TALKING TECHNOLOGY

Dr Matthew Aylett
Royal Society Industrial Fellow, Informatics, University of Edinburgh

"Speech is the mirror of the soul; as a man speaks, so is he" – Publilius Syrus Roman author, 1st century BC

We know there is something special about speech. Our voices are not just a means of communicating, although they are superb at it, they also give a deep impression of who we are. They can betray our upbringing, our emotional state, our state of health. They can be used to persuade and convince, to calm and to excite. The power of speech is key to our parliamentary democracy. A debate requires the ability to speak, the rough and tumble of parliamentary life revolves around what politicians say rather than what they write. So, what can we do if a person loses the power of speech? What can we do if we want to grant the power of speech to our machines and tools? The answers lie with speech synthesis (also known as text to speech – TTS) technology.

Speech synthesis has progressed enormously since the trademark Stephen Hawking voice which was based on synthesis developed in the mid-eighties. Hawking has retained the same system, despite issues with naturalness, because it has become his personal voice. If he used a modern more natural system no one would recognise him. This issue of personalisation has become a major driving force behind modern speech synthesis technology. In the past a company would only offer a Male and Female British RP accent. Now companies offer many voices with many regional accents. CereProc, an Edinburgh company, even offers a Glaswegian accented system for Android. However, for people suffering from a speech disability, the voice they are really searching for is their own.

Roger Ebert, arguably America's most famous film critic, lost the ability to speak after a thyroid cancer operation. Although he used speech synthesis available on his Apple Mac to communicate, he was frustrated because the voice did not sound like him. CereProc stepped in to help him. Using hours of Roger's commentaries from DVDs, they were able to create a voice that mimics his original speaking style. These techniques were also used by CereProc to create a satirical 'Bush-o-matic' website which mimicked the speech of George W Bush, and, during the US presidential elections, a version of Barack Obama and Mitt Romney.

. . . something special about speech . . .

However many people do not have a large bank of clean recorded audio with which to build a synthetic version of their voice. Current speech synthesis research is exploring how to mimic a subject's voice with less audio, and for that audio to be less cleanly recorded. Currently 3-5 hours of speech is required to produce a voice that sounds almost completely natural, however with a new statistical modelling approach good quality voices can be produced with 40 minutes of audio, and voices that sound like a person, although with reduced quality, with even less data. This ability to clone voices with less data raises a host of ethical and legal issues. To what extent does your voice belong to you? Audio recorded by radio, TV and for audio books is typically owned by the company producing the audio, not by the speaker. The recent GOS foresight report on 'Future Identities' highlighted the blurring and complexity of identity caused by hyper-connectivity, the ability to seamlessly create synthetic copies of voices from limited data presents an even greater challenge to understanding, controlling and facilitating digital identity.

For Roger Ebert, having a synthetic voice **mimic** his speaking style was not enough:

"I now propose a test for computer voices – the Ebert test. If a computer voice can successfully tell a joke and do the timing and delivery as well as Henny Youngman, then that's the voice I want."
TED Talk – Roger Ebert: Remaking my voice

Offering a sensitive means of controlling the artificial voice as well as building a voice which allows emotional variation is still

. . . the voice they are really searching for is their own . . .

very much a research question. Commercial speech synthesis can offer some emotional variation but it is very much limited compared to virtuosity and flexibility of the human voice. Current systems are typically controlled either by eye-gaze or by typing which make fluid conversation almost impossible.

Despite the social and medical need for voice replacement, this has not been the driving force behind increased current commercial and academic interest in speech synthesis. Instead it has been the popularity of mobile devices and the advent of pervasive computing. Apple's SIRI has increased the profile of using synthetic speech synthesis to allow our tools and machines to communicate with us. A sub-plot of an episode 'The Big Bang Theory' explored the idea of one of the characters falling in love with SIRI. Indeed, a machine that speaks and interacts through speech can be a disconcerting experience. However the potential power for speech to be used to make devices easier to use, and to help us manage the ever increasing sea of digital data that surrounds us is huge. Speech interfaces can also offer a non-technical interface that can be used more readily by sectors of the community which have encountered barriers to using modern technology.

American companies have been quick to see the potential of this new speech synthesis technology. Nuance bought two European rivals in 2011, with both Google and Amazon following suit. Europe, with its high technology infrastructure and multiple languages has previously been a dominant player in Language technology. With stiff competition from Asia together with buying power from the US, this may be about to change. There is now only one independent speech synthesis company in the UK (CereProc) and only one other in the rest of Europe (Acapela).

. . . increased commercial interest in speech synthesis . . .

SIRI has demonstrated how powerful speech technology becomes when connected with language processing and search technology. The ability to offer users information when they are on the move *eyes-free* is only worthwhile if you have information to give them. Here natural language processing (NLP) systems are critical. Such systems can summarise, search out and organise information that is of personal interest, speech synthesis can then generate personalised podcasts in the same way political researchers summarise information for MPs and Ministers. Being able to deliver intelligently summarised information as audio can help build communities, educate, and allow busy professionals a means of keeping up to date with a rapidly changing world.

As we bring ever increasing artificial intelligence (AI) algortihms together with both speech technology and computer animation, we are able to produce virtual characters and virtual representations of ourselves. Such technology is already being used in computer games, virtual training systems and in social computing. A natural sounding, flexible synthetic voice is a key to these applications.

The ability to give a natural sounding voice to animated characters, virtual agents and robots using speech synthesis is a reality. The scope for delivering information using synthesis is immense. However, just as our own power of speech reflects our own humanity, so speech synthesis can add a touch of humanity to our machines and tools, and, in the end, this sensation of seeing ourselves in our machines is perhaps the most strange and fascinating aspect of current speech synthesis technology.

HOW DOES SPEECH SYNTHESIS WORK?

Most commercial speech synthesis systems have a neutral speaking style and are an example of *unit selection* or *concatenative* synthesis. In simple terms, the synthetic speech is made from taking lots of small pieces of speech, taken from recordings of a human voice, and sticking them together in order to create the required series of sounds, intonation and voice quality for a new message. Such synthesis systems have four main components, a large database of recordings in the order of 3-5 hours of speech, a set of features that describe a new phrase or sentence, a search algorithm that finds the best pieces of speech in the database to match these features, and a method to smoothly glue these pieces together to produce the new phrase.

The advantage of using this approach is that the normal voice quality of the speaker is retained, and with enough material, the joins are not noticeable. However the system can only produce speech in the same style it was recorded in and, if some sounds are missing, they cannot be reproduced.

An alternative approach using a statistical model to abstract the sounds in a speech database with reference to the context the sound appears in. This model is then used to recreate completely a speech waveform using digital signal processing techniques. One advantage is that because no single unit is used, an error in the data will be absorbed into the model and its impact reduced. Another is that if a sound or transition doesn't exist in the data, it can be extrapolated from another speaker's data. This has three main effects, less data can be used to produce an acceptable quality voice, the synthetic voice is very stable and intelligible, however the voice quality does not sound as natural.

Current research is also interested in using a hybrid version of these systems in order to try and gain the advantages of both.