Institution: University of Edinburgh

Unit of Assessment: B11 — Computer Science and Informatics

Title of case study: Clinical and commercial applications of text-to-speech synthesis technologies

1. Summary of the impact

Edinburgh's research in multilingual speech synthesis has had clinical and commercial impact, and has resulted in a large and diverse community of users.

Clinical applications: Our research has enabled the construction of natural-sounding, personalised synthetic voices from recordings of speech from people with disordered speech due to conditions such as Parkinson's disease or Motor Neurone Disease. These synthetic voices are used in assistive technology devices that allow sufferers of these conditions to communicate more easily and effectively.

Commercial take-up: Our research has achieved commercial impact through the licensing of technology components, and through the activities of start-up companies.

Community of users: The Festival Speech Synthesis System (v2.1 released in November 2010) is a complete open-source text-to-speech system released under an unrestrictive X11-type license, and is distributed as part of many major Linux distributions.

2. Underpinning research

Key researchers at the University of Edinburgh developing the underpinning research:

Robert Clark (research fellow 2002–2004; lecturer 2004–2009; research fellow 2009–date)
Simon King (lecturer 2000–2007; reader 2007–2010; professor 2010–date)
Steve Renals (professor 2003–date)
Korin Richmond (research fellow 2002–date)
Paul Taylor (lecturer 1997–2001)
Junichi Yamagishi (research fellow 2007–2011; EPSRC Fellow 2011–date; Lecturer 2013–date)

The impact of Edinburgh speech synthesis arises from a number of research findings in *text-to-speech synthesis*, the automatic conversion of written language into speech. There are two main approaches to the problem: concatenative (unit-selection) speech synthesis (a), and statistical parametric (HMM) speech synthesis (b). We have made research advances in both (a) and (b) that have resulted in research impact. Within the statistical framework we have made key developments that have enabled the construction of personalised voices (c) from small amounts of data and taking different accents into account (d). A significant application of Edinburgh research findings has been the development of personalised synthetic voices for people with disordered speech due to neurological disorders (e).

a) ***Concatenative text-to-speech synthesis*** [1,2] 1994-2010.
The basic techniques for concatenative speech synthesis, and their software embodiment in the Festival system, were developed during 1994-2001. This resulted in a number of technical advances in all areas of speech synthesis (including letter-to-sound mappings, intonation modelling, and unit selection algorithms), and a more general formal framework for speech synthesis [1], which formed the basic structure of Festival and later commercial systems (e.g. Phonetic Arts, §4). This work was supported by six EPSRC responsive mode grants (*GR/K54229/01*; *GR/L53250/01*; *GR/L50341/01*; *GR/R94688/01*; *EP/D058139/1*; *EP/E031447/1*) during 1997-2010 (total value: £1,382k).

b) ***Adaptive statistical parametric speech synthesis*** [3] 2006-2013.
Within this framework, we have developed new algorithms to adapt an "average voice" synthesis system (trained using speech from multiple speakers) to the voice of a new speaker

using much less speech from the target speaker compared with the previous concatenative systems [3]. This approach allows more control over the synthesised speech, enabling automatic adaptation to new speaking styles and emotions. This work was supported by EPSRC responsive mode grant *EP/E027741/1* (2006-2009; £287k), EPSRC programme grant *EP/I031022/1* (2011-2016; £6,236k), and EPSRC Career Acceleration Fellowship *EP/J002526/1* (2011-2016; £741k).

c) ***Personalised speech synthesis*** [4] 2008-2013.
Using the adaptive framework we have developed systems which can automatically create a personalised synthetic voice for a target speaker using just a few minutes of data ("voice cloning"). We demonstrated this approach by creating thousands of personalised synthetic voices [4], and have also shown how some of the techniques can be applied with unit selection systems. In addition the techniques developed in [3] may be applied to lower quality recordings (e.g. web videos) than was previously feasible for speech synthesis development. This work was supported by EPSRC programme grant *EP/I031022/1* (2011-2016; £6,236k), EPSRC Career Acceleration Fellowship EP/J002526/1 (2011-2016; £741k), and two EU FP7 grants coordinated by Edinburgh (*EMIME*, *Simple4All*; €6,000k).

d) ***Accent-specific pronunciation lexicon*** [5] 2005-2010.
Synthesising speech across different accent groups is a key aspect of a general approach to personalised synthesis. This requires automatic adaptation of the pronunciation lexicon [5] as well the acoustic components of the system (b, c). *Combilex* is a high-quality pronunciation lexicon for speech technology applications, developed from scratch since 2005 in the Centre for Speech Technology Research, University of Edinburgh. It is based on an accent-independent top-level lexicon, from which accent-dependent surface lexica may be automatically generated. This work was supported by a Proof-of-Concept grant from Scottish Enterprise (2005–2007; £128k).

e) ***Voice reconstruction*** [6] 2010-2013.
From 2010-2013 we have further developed personalised speech synthesis to enable voice reconstruction in a clinical setting in which the target speakers have disordered speech due to a neurological condition such as motor neurone disease. The resulting synthetic speech repairs the disordered aspects, resulting in normal-sounding, intelligible, personalised speech. The key modelling and algorithmic advances were made at Edinburgh, with initial trials carried out in collaboration with the University of Sheffield. This work was supported by EPSRC programme grant *EP/I031022/1* (2011–2016; £6,236k), EPSRC Career Acceleration Fellowship EP/J002526/1 (2011–2016; £741k), and by the Euan MacDonald Centre for Motor Neurone Disease (MND) Research (2010–date.)

3. References to the research

1. P Taylor, AW Black, and R Caley (2001). "Heterogeneous relation graphs as a formalism for representing linguistic information", *Speech Communication*, **33**, 153-174. *http://dx.doi.org/10.1016/S0167-6393(00)00074-1*
2. RAJ Clark, K Richmond, and S King (2007). "Multisyn: Open-domain unit selection for the Festival speech synthesis system", *Speech Communication,* **49**, 317-330. *http://dx.doi.org/10.1016/j.specom.2007.01.014*
3. J Yamagishi, T Nose, H Zen, Z Ling, T Toda, K Tokuda, S King, and S Renals (2009). "Robust speaker-adaptive HMM-based text-to-speech synthesis", *IEEE Transactions on Audio, Speech, and Language Processing,* **17**, 1208-1230. *http://dx.doi.org/10.1109/TASL.2009.2016394*
4. J Yamagishi, B Usabaev, S King, O Watts, J Dines, J Tian, R Hu, Y Guan, K Oura, K Tokuda, R Karhila, and M Kurimo (2010). "Thousands of voices for HMM-based speech synthesis – analysis and application of TTS systems built on various ASR corpora", *IEEE Transactions on Audio, Speech, and Language Processing*, **18**, 1005-1016. *http://dx.doi.org/10.1109/TASL.2010.2045237*
5. K Richmond, R Clark, and S Fitt (2009). "Robust LTS rules with the Combilex speech

technology lexicon", *Proc Interspeech,* 1295-1298. *http://www.era.lib.ed.ac.uk/handle/1842/3958*

6. S Creer, S Cunningham, P Green, and J Yamagishi (2013). "Building personalised synthetic voices for individuals with severe speech impairment", *Computer Speech and Language*, **27**, 1178-1193. *http://dx.doi.org/10.1016/j.csl.2012.10.001*

References [1], [2], [3], [4] and [6] are papers in the three most important journals in the speech processing research field. Reference [5] is a paper in the leading international speech processing conference. References [2], [3], and [4] are most indicative of the quality of the research.

4. Details of the impact

4.1. Clinical applications

Neurological diseases such as MND or Parkinson's Disease can result in deterioration in speech production due to a loss of coordination and control of the speech articulators. It is currently estimated that 170 people per 100,000 are affected by dysarthria (speech motor disorder); about 5,000 people in the UK have MND, with 2 people per 100,000 newly diagnosed each year. People with such speech disorders lose not only a means of communication, but also vocal expression of individual and social identity. A number of *Augmentative and Alternative Communication* (AAC) devices are now available to enable people with such conditions to communicate by speech, for example using eye-tracking interfaces. However these devices come with a very limited range of synthetic voices: sometimes users do not even have a choice of male or female voice, let alone a voice with their accent and speech characteristics.

In conjunction with MNDA Scotland, we have developed a "voice banking" service containing recordings of several hundred speakers from across Scotland. The main aim of this is to enable accent-specific average voices to be constructed which can then be better adapted to the target speaker, but it also means that donors will have an 'insurance policy' should they ever require a personalised synthetic voice. Voices banked include the First Minister of Scotland, and many other MSPs (corroboration: [C], [D]).

Our research in personalised speech synthesis and voice reconstruction has resulted in a collaboration with the Euan MacDonald Centre for MND Research at Edinburgh. The Euan MacDonald Centre was established in Edinburgh, in 2007, by the generosity of MND patient Euan MacDonald and his father, Donald. Initially we carried out a pilot study with Euan MacDonald for whom just three minutes of (disordered) speech was available. We were able to reconstruct a personalised synthetic voice, which is installed on his eye-tracking based AAC device and is in daily use. Euan MacDonald campaigns on behalf of the disabled [B] writing that "I feel that a person's voice is one of the most personal things that they possess and the Voicebank project is another project that I feel passionately about.'' [C].

Since then, in an extended trial, we have successfully provided ten patients with a reconstructed voice that they use via an internet-connected device (e.g. iPad). Current trials involve a prototype user interface in which everything runs locally. This work has had considerable media coverage, for example a special feature in the prime time (9% audience share) Japanese programme "Close-Up Gendai" (corroboration: [F]). The work is being extended into a clinical trial phase supported by an MRC Confidence in Concept award, and funding from the charity MNDA.

The recently opened Anne Rowling Clinic (founded by donations from J.K. Rowling) will have a recording facility specialized for voice banking purposes, and incorporated into the design as a direct result of our voice banking and reconstruction research (corroboration: [E]).

4.2. Commercial take-up

Since 2008, the Combilex multi-accent lexicon has been commercially licensed to ten companies and organisations in seven countries. These are MModal in the USA; IVO Software in Poland; Phonetic Arts in the UK; Toshiba Research Europe in the UK; Illumina Digital in the UK; the University of Alberta in Canada; NICT in Japan; Amazon in the USA; Google in Ireland; Samsung

Beijing R&D in China. In addition to these a further five evaluation licenses have been acquired, resulting in revenue of £31,000.

Following an initial exploratory consultancy contract with the University (2011), Orange / France Telecom (UK) Ltd initiated a Knowledge Transfer Partnership (2012–13) whose aim is to improve automatic voice building through development/integration of novel automatic speech recognition techniques and build commercial-grade systems for bringing personalised speech technology to Orange customers. Building on this Orange / France Telecom recently funded custom development of Swahili accent English TTS voices and Kiswahili (Swahili Language) TTS for trials with customers in Kenya.

SMEs have also been developed based on the research produced at Edinburgh. Paul Taylor founded Phonetic Arts in 2007; building on the concatenative synthesis structures he developed while a lecturer in Edinburgh (1997–2001). The 15-person company specialised in the development of high-quality speech synthesis for computer game applications and were acquired by Google in December 2010 for an undisclosed amount (corroboration: [A], [H]).

4.3. Community of users

We have developed a broad and diverse community of users through the release of software toolkits and synthetic voice libraries.

We are the coordinating site for the open-source speech synthesis toolkit, Festival and the associated Edinburgh Speech Tools package. Festival is distributed as default in a number of standard Linux distributions including Arch Linux, Fedora, CentOS, RHEL, Scientific Linux, Debian, Ubuntu, openSUSE, Mandriva, Mageia and Slackware, and can easily be installed on any Linux distribution that supports apt-get. More recently our work on statistical parametric speech synthesis and the algorithms for adaptation have been incorporated in the HTS toolkit (one of the coordinators (Yamagishi) is from Edinburgh), which integrates with Festival. These toolkits are the most used open-source speech synthesis systems (Corroboration: [G]). These open-source toolkits have also formed the high performing baseline systems for the international Blizzard evaluation of (commercial and research) speech synthesis also organised by Edinburgh.

Although our core speech synthesis software is open source, we licence a specifically-produced high quality synthetic voice library separately (e.g., *http://licensing.research-innovation.ed.ac.uk/2948*), free for non-commercial research usage. Between December 2010 and July 2013 it was licensed to 162 researchers from a variety of organisations in 25 countries.

5. Sources to corroborate the impact

A. Research Manager, Speech Synthesis, Google; founder of Phonetic Arts — can corroborate that technology developed by Phonetic Arts builds on research done in Edinburgh.
B. "Euan just wants to go places", Blackwood Foundation Bespoken forum, *http://www.bespoken.me/forum/topics/euan-just-wants-to-go-places*, July 2013.
C. "The Voice Bank Project", Blackwood Foundation Bespoken forum, *http://www.bespoken.me/forum/topics/the-voicebank-project*, July 2013.
D. *http://www.smart-mnd.org/voicebank/* and *http://www.euanmacdonaldcentre.com/voicebank.html*
E. *http://annerowlingclinic.com/research.html*
F. NHK (Japan Broadcasting Corporation), Close-up Gendai, 28 Feb 2012, Medical applications of speech synthesis technologies. *http://www.nhk.or.jp/gendai/kiroku/detail02_3166_all.html*
G. H Zen and K Tokuda (2009). "[Best of the Web] TechWare: HMM-based speech synthesis resources", *IEEE Signal Processing Magazine,* **26(4)**, 95-97. *http://dx.doi.org/10.1109/MSP.2009.932563*
H. *http://www.computescotland.com/google-absorbs-phonetic-arts-3868.php*

Archive copies of web page sources are available at *http://ref2014.inf.ed.ac.uk/impact*