Hidden Treasures Lost Forever? Speech technology & the disclosure of Dutch audiovisual archives

Mies Langelaar, Willemijn Heeren, Gemeentearchief Rotterdam, Rotterdam, The Netherlands & Universiteit Twente, Enschede, The Netherlands

The problem: Given the enormous backlog at audiovisual archives, digitized and digital born, and the generally high levels of item description, collection disclosure and item access are both at risk. Technology can play a role in improving disclosure and access of digitized spoken word collections during and after transfer to the archive. Therefore, the potential of automatic annotation and search technology are being investigated in the ongoing NWO CATCH project CHoral, a unique cooperation between speech technology researchers from the University of Twente and archivists from the Rotterdam Municipal Archives. In this paper we will present ongoing research in CHoral. The test case: One collection that is typical of A/V archives in the cultural heritage (CH) domain is the archive of the city of Rotterdam's regional radio channel `Radio Rijnmond'. This collection consists of recordings since its initial broadcast in 1983, amounting to over 60.000 hours today. Only a few percent of the collection has been disclosed on a fairly high level, while the large majority remains in the deposits, undisclosed. Making detailed annotations for disclosure would take long and would be very costly. The main problems with this example collection are: 1)there is a large backlog of undisclosed, i.e. un-accessible, material 2)most of the available annotations are fairly unspecific, restricting their use for answering information needs 3)most of the audio is being kept on analog data carriers or on CDs, i.e. data carriers that do not support online search of the collection. Our approach: The first two problems are being addressed by researching and developing technologies for automatic metadata generation. Specifically, spoken document retrieval technology is being researched and developed for application in Dutch audiovisual archives. Experiments so far have shown that annotation and disclosure through speech and search technology are possible, and of course we will show some examples. On the other hand, the archive has developed a standard metadata scheme to ensure that all information is disclosed along the same lines, and that newly received data is up to this standard. Moreover, to allow end users online access to audiovisual documents, tools are being developed to facilitate interpretation of and interaction with the data. The third problem is being addressed by digitisation of the analog data carriers, and transfer of data on CDs to hard disks. Along with this a trusted digital repository is being developed to ensure long-term preservation of the digital audio material. Finally, we try to make agreements with the companies and institutions that deliver A/V material about the kind of metadata we need and the disclosing methods. What's left for discussion: We will argue that the work-flow and daily practice at audiovisual archives on the one hand, and the state-of-the-art in technology on the other need careful tuning to be mutually beneficial, but it can be done! We will moreover reflect on the envisioned benefits that ongoing cooperation can bring.