Voice Banking

Sarah Creer
Sarah Creer is a PhD student in the Clinical Applications of Speech Technology (CAST) group at the University of Sheffield, UK. She is working on synthesising personalised voices for individuals with progressive speech loss.

Professor Phil Green
has a Personal Chair at the department of Computer Science, University of Sheffield, UK.

Dr Stuart Cunningham
is a lecturer in the department of Human Communication Sciences, University of Sheffield, UK.

Correspondence to:
Sarah Creer,
Department of Computer Science
University of Sheffield
Regent Court
211 Portobello
Sheffield, S1 4DP.
Email: s.creer@dcs.shef.ac.uk

Acknowledgments:
Sarah Creer's PhD work is funded by EPSRC. The HTS software was provided by Junichi Yamagishi, CSTR, University of Edinburgh.

Degenerative diseases such as Parkinson's (PD) or motor neurone disease (MND) can lead to the partial or complete loss of speech. The time from diagnosis to a deterioration of the individual's control of their musculature varies but can be short and, as one of the first symptoms to present, speech deterioration frequently coincides, with or quickly follows, diagnosis. At diagnosis, thoughts about the future without speech or with speech that is no longer recognisable as your own can be difficult to focus on. This is a time, however, to take steps to make recordings of the voice (ahead of the deterioration) for later use in a prosthesis. This `banking' process makes it possible to produce a speech synthesiser that does not sound like someone with a different geographical, social, ethnic background or someone of a different age or gender, but can provide a voice with which the individual can identify and be recognised. This article examines the options available for voice banking; the current research in this area and also will provide some guidelines on making your own recordings.

The most basic type of speech synthesis is to store and playback pre-recorded phrases. Most communication aids can store digitised phrases for easy and quick access. This approach is limited in what can be said and how much can be stored.

A slightly different approach can be taken in banking a voice to create a fully synthesised output that can produce any utterance. High quality voices are currently available on communication aids such as the Toby Churchill Lightwriter (**http://www.toby-churchill.com/**). The quality of Acapela (**http://www.acapela-group.com/**) concatenative synthesis voices, for example, is due to a very large database of recordings produced by a professional speaker with high quality recording equipment, good recording conditions and consistent voice quality. This technique requires a dataset to be recorded that contains all the sounds in the target language so that it can produce any required output. The database is segmented into smaller units which can then be recombined to produce new utterances. Recording these sounds in context captures the variation in production depending on the surrounding speech making the speech more natural sounding and intelligible. Producing this quality of output is unlikely to be achieved by a non-professional speaker but as a trade-off of quality against a degree of personalisation, options are available for people wanting to bank their own voices using a smaller amount of data and in their own homes.

Festvox, (http://festvox.org/), from Carnegie Mellon University, Pittsburgh, is a voice building tool for researchers and as such, is not designed to be easy for someone without specialist phonetic and computational knowledge to use. A reasonable synthesised voice can be produced with around 1200 sentences or 1hr 20 mins of continuous speech. With minimal data input, however, it can be inconsistent and sometimes unintelligible. Examples of my own voice built with the Festvox tool and around 600 sentences (40 mins of speech) are available on the ACNR website (**www.acnr.co.uk/rehabfiles**) (fest_cultural, fest_cupoftea), with an example of the original recorded speech (original_smc).

More specifically aimed at people with progressive speech loss, ModelTalker, developed by the Nemours Speech Research Laboratory at the Alfred I. duPont Hospital for Children in Wilmington, Delaware, USA, is a free voice building service which can be used on any home computer (**www.modeltalker.com**). The data collection tool, ModelTalker Voice Recorder (MTVR), requires around 1800 utterances to be recorded. MTVR prompts the individual to produce an utterance, screening it for consistency of pitch, loudness and pronunciation, aiming to collect only good quality usable data. The system is optimised for US English and results have been shown to be mixed for British English speakers, particularly as the pronunciation check may have to be switched off for UK speakers. Examples of my ModelTalker voice are on the ACNR webpage (mt_cultural, mt_cupoftea). The user can upload the recordings to the developers and within a short time, the voice becomes available to download. The interface is easy to use and the voice building process is done entirely by the developers. The voice can be used on any home computer and is portable onto other communication aids that support SAPI 5.1 voices.

Commercially, voice building has been reserved for providing a service for companies due to the high cost. Cereproc (**http://www.cereproc.com/**), based in Edinburgh, estimate a voice built specifically for a company would cost upward of £20k, making it non-viable for individuals. However, Cereproc are currently looking for investment for production of software for voice banking and voice building for text to speech synthesis for individuals with progressive speech loss, so the market is becoming aware of the need for personalised synthesised voices.

On the research front, Hidden Markov Models (HMMs) are being used to create robust statistical representations of speech which can then be adapted towards an individual's speech. The HTS toolkit (http://hts.sp.nitech.ac.jp/) has been developed by Nagoya and Tokyo Institutes of Technology, Japan. The amount of data required for this technique is significantly reduced to around 100 sentences, or approximately 6-7 minutes of continuous speech. These models can generate speech for any utterance in a consistent way that sounds similar to the person's speech to which it has been adapted. The speech has a slightly more robotic quality to it but is much more consistent than concatenative techniques. It is easier to manipulate in terms of prosody and speed, making it easier to tailor the speech output to the individual's own needs. Using 500 sentences as adaptation data, examples (hts_cultural,

References: 1. Rudick RA *et al*. Neurol 1997; **49:** 358-63. 2. Jacobs LD *et al*. Ann Neurol 1996; **39:** 285-94. 3. Rudick RA *et al*. Poster presented at ECTRIMS. October 2007; Prague, Czech Republic. 4. Bermel RA *et al*. Poster presented at WCTRMS. September 2008, Montréal, Canada.

hts_cupoftea) are available to hear made from my own voice.

Current research in progress in the CAST (Clinical Applications of Speech Technology) group at the University of Sheffield (**http://www.shef.ac.uk/cast/**) is using HTS to bank and recreate voices where deterioration has begun. Where voices have started to deteriorate, features of the voice that hold the speaker's characteristics are retained and those features that are affected by the speech disorder, such as energy or duration, are substituted from the robust statistical model. This work is showing promise as a potential for voice banking pre- and post-deterioration for eventual use in personalised voice output communication aids.

For voice banking, steps can be taken to ensure that recordings could be used in the future when new technologies become available, as well as for currently available sources and techniques. If an individual wishes to make recordings onto their own computer, here are some guidelines to ensure that the quality is high and the output is as usable as possible.

1. The recordings should be as high quality as possible using a non-compressed format, i.e. WAV files not mp3. Successful voices have been built using recordings done on a home computer in a quiet room, which is usually of sufficient quality. You could also contact your local university Linguistics or Speech Science department as they may have suitable facilities or equipment that they may be willing to offer.

2. Recordings should be done either in one sitting or at the same time of day over a short period of time. Do not record if you have a cold or are fatigued and keep the recording conditions as consistent as possible. A head-mounted microphone usually improves consistency.

3. Record phrases that you use frequently, including names and places. Record tokens such as "yes", "no" and "mm" so these can be used directly as recorded for social interaction. Apart from these tokens, avoid recording isolated words; it is more useful for the voice building process to put them into contexts in longer phrases. Record favourite songs or phrases in different tones of voice as these are difficult to reconstruct with a synthesiser. Other useful phrases could be recorded such as this list devised by Beukelman and Gutmann in 1999 (http://aac.unl.edu/vocabulary.html) on which part of the ModelTalker inventory is based. Another suggestion is to spend a few days becoming aware of your own communications and note down what you might need to access quickly or say well. Talk to family and friends as they may be more aware of your idioms and care providers about what you may find useful as your condition progresses.

4. Try also to record a set of data that has a wide phonetic coverage of English. The ModelTalker database is balanced for the phonetic coverage of US English. Once downloaded, it provides a useful interface for collecting recordings, which are stored on the computer once the recordings are uploaded. It provides a recorded database whether or not the final synthesised voice from ModelTalker provides a good result. Another suggestion is to record set A of the Arctic database which is around 600 sentences. This was designed to have full coverage of the sounds of English specifically for a speech synthesis task. This can be found at http://festvox.org/cmu_arctic/.

5. Try and sound as natural as possible.

The synthesised voice end result will not be able to recreate your own voice exactly but banking as much data as possible will at least provide a starting point for a prosthetic voice to retain some of your own characteristics in that speech, either now or when the technology becomes available. ◆