# The needs of academic researchers using speech-to-text systems

Talk given at Opening up Speech Archives conference, British Library, 8 February 2013

**Luke McKernan**

There is coming, I think, a great change in how we discover things on the Internet. It is one which will play a major part in making the moving images and sound recordings all the more central to knowledge and research. The great change will be brought about by speech-to-text technologies.[1]

Technologies that convert the spoken word into readable text have been around for a while now. Dictation tools such as those produced by Dragon do a fine job for the single voice which the software has been trained to recognise, and the new generation of smartphones now incorporates voice command technologies working on much the same principles. Speech-to-text systems are used in call centres, and by broadcasters to generate a rough transcript from which subtitles are then produced. But the great challenge has been how to apply speech-to-text to large-scale collections of speech-based audio and video, such as are held by broadcasters, archives and libraries.

Take the British Library for example. We have around one million speech recordings in our Sound Archive. They have catalogue records, so you can find out basic information about their contents, but providing more detailed descriptions, or even transcriptions, is enormously time-consuming, labour-intensive and slow – with the rate of production naturally falling way behind the rate of acquisition. On the video side of things, we have a rapidly growing collection of television news, amounting to nearly 25,000 hours. Around half of this comes with subtitles captured as part of our off-air recording system, so we can offer a pretty accurate, word-for-word (and word-searchable description) for those programmes. But for the other half – such as most 24-hour news channels – there are no subtitles. All you get is a one-line description taken from the Electronic Programme Guide which says something like “the latest news from around the world”, and that’s it. We need to open up those recordings to match the level of discovery we can offer for subtitled news programmes.

This isn’t just about opening up speech archives – it’s about levelling the playing field. The digitisation and digital production of text means that full-text searching across a vast corpus is now a reality, as we see with such sites as Project Gutenberg, Hathi Trust, Gallica, Trove, British Newspaper Archive, Papers Past, the Internet Archive and more. If video and sound are to be treated equally by libraries and archives, then that means they need to be discoverable to an equivalent level of depth, and for researchers to be able to pursue subjects through books, manuscripts, newspapers, web pages, video and sound recordings on an equal footing. We need to know what those audiovisual records are saying.

[1] This talk is adapted from my blog post, 'Opening up Speech Archives', http://lukemckernan.com/2013/02/05/opening-up-speech-archives.

Over the past couple of years speech-to-text technologies have developed to a stage where we are very close to achieving such a goal. University departments, broadcasters' R&D divisions, the video industry, and the major web companies have all been in pursuit of this particular holy grail. It is no easy matter, as the human voice is a complex thing, and the huge variety of voices represented by any large video or sound archive will covers many different accents, languages, arrangements (i.e. multiple voices), instances of background noise, and so on. It's interesting to see what's driving much of this activity. It is not an idealistic wish to push back the barriers of research. Go to the websites of the developers and service providers and again and again you'll see the same thing – demonstrations on how good they are at reading Arabic. It is surveillance demands that are pushing this particular industry forwards, with governmental departments, police forces and the military being the major customers.

Over the past year we have been hosting a project here at the British Library, entitled 'Opening up Speech Archives', which has looked at the application of speech-to-text technologies for research, particularly in the art and humanities. Funded by the Arts & Humanities Research Council, the project has not been about assessing the best technical solutions. There are plenty of other people highly expert in the field who are working hard on various solutions. Happily a number of them are in the audience and will be speaking to you today. Nor is it a project about selecting one system over another. Our aim is not to pronounce on what might be the best such service for the British Library, for UK higher education, or anyone else.

Instead the project has been looking at things from academic user's perspective, and asking some basic questions. How useful are the results to researchers? What are the methodological and interpretative issues involved? And how can speech-to-text technology be adopted in UK research in a form that is readily accessible and affordable? So it is that the audience we have here today is more diverse that you might usually expect to find in an event on technological developments. We have developers here, service providers and vendors, but we also have librarians, archivists and researchers from a range of disciplines. We have people from oral history, socio-linguistics, legal studies, journalism studies, theatre studies, radio studies, sociology, modern languages and ethnomusicology. Everyone of them, I can guarantee, will see the solutions and services presented today in a different way, with different ideas about how they may or may not serve their particular subject area.

For the Opening up Speech Archives project we have been interviewing researchers, either on a one-to-one basis, or in group sessions, and getting them to try out research topics on a variety of speech-to-text and related systems. I'm very pleased that some of those researchers are with us here today. The process of testing hasn't always been that easy since the demonstration versions we had to work with necessarily only presented a selection of video and sound content, so it was necessary for the researchers to use a degree of imagination in transposing limited results to their own particular area of interest.

Nevertheless we have established some basic findings. Firstly, such services were recognised as being  unquestionably useful for researchers across many disciplines, with all recognising that opening up speech archives in this way would be a game-changer as far as research was concerned. Not all necessarily welcomed such change unequivocally. Some felt that they were quite happy with things as they are, and several complained about the confusing nature of existing resource discovery systems, such as library catalogues, feeling that adding another tier would make them feel all the more bewildered by choice. But at the same time few felt intimated by the individual systems that we asked them to try, each picking up on how the system operated and what it could deliver very quickly.

One point made that particularly interested us was a desire for such systems to make clear the process by which the results were being delivered to them, including errors in transcripts and the existence of 'false positives' - that is, results which look like they provided the right answer to your query but in fact did not, an inevitable hazard with speech recognition systems. Speech-to-text, crudely speaking, is delivered in one of two ways - dictionary-based or phoneme-based. A dictionary-based system will have a large corpus of words fed into it, and when one of these words is spoken it matches it to what it finds in its dictionary. If the word is not in its dictionary, then it goes for the whatever it can find that fits best. This occurs in particular with proper names, and is one of the small joys of working with speech-to-text systems, such as the Microsoft-developed system we trialled which translated Gordon Brown as Golden Brown, or another system which solemnly reported on the Islamist invasion of Tim Buckley (meaning Timbuktu).

Phonetic systems don't look for words, but rather for parts of sounds. So, instead of looking for 'Barack Obama', such a system looks for BUHR-OCK-OH-BUH-MUH, seeking instances in the files it has indexed where those particular sounds occur close together. This brings about its own comic joys, as we have discovered by searching words like 'turnip' across a collection of American news programmes and finding that the system reported several hits. The system was doing its best, coming up with the nearest sounds it could find, which were variously 'turn up' or 'turn in'. Those are false positives, but as said some of our researchers welcomed these. That's because they permit the researcher to judge more accurately the value of the correct matches. If all you are presented with is perfect results, how can you get a perspective on what you are researching? You need the bad to judge the value and frequency of the good.

Likewise with transcripts, which dictionary-based systems provide but phonetic systems cannot. Researchers wanted to see these warts and all. Some systems attempt to tidy up the language, or to make only parts of the transcribed text available. But the message we got was that it was better to see where the speech-to-text transcription was in error. That way you understood how the transcript had been created, and that it was not perfect. However, there is concern among some involved in teaching that students could cut and paste a transcript without bothering to listen to the sound or video recordings.  We had the example in one service we tried out of a politician being quoted as saying that there would be 'no tax breaks for married couples'. If you watched the actual video, the words spoken were 'new tax breaks for married couples'. It would be all too easy to copy the incorrect text and not bother to listen to the audio file for its own sake, which is of course what we want to see happen.  Education about how such records are created will be essential.

Speech-to-text is not perfect, and probably never will be. Accuracy rates tend to range from 60-90%, depending on whether it is one person speaking to camera, or multiple voices. One reason that news programmes feature so heavily in such services is because the use of news presenters is ideal for them. You get one person addressing the camera in  a clear voice, without any background interruptions. Speech-to-text systems are not about perfect transcriptions in any case, and it is misleading to think of them in that way – they are about improving search. You have to concentrate on what they get right, and where that leads you.

One thing we noted is that not all of the researchers could quite pick up on the huge potential of such systems, beyond finding search terms quickly. That's only the start of it. It is how such terms may be linked to a wider world of discovery that is where the real value lies. I hope you'll be able to pick up on that in some of the talks we'll have today, and in the demonstrations of six services which will be featuring during the conference lunch break. Do please take time to try them out, and imagine the possibilities. One of the things that particularly intrigues me is how the systems for extracting meaning from video recordings are offering more in the way of features than most resource discovery systems that focus on print media. So much of what is offered for the latter is not much more than the traditional catalogue experience - you look for a subject, you are presented with a list of possibles, you pick one of them, end of story. But look at a fully-featured audiovisual discovery system and you can get not just catalogue search, but speech-to-text transcriptions, subtitle search, entity extraction (that is, presenting themes derived from the video document's textual record), image recognition, face recognition, speaker recognition, melody recognition, story segmentation, story summaries, gender detection ... Once you have got your content in a digital form, with structured metadata underpinning it, the world that can be opened up for you is extraordinary. And now sound and video are going to be a part of that world.

How big will this get? A few years ago there was a bit of excitement about Gaudi, Google's own speech-to-text system which turned up on its Lab site. It let you search across videos of the first Obama election using speech-to-text, with instances of the word marked along the timeline of the video. But then it disappeared. But this doesn't mean Google has abandoned speech-to-text. Quite the opposite – the indications are that it is planning something major. There was the almost casual mention in a recent *Guardian* piece on Amit Singhal, head of Google Search, that it "has been assiduously accumulating as much human voice recording as possible, in all the languages and dialects under the sun, in order to power its translation and voice recognition projects".[2] Then there's the news of the recruitment of Ray Kurzell, language processing expert, as director of engineering, with a brief to "to create technology that truly understands human language and its real meaning"[3] and a blank cheque to enable him to do so, tend to point to something big, very big, eventually.

There are other indicators of big-ness. In January 2012 a US law was passed which says that all TV broadcasts from the USA when published on the Web need to come with closed captions, to enable accessibility for all.[4] This isn't speech-to-text, but it is making a major tier of web video word-searchable, and where such a law can be passed in the USA, can Europe be far behind in its thinking? Already you can see on any YouTube video a new automatic captions service provided on the navigation bar. The results are frequently ridiculous, but they will improve. Online services such as Amara, which offer a crowdsourcing method for captioning and translating any web video, could make a huge difference.[5] The word is out that videos contain words.

[2] Tim Adams, ' Google and the future of search: Amit Singhal and the Knowledge Graph', http://www.guardian.co.uk/technology/2013/jan/19/google-search-knowledge-graph-singhal-interview.

[3] Rip Empson, ' Imagining The Future: Ray Kurzweil Has "Unlimited Resources" For AI, Language Research At Google', http://techcrunch.com/2013/01/03/imagining-the-future-ray-kurzweil-has-unlimited-resources-for-ai-language-research-at-google.

[4] This implemented provisions of the Twenty-First Century Communications and Video Accessibility Act of 2010 (CVAA). See http://www.fcc.gov/guides/captioning-internet-video-programming.

[5] http://www.amara.org.

Will all of this change how we discover things on the Internet in a truly significant way? I think it will. It's not just that we'll be able to uncover huge amounts of speech-based content (assuming that these systems become affordable on a mass scale). It is how these records will be discoverable alongside all the other text-based records in libraries, archives, and the Web, that is going to be so revolutionary. I think we will have two levels of discovery – the basic level (a catalogue description, essentially), and the full-text level, in which every word in a document, of whatever medium, is discoverable. And from the words our systems then build, or enable us to build, will come further associations by the extraction of key terms – subjects, names, locations, dates, time periods, concepts – which can then create links to other files, and be used for themselves to visualise data, to map associations, to learn new things about the familiar and to discover the hidden and unsuspected.

Such interconnected systems won't just make us able to do what we do now, which is to search in a rather linear way, but will immerse us in data, radiating out from whatever it is we are thinking about. We will have to see things differently, ask new questions, discover things we hadn't even realised we were looking for. But if we in the world of academic research recognise that we could benefit greatly from such systems, then we have to begin a dialogue with those who are developing them. We can really help by explaining what we are looking for, or those whose subject areas we curate or represent are looking for - the questions they try to answer, or the questions they haven't been able to answer as yet, or haven't begun to think of, but could now do.

Today is about bringing the worlds of research, technological development and service provision together. You will hear talks on commercial products and R&D developments that take radically different approaches to the challenge of opening up archives using speech-to-text tools. You will hear about university projects with very specific goals in mind. You will hear about the importance of time-based metadata and the importance of the complementary field of subtitle capture. You will hear about how audiovisual services operate in higher education, and the crucial element of citation for sound and moving image. And of course you will have the opportunity to test out these systems for yourselves. The bigger picture should emerge by the end of it, hopefully.

To end with a personal note, I am thrilled that the moving image - my medium - is central to all this. Of course moving images aren't all about speech, and their unique quality ultimately lies in what they do not, in a literal sense, say. But words give moving images their specificity, and they connect the medium to the traditional modes of discovering knowledge, which is to say through books and manuscripts. 120 years or so after motion pictures were first invented, we might finally be in a position to start learning from them.