Institution: The University of Edinburgh

Unit of Assessment: 28B Linguistics

Title of case study: Commercial and clinical impact of speech synthesis

1. Summary of the impact

Our research on speech synthesis is embodied in software tools which we make freely available. This has led to widespread use and commercial success, including direct spinouts, follow-on companies and use by major corporations. This same research benefits people who lose the ability to speak and have to rely on computer-based communication aids. Unlike existing aids, which provide a small range of inappropriate voices which are often not accepted by users, our technology can uniquely create intelligible and normal-sounding personalised voices from recordings even of disordered speech, and so enable people to communicate and retain personal identity and dignity.

2. Underpinning research

Text-to-speech (TTS) is the automatic conversion of written language into speech. This involves *text analysis*, followed by *waveform generation*. The main approaches to the second stage are playback of recorded speech sounds (*concatenative synthesis*), or a statistical model (*H[idden] M[arkov] M[odel]-based synthesis*).

The Centre for Speech Technology Research (CSTR), a joint research centre of LEL and Informatics, has been pioneering TTS for well over 20 years, but the key components underpinning the impact described here were developed from 1996 onwards. Our general, well-established framework performs text analysis and concatenative waveform generation and is embodied as the **Festival software toolkit** (http://www.cstr.ed.ac.uk/projects/festival, archived at http://tinyurl.com/phm5tpt), alongside a separate line of research on HMM-based waveform generation started in Japan, and now involving CSTR, embodied in the **HTS software toolkit** (http://hts.sp.nitech.ac.jp, archived at http://tinyurl.com/o9tensp).

The **Festival toolkit** offers a complete framework for TTS and incorporates the research of numerous members of CSTR. The first version was created in 1996 by Black (researcher: 1996-99) and Taylor (researcher; lecturer: 1993-2001) later joined by Caley (developer: 1997-2001). Festival embodies research results from CSTR produced from 1996 onwards by Black, Taylor, Isard (director: left 1999), Clark (PhD student; researcher; lecturer: arrived 1996), King (PhD student; lecturer; reader; professor; director: arrived 1993), Richmond (PhD student; researcher: arrived 1997), and Yamagishi (researcher; lecturer: arrived 2006). These results include significant advances in letter-to-sound prediction, intonation, signal processing, unit selection, etc. (e.g., Taylor, Black & Caley, 1998; Clark et al, 2007), and were achieved with funding from EPSRC, commercial sources and the EC.

The **HTS toolkit** is for HMM-based systems, used in conjunction with *Festival*. HTS is co-maintained by CSTR, with annual software releases embodying novel research conducted in CSTR and a few other groups. The key developments underpinning the claimed impact were made at CSTR, with funding from EPSRC and the EC, from 2006 onwards (Yamagishi et al, 2009; Yamagishi et al, 2010; Watts et al, 2010).

These developments concern the ability to adapt the statistical model to new speakers using just a few minutes of speech. The technique can be used to create speaking styles, emotions, etc. (e.g., Watts et al, 2010), opening up novel applications. The major advantage of our techniques over existing methods is the possibility of using lower-quality recordings (e.g., home video), less data (minutes, not hours), and speech that is disordered (e.g., due to Motor Neurone Disease) while still creating normal-sounding, intelligible, personalised synthetic speech. The key breakthrough enabling the use of disordered speech was made at CSTR in the period 2008-2012. Initial tests were made in collaboration with the University of Sheffield (Creer et al, 2009). CSTR deployed the first “voice reconstruction” system in a clinical setting in Edinburgh in 2010.

This recent clinical application of the technology is underpinned by the sustained earlier period (1996-present) of work on general-purpose TTS, which has itself made a substantial impact on a wider range of applications including telephone services, computer games, and facial animation, via take-up by industry and our spinout companies, as described in section 4.

3. References to the research

Clark, R. A. J., K. Richmond, and S. King (2007). Multisyn: Open-domain unit selection for the Festival speech synthesis system. *Speech Communication*, 49(4):317-330. DOI: 10.1016/j.specom.2007.01.014

Creer, S., P. Green, S. Cunningham, and J. Yamagishi (2009). Building personalised synthesised voices for individuals with dysarthria using the HTS toolkit. In John W. Mullennix and Steven E. Stern, editors, *Computer Synthesized Speech Technologies: Tools for Aiding Impairment*. IGI Global. ISBN 978-1-61520-725-1 (pdf of chapter available from University of Edinburgh)

Taylor, P., A. Black, and R. Caley (1998). The architecture of the Festival speech synthesis system. *Proc. Third ESCA/COCOSDA Workshop on Speech Synthesis*, Australia. Handle: http://hdl.handle.net/1842/1032

Yamagishi, J., 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(6):1208–1230, August. DOI: 10.1109/TASL.2009.2016394

Yamagishi J., 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(5):984-1004, July. DOI: 10.1109/TASL.2010.2045237 Output returned in the REF

Watts, O., J. Yamagishi, S. King, and K. Berkling (2010) Synthesis of child speech with HMM adaptation and voice conversion. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(5):1005-1016. DOI: 10.1109/TASL.2009.2035029 Output returned in the REF

Grant funding for the underpinning research is extensive, including a number of EPSRC research grants from the late 1990s onwards (e.g. COUGAR (PI) 2002-05, £181K; TESSa (PI) 2005-08: £257K; ePHONES (PI) 2006-09: £238K; ProbTTS (Co-I) 2006-09: £359K; Attaca (PI) 2007-10: £351K, NST (Co-I) 2011-16: £7,6M), donations to CSTR from Sun Microsystems (£51K), EC-funded projects involving Clark in the early 2000s that integrated speech synthesis into applications (e.g. M-PIRO, 2000-03: £268K) and more recent EC FP7 projects such as EMIME (2008-11: £606K) and Simple^{4}All (2011-14: £1.05M) co-ordinated by King. Further industry support has come from France Telecom R&D UK Ltd, supplemented by license income [5.LIC].

4. Details of the impact

Impact was achieved first commercially and then in a clinical application. Although there have been minor contributions from collaborators regarding the clinical work, all impact claimed here is a direct result of research conducted only at the University of Edinburgh [5.GRE]. This starts from the basic components required to convert text into speech, implemented as the Festival toolkit. This is combined with the speaker adaptation and noise robustness of Yamagishi's work embodied in the HTS toolkit. Together, Festival and HTS have had substantial impact on the speech technology industry [5.CON; 5.TAY]. Adding to this our unique ability to *repair* disordered speech has led to an assistive technology application which enhances the quality of life of people with speech disorders [5.DON]. Our techniques work automatically from data: they can be deployed widely and cost-effectively. A small-scale clinical service has already benefitted patients. There is a funded roadmap to a full service for patients (funding: MNDA & MRC).

Commercial R&D: Our tools are widely used as a research & development framework by industry [5.TAY; 5.CON; 5.FIN]. This represents a major form of impact for every individual research contribution embodied in them and a reach extending to most of the large industry players. The release of working implementations has higher impact than the published papers, since re-implementing complex techniques is time-consuming and expensive [5.TEC]. Evidence for the reach and significance of this impact for both corporations and other companies during the period 2008–2013 can be found at any workshop or conference, where typically around half of all papers presented are based on research performed using the Festival and HTS toolkits with an estimated one quarter of all papers coming from industry groups. A typical example is *Proc Interspeech*

2010, where the majority (9/14) of papers on speech synthesis authored by researchers in industrial labs used either Festival or HTS to conduct their experiments. The documentation for the various releases of Festival (online manuals and papers describing the architecture) have been cited over 400 times between 1 Jan 2008—31 July 2013 (source: Google Scholar), again with an estimated one quarter of these being from industry. According to a senior researcher at AT&T "almost every industrial researcher in the field has used or is familiar with both Festival and HTS" [5.CON].

Commercial products: Festival is released as Open Source under an unrestrictive license. It has formed the basis of products and led to company formations [5.SPN]. A direct spinout from CSTR, Rhetorical Systems, led to follow-on companies (Phonetic Arts, CereProc)—see below for more detail—and to continued use of Festival and HTS by major corporations including AT&T [5.CON], Nuance, Google [5.TAY] and Microsoft. We also license specific technologies on a commercial basis to a wider group of companies. Our Combilex dictionary system has been licensed to companies in the UK, Eire, Switzerland, Poland, USA, China (£31K to date); our voice databases and the tools developed for our clinical application, which are also available for non-clinical use on healthy voices, have been licensed to companies in the UK, Poland, USA (£9.5K to date) [5.LIC].

Google's current speech synthesis group and the speech synthesis company CereProc both have their roots in Festival. Taylor (CSTR 1993–2001) founded Rhetorical Systems in 2000, which used Festival as the basis for its commercial product rVoice. Rhetorical Systems was then acquired by Nuance in 2004 for £3.6 million. Taylor then founded follow-on company Phonetic Arts in 2006; in 2010 it had a turnover of £154K from products including a unit-selection text-to-speech system closely following the approach in Festival. Google's current speech synthesis group was formed by acquiring Phonetic Arts in 2010 for an undisclosed sum. As their Technical Lead of TTS has stated, the impact of TTS at Google is "huge," with millions of unique users of their TTS systems every day; Festival has been "highly influential" for their system, and their speech synthesizer "has its roots in HTS" [5.TAY]. Aylett (CSTR 1999–2000, 2006–09, 2012–ongoing) was also involved in Rhetorical Systems (2004–05), and founded CereProc in 2005. This company is still trading and has a unit-selection product which closely follows the Festival approach, and a statistical parametric product based on the HTS code.

AT&T continues to develop their own commercial product based on the Festival architecture [5.CON].

CSTR's recent speech animation spinout Speech Graphics, formed in 2010 and still trading, is based on research conducted in CSTR, including the speech synthesis research outlined in section 2. Its customers include Supermassive Games and Havok; in 2011 it was awarded a prestigious John Logie Baird Innovation Award for Knowledge Transfer, in 2012 it was a finalist in the TIGA (trade body of the UK Games Industry) Awards.

Benchmarking and evaluation: Festival and HTS are both important reference implementations for industry well beyond our own spinouts [5.TAY]. Our systems have become the benchmarks by which other systems are judged, mainly because of the high quality speech they generate but also because they are publicly available and provide reproducible results. Every year since 2005, they have been used as benchmarks in the Blizzard Challenge, a competitive evaluation, organised by King, of systems from companies including Microsoft, IBM, iFLYTEK, IVONA, Voiceware, Nokia alongside those from leading research groups worldwide [5.BLZ]. This is the only place where direct comparisons can be made between commercial systems. The Challenge is funded by industry subscriptions, cash awards from Google [5.FIN; 5.TAY] and contributions in kind from Phonetic Arts, Toshiba, Lessac Technologies, ATR, IVONA/Amazon and Loquendo [5.BLZ]. Evidence of the importance of Blizzard to the industry is demonstrated by the high levels of participation in the Challenge itself and the attendance at the workshop of senior industry figures. "The most prestigious event in the calendar facilitating an exchange of ideas among those conducting research into speech synthesis […] a unique opportunity to compare different state-of-the-art TTS technologies with a view to discovering innovative solutions aimed at improving the quality and accuracy of text to speech" (Paul Coppo of Loquendo [5.BLZ]).

Clinical: The first pilot study with Motor Neurone Disease sufferer Euan MacDonald using a 3-minute sample of his voice was conducted in 2010; the resulting voice is now installed on his eye-

tracking-based communication device and is in daily use [5.DON]. The next phase involved a more extensive trial using a voice banking service (one hour of speech from each of 600 people, including the Scottish First Minister and many MSPs) to gather the data needed to train the underlying statistical model, and treating more patients. Everyone who contributes their voice for use in reconstructing patients' voices, also has an insurance policy in the event their own voice becomes disordered. We have successfully provided 10+ patients with a reconstructed voice that they can use on a smartphone or tablet [5.DON]. Further evidence of the impact includes funding awarded because of the success of initial trials, including donations made by the MacDonald family in 2010-12 [5.DON], an MRC Confidence in Concept award (awarded early 2013) and funding from the charity MNDA (awarded late 2012) sustaining this project into the clinical trial phase; and purpose-built recording facilities designed to our specification at the new Anne Rowling Regenerative Neurology Clinic (opened 2013), funded from a donation made by J. K. Rowling.

5. Sources to corroborate the impact

Individuals who can provide corroboration of claims made in this impact case study:

5.CON Impact of Festival and HTS for commercial R&D in the Speech Technology Industry: Factual statement from AT&T, available from the University of Edinburgh

5.DON Utility of clinical application to patients and verification of the sustainability of the clinical programme: Patient family

5.GRE Responsibility of CSTR for research underpinning clinical application: Personal Chair at the Department of Computer Science, University of Sheffield

5.TAY Impact of Festival and HTS for commercial R&D in the Speech Technology Industry: Factual statement from Google, available from the University of Edinburgh

Other sources of corroboration:

5.BLZ The Blizzard Challenge:
a. Participation details for every year of the Challenge, with linked papers by industry participants: http://www.synsig.org/index.php/Blizzard_Challenge, archived at http://tinyurl.com/pj8n8s9;
b. Example of in-kind industry support: http://www.gitex.com/press/Loquendo-hosts-The-Blizzard-Challenge-2011-Workshop, archived at http://tinyurl.com/nm5em4u

5.FIN Commercial financial support for Festival which assisted its release as Open Source and its availability to industry for R&D, and support for related activities such as the Blizzard Challenge have come from Sun Microsystems and Google; charitable support has been given by Donald MacDonald via the Euan MacDonald Centre for pilot work enabling the clinical impact. Full details of the finances available from the University of Edinburgh

5.LIC Companies who have bought licenses for Combilex, and for the voice databases and tools developed for clinical application: Full details available from the University of Edinburgh

5.SPN Spinout and follow-on companies:
a. Acquisition of Rhetorical Systems by Nuance http://www.nuance.com/news/pressreleases/2004/20041115_rhetorical.asp, archived at http://tinyurl.com/p7x68cu
b. Acquisition of Phonetic Arts by Google http://googleblog.blogspot.com/2010/12/can-we-talk-better-speech-technology.html, archived at http://tinyurl.com/p2b3vjt
c. CereProc: www.cereproc.com, archived at http://tinyurl.com/ntozeok
d. Speech Graphics: www.speech-graphics.com, archived at http://tinyurl.com/pf33dn6

5.TEC Independently-authored article evidencing some of our commercial, research and benchmarking impact: "TechWare: HMM-Based Speech Synthesis Resources," *IEEE Signal Processing Magazine* 26(4): 95–97, July 2009. DOI: 10.1109/MSP.2009.932563.