Facilitating Digital Humanities Research in Central Europe through the Advanced Speech and Image Processing Technologies

Jan Švec[1], Petr Stanislav[2], Marek Hrúz[3],
Aleš Pražák[4], Josef V. Psutka[5], Pavel Ircing[6]

In the recent decades, there is a constantly growing amount of multimodal data being collected and stored in order to be preserved for a future use. These data include – among other things – videotaped oral history interviews, archived footage of various TV broadcasts and a plethora of scanned hand-written and typed documents and photographs. The resulting archives present invaluable resources for many branches of the humanities (history, linguistics) and social sciences (political science, communication studies).

However, almost all of such archives share the same problem; unless they are really thoroughly equipped with the detailed metadata (describing e.g. topics of the individual documents, names and place appearing in the documents and/or being mentioned there, etc.) – and it is only rarely the case – it is almost impossible to find the desired information in the vast amount of data.

Our research lab has been participating in several projects (in some of them as the leading partner) that applied first the speech processing technologies and later also the image processing ones to facilitate the access to the information content of

[1] Department of Cybernetics & NTIS – New Technologies for the Information Society Faculty of Applied Sciences, University of West Bohemia, Plzeň, Czech Republic

[2] Department of Cybernetics & NTIS – New Technologies for the Information Society Faculty of Applied Sciences, University of West Bohemia, Plzeň, Czech Republic

[3] Department of Cybernetics & NTIS – New Technologies for the Information Society Faculty of Applied Sciences, University of West Bohemia, Plzeň, Czech Republic

[4] Department of Cybernetics & NTIS – New Technologies for the Information Society Faculty of Applied Sciences, University of West Bohemia, Plzeň, Czech Republic

[5] Department of Cybernetics & NTIS – New Technologies for the Information Society Faculty of Applied Sciences, University of West Bohemia, Plzeň, Czech Republic

[6] Department of Cybernetics & NTIS – New Technologies for the Information Society Faculty of Applied Sciences, University of West Bohemia, Plzeň, Czech Republic

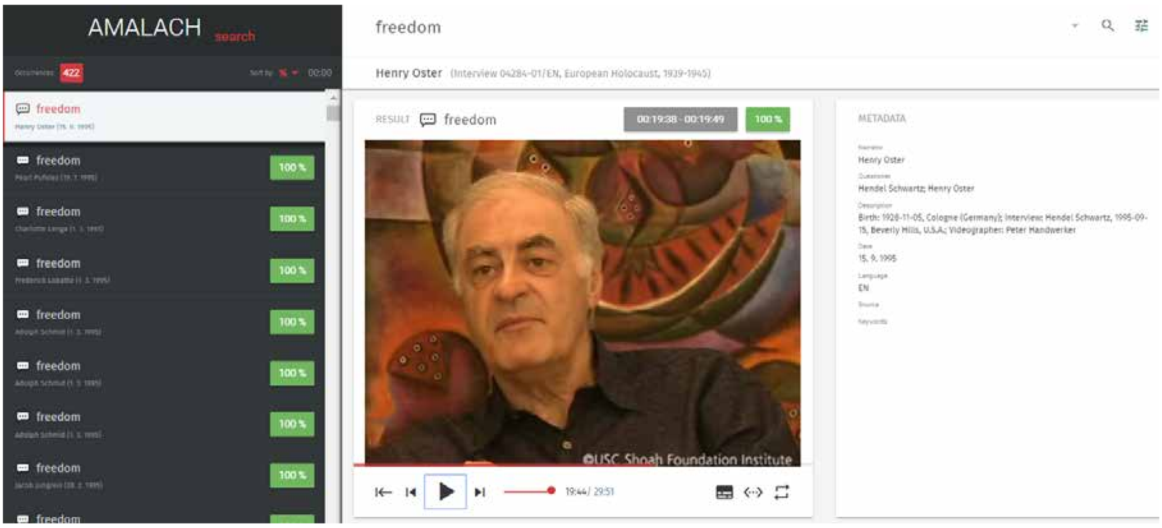
the archives.

First, we started with building the systems for automatic speech recognition (ASR) for the Czech, Slovak, Russian and Polish recordings from what is nowadays the USC Shoah Foundation Visual History Archive (http://vhaonline.usc.edu). The original plan was to transcribe the audio track of the video recordings into the plain text format and then use the traditional information retrieval techniques to find elaborately crafted search topics. This has proven to be problematic due to a relatively poor recognition accuracy (more than 30% Word Error Rate) on the challenging data and we have resorted to using an approach called *spoken* **term detection (STD)** instead.

In STD, the task is to find any occurrence of a specified word or a short phrase (**the query**) in the archive and return the exact time point where such occurrence takes place. Ideally, there is also a graphical interface that allows user to play the relevant segment from the recording instantly, as is shown in Figure 1.

Since the user does not need to read the text produced by the ASR as he/she watches directly the original passage from the recordings, we can use more than just the usual sequence of words that the ASR engine deems most probable (so called one-best transcription). Instead, we can take into account the entire set of recognition hypotheses stored in the form of a lattice – a directed acyclic graph where nodes represent the time instants and edges "carry" words together with a confidence score that represents a probability that a given word actually occurs within the given time interval.

In order to allow efficient searching, the lattices are processed into the data structure called the **(inverted) index** using the following procedure: The individual edges of the lattice are subject of a two-stage pruning. The first stage takes place at the beginning when all the edges whose confidence scoreis lower than a threshold $\theta_w = 0.05$ are discarded. Each of the remaining edges is represented by a 5-tuple (*start_t, end_t, word, score, item_id*) where *start_t* and *end_t* are the beginning and end time, respectively, *word* is the ASR lexicon item associated with the edge, *score* is the aforementioned confidence score and finally *item_id* is the identifier of the original video file (*start_t* and *end_t* represent the offset relative to the beginning of this file). The index is further pruned by removing similar items; that is, if there are two edges labeled with the same word that are either overlapping or are being less than $\Delta\, t_w = 0.5\, s$ apart, only the edge with the higher score is retained. It follows from the description that the indexing procedure omits the structural properties of the original lattice but, on the other hand, makes a compact and efficient representation of the recognized data. The resulting index is stored in a MongoDB database.


AMALACH search
422
00:00
freedom
100 %
freedom
Henry Oster (Interview 04284-01/EN, European Holocaust, 1939-1945)
RESULT
00:19:38 - 00:19:49
©USC Shoah Foundation Institute
19:44/ 29:51
METADATA

Henry Oster
Hendel Schwartz; Henry Oster
Birth: 1928-11-05, Cologne (Germany); Interview: Hendel Schwartz, 1995-09-15, Beverly Hills, U.S.A.; Videographer: Peter Handwerker
15. 9. 1995
EN

Figure 1 - The latest version of the user interface for STD in the USC Shoah Foundation Visual History Archive

The current version of the search system depicted in Figure 1 is able to search in all the Czech and Slovak recordings from the USC Shoah Foundation Visual History Archive (approx. 1 000 hours and 800 hours, respectively) and about 2 000 hours of the English ones (in this case, it amounts to only about 4% of all the available interviews). All Czech, Slovak and English versions use the same core ASR, indexing and visualization technologies but of course there is also a substantial amount of localization to the individual languages – from the obvious ones such as different ASR phoneme sets and lexicons to the completely distinct approach to lemmatization used for Czech and Slovak on one side and English on the other.

In the last decade, we have started developing similar search engines for Czech institutions that have recorded similar audiovisual archives. The processing of the audio track of the video recordings was the main task in all cases but there were also some additional "data streams" to be explored. The first of the archives – the collection put together by Institute for the Study of Totalitarian Regimes (https://www.ustrcr.cz/en/) – includes also a substantial amount of scanned written documents that we wanted to make searchable in the similar manner as the audio content. The second archive – the ever-growing collection of the main news broadcast of the Czech public television broadcaster – contains the video recordings only but the broadcaster asked us to develop a technique that would allow to search also in the "visual" track. That is, to look for a specific person appearing in the video and to find the occurrence of words showing on the screen (running headlines, text documents shot by the camera or even banners in the crowd).

Thinking about those requirements carefully, we have found out that all of them can be rather easily indexed as in the case of audio-only indexing. Once we have a sequence (or lattice) of words, we can create a searchable index in the same manner as described above for the speech data. The challenge is then reduced to getting the text representation from all the modalities – this can be done using the **optical character recognition (OCR)** when dealing with scanned documents, **face detection and recognition** when searching for faces and a so-called **reading text in the wild** techniques when processing the text from the video footage. That way we can create multiple indices and have the application to search in all of them. Note that for the faces and the text captured in the video footage the individual components of the index 5-tuple have the same meaning as for the speech track, in the case of the words recognized in the scanned documents, we use the coordinates of the word bounding boxes instead of *start_t* and *end_t* labels.