Opening up Speech Archives

Mari King
The British Library
marikingface@googlemail.com

Luke McKernan
The British Library
luke.mckernan@bl.uk

Abstract

This paper describes the work undertaken at the British Library for its project 'Opening up Speech Archives', the aim of which has been to look at the application of speech-to-text technologies for research, particularly in the Arts and Humanities.[1] *It outlines the importance of speech recognition technologies which convert speech audio into word-searchable text for making audio archives more readily available to researchers, in a form where they can be searched alongside other, text-based media. It describes the approaches undertaken by the project, including the different services tested; and it considers how transformative an effect speech-to-text technologies may eventually have on how we search for information.*

Keywords: *Archives; speech-to-text; audio; video; research*.

Introduction

The project was funded 2012/13 by the Arts and Humanities Research Council as a part of the AHRC's Digital Transformations in Arts and Humanities

[1] This paper originated in a blog post, Luke McKernan, 'Opening up Speech Archives', http://lukemckernan.com/2013/02/05/opening-up-speech-archives. It was presented by Mari King in a revised form at the Copenhagen Speech Event, University of Copenhagen, 20-22 March 2013.

theme, which aims to contribute to a package of measures aimed at developing innovative approaches to archiving, accessing and using data for research in the arts and humanities.[2] The project has not been about assessing the best technical solutions. There are many who are highly expert in this field who are working hard on various solutions.[3] Nor has it been a project about selecting one system over another. Our aim has not been 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 the academic user's perspective, by communicating directly with the arts and humanities research community and examining their information needs, uses and behaviours.

Why speech-to-text?

Technologies that convert the spoken word into readable text have been around for some while. Many people now use Apple's Siri and similar on smart phones; speech-to-text systems are used in call centres, and by broadcasters to generate a rough transcript for subtitles.

However, one major challenge that remains is how to apply speech-to-text to large-scale collections of speech based audio and video held by broadcasters, archives and libraries. The British Library, for example has around one million speech recordings in its Sound Archive. These have catalogue records, so one can find out basic information about their contents, but providing more detailed descriptions, or even transcriptions, is time-consuming, labour intensive and slow.

On the video side of things, the Library has a rapidly growing collection of television news, amounting to over 25,000 hours. Around half of this comes with subtitles captured when we record from the broadcast signal, so we can offer a reasonably accurate, word-for-word (and word-searchable) transcript for those programmes. But for the other half – such as most 24-hour news

---

[2] The project team has comprised lead investigator Luke McKernan (Lead Curator, News and Moving Image), co-investigator Paul Wilson (Curator, Radio) and project researcher Mari King,

[3] For a listing of some of the services available or in development, see http://playback.ning.com/group/speech-to-text.

channels – there are no subtitles. All one gets is a one-line description taken from the Electronic Programme Guide which says something such as 'the latest news from around the world', and that is all. We need to open up those recordings to match the level of discovery we can offer for subtitled news programmes.

This is not just about opening up speech archives – it is 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.[4] 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, for their comparative content, and for their unique content.

Thanks to significant leaps forward in computer processing power in recent years, speech-to-text technologies have finally begun to develop to a stage where we are very close to achieving such a goal. IT manufacturers, university departments, broadcasters' research and development divisions, the video industry, and the major Web companies have all been in pursuit of this particular holy grail. It is no easy matter, and as the human voice is a complex thing, and the huge variety of voices represented by any large video or sound archive will cover many different accents, languages, arrangements (i.e. multiple voices), instances of background noise, and so on. It is interesting to see what is driving much of this activity. If one goes to the websites of the developers and service providers and again and again one will 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.

---

[4] Project Gutenberg, Hathi Trust, Gallica, Trove, British Newspaper Archive, Papers Past, Internet Archive.

Our approach

The project began by asking some basic questions:

- How useful are the results to academic researchers?
- What are the methodological and interpretative issues involved?
- How can speech-to-text technology be adopted in UK research in a form that is readily accessible and affordable?

To assist us in answering these questions we interviewed researchers, either on a one-to-one basis, or in group sessions, and got them to try out research topics on a variety of speech-to-text and related systems. For internal use only we have also been conducting a speech-to-text technology survey, with over forty unique entries and growing.

To keep things manageable we focused on user testing five speech-to-text demonstration services, selected to represent the different kinds of services available.

- GreenButton – InCus (developed out of Microsoft Research's MAVIS system)
- Nexidia – Dialogue Search
- BBC R&D – World Service Radio Archive Prototype (using CMU Sphinx open source software)
- Autonomy HP – Virage/IDOL
- SAIL Labs – Media Miner

The process of testing such services with real-life research topics proved challenging. We worked with researchers from oral history, socio-linguistics, legal studies, journalism studies, theatre studies, radio studies, sociology, modern languages, and other fields. The demonstration versions made available to us only presented a very small selection of video and sound content (some, but not all, from the British Library itself), so the researchers had 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 academic research was concerned. But not (?) all welcomed such change unequivocally. Some felt that they were quite happy with things as they are. 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 they felt intimidated 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 - vocabulary-based or phoneme-based.[5] A vocabulary/grammar-based system (LVCSR – Large Vocabulary Continuous Speech Recognition) 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 vocabulary, it selects whatever term 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 one 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 do not look for words, but rather for parts of sounds, which are then matched to search terms expressed as words. Hence if one submits a search request for 'Barack Obama', such a system undertakes a syllabic analysis of the phonetic terms BUHR-OCK-OH-BUH-MUH, seeking instances in the files it has indexed where those particular sounds occur close

[5] Microsoft Research, 'Speech Recognition for Audio Indexing: Backgrounder'.

together. This brings about its own comic joys, as we have discovered by searching words such as 'turnip' across a collection of American news programmes (which are unlikely to have reports on root vegetables) 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. This is because they permit the researcher to judge more accurately the value of the correct matches. If all one is presented with is are perfect results, how can one get a perspective on what one is researching? We 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 55-90%, depending on whether it is one person or multiple voices speaking to camera. 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. Many in academia and archives who have expressed an interest in speech-to-text and speech recognition up to now have had hoped of systems able to deliver faultless transcriptions, to cut down on the laborious business of producing these manually. But this is to overestimate what the

technology can possibly provide, and to misunderstand where its real benefit lies. Such systems are fundamentally about improving search. We have to concentrate on what they get right, and where that leads us.

One thing we noted is that not all of the researchers could pick up on the huge potential of such systems, beyond finding search terms quickly. Finding information quickly is only the start of the process. It is how such terms may be linked to a wider world of discovery that is where the real value lies. Once you have got your content in a digital form, with structured metadata underpinning it, the world that is opened up becomes transformative for academic research. Now sound and video are going to be a part of that world.

How big will this get?

A few years ago there was some excitement about Gaudi, Google's own speech-to-text system which turned up on its Labs site.[6] It let users search across videos of the first Obama election using speech-to-text, with instances of the word marked along the timeline of the video. It has now been taken down from the Google Labs site and is no longer available.

But this does not 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 interview with 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.[7] Then there is the news of the recruitment of Ray Kurzweil, language processing expert, as director of engineering, with a brief to 'to create technology that truly understands human language and its real meaning' and a blank cheque to enable him to do so, tend to point to something big, very big, eventually.[8]

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

---

[6] Google Official Blog, Google Audio Indexing now on Google Labs, 16 September 2008.
[7] Tim Adams, 'Google and the future of search: Amit Singhal and the Knowledge Graph'.
[8] Rip Empson, 'Imagining The Future: Ray Kurzweil Has "Unlimited Resources" For AI, Language Research At Google'.

the Web need to come with closed captions, to enable accessibility for all.[9] This is not 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 one 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.[10] The word is out that videos contain words.

Will all of this change how we discover things on the Internet in a truly significant way? We think it will. It is not just that we will 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. We will probably 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. The systems we then build, based on the principles of Linked Open Data, will enable us to generate further associations by extracting key terms – subjects, names, locations, dates, time periods, concepts – which will create links to other files, and will be used to visualise data, to map associations, to learn new things about the familiar and to discover the hidden and unsuspected.

Such interconnected systems will not only make us able to do what we do now, which is to search in a rather linear way (query > listing > answer), 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 had not even realised we were looking for. This is the vision that

---

[9] Federal Communication Commission, 'Captioning of Internet Video Programming'. These rules implement provisions of the Twenty-First Century Communications and Video Accessibility Act of 2010 (CVAA).

underpins the Semantic Web, but the potential for having audiovisual media as central to this vision is insufficiently appreciated. But it will happen. Of course moving images and sounds are not all about speech, 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 picture film was established, and 150 years after sound recording technology was first developed, we might finally be in a position to start learning from them.

References

Adam Tim, 'Google and the future of search: Amit Singhal and the KnowledgeGraph'.

Empson, Rip, 'Imaging the Future: Ray Kurzweill has "Unlimited Resources" for AI, Language Research at Google',

Federal Communication Commission, 'Captioning of Internet Video Programming'.

Google official Blog, 'Google Audio Indexing now on Google Labs', 16 September 2008.

McKernan, Luke, Opening up Speech Archives.

Microsoft Research, Speech Recognition for Audio Indexing: Backgrounder.

Websites

Amara
BBC World Service Radio Archive Prototype
British Library Sound Archive
Gallica
Hathi Trust
InCus
Internet Archive
Linked Data
Media Miner
Nexidia
Papers Past
Project Gutenberg
Trove
Virage/IDOL