'Googling' through unique audio material: towards a better search result

July 4 2012

Searching and finding in audio archives can be improved if we take a different look at the underlying technology and allow for how the results are used. This provides a better picture of the problems and the points for improvement. Laurens van der Werff demonstrated this in his PhD thesis 'Evaluation of Noisy Transcripts for Spoken Document Retrieval', which he will defend on 5 July at the University of Twente.

Van der Werff's research was carried out within the project CHoral, which focuses on making spoken audio material from the past accessible. Dutch archives and other heritage institutions look after many hundreds of thousands of hours of audio material such as interviews with witnesses of a special event but also, for example, all transmissions of national and regional radio organisations.

If this unique audio material can be disclosed well then it will make a valuable contribution to research in the area of language use and dialect, regional and national politics, and history. CHoral is one of 18 projects from the NWO research programme CATCH (Continuous Access to Cultural Heritage) which has a total budget of more than 15 million euros and is working on the accessibility of Dutch cultural heritage.

Improved evaluation of transcripts

Automatic speech recognition in combination with search technology offers the possibility of searching through sound files: spoken word is

converted into a written text (transcript) that you can subsequently search as 'usual'. Many research labs worldwide are working hard on improving the quality of automatic speech recognition. However, for applications in search systems - and certainly for heritage collections - these improvements do not always deliver a maximum benefit.

For heritage collections, Van der Werff proposed a new way of evaluating the quality of automatically generated transcripts that pays more attention to how historians and other end-users want to use the search results. This offers the possibility of an improved analysis of where problems occur and provides leads for optimisation. Due to the limited frame of reference in the heritage sector on which optimisations can be based, this approach is a most welcome step forwards.

Specific challenges of heritage material

The audio material in heritage collections has a number of special characteristics. Many sound tapes are not digitised, they have mostly not been manually transcribed and they have no or only superficial metadata. Furthermore, it often concerns recordings from non-professional speakers with a lot of noise in the background. And many of the speakers only occur in a single sound fragment and so very little training material is available for a computer – a typical problem within cultural heritage that is exacerbated by the small geographic area Dutch is spoken in. Another complicating factor is that this heritage data is mostly used in a highly specific manner. As a result of all of these special characteristics, an approach that works well with news data, for example, cannot be automatically applied to this unique material.

Applications of the optimised technology

The techniques from the Choral project were, for example, used on

collections from the Rotterdam Municipal Archive (transmissions Radio Rijnmond; website 'Brandgrens' with eyewitness accounts about the bombing of Rotterdam), the NIOD (Radio Oranje with speeches from Queen Wilhelmina during World War II; eyewitness accounts of survivors from Buchenwald) and the interview archive of Aletta/IAVV.

The knowledge and techniques from CHoral have also helped to lay the basis for the open source speech recognition package SHoUT (University of Twente) that has been further developed within the CATCH valorisation programme CATCHPlus (www.catchplus.nl). Using this software each archive can now, in principle, make its audio sources accessible without the need for its own in-house specialists. SHoUT is already being used for the national website 'Verteld Verleden' ['Spoken Past'], through which all audio sources in the Netherlands will be accessible in the future.

Further information: www.nwo.nl/catch and www.nwo.nl/catch/choral

Provided by Netherlands Organisation for Scientific Research (NWO)