embraces our management tool sees an increase in sales of over USD100,000 a month.*" **[S4]**.

5. Sources to corroborate the impact (indicative maximum of 10 references)

S1. Confidential testimonial from the CEO of VoiceBase (2020). Corroborates a) the role of Sheffield's research in Voicebase developments, b) company growth and value, and c) a list of key customers.

S2. Federal Communications Commission press release. Provides an example of the size of financial penalties for call centre fraud and illegal activities. (Accessed 16th June 2020). https://docs.fcc.gov/public/attachments/DOC-332911A1.pdf

S3. [Text removed for publication]

S4. YouTube video of the COO of PDS. Corroborates a) the value of VoiceBase technology to PDS, b) how the technology is used, and c) improvements in the business performance of dealerships using this technology. (Accessed 16th June 2020). https://www.youtube.com/watch?v=4Veq-ge_6Vc&t=52s