INTERSPEECH 2016: Show & Tell Contribution
September 8–12, 2016, San Francisco, USA

An Engine for Online Video Search in Large Archives of the Holocaust Testimonies

Petr Stanislav, Jan Švec, Pavel Ircing

Department of Cybernetics,
University of West Bohemia, Pilsen, Czech Republic
{pstanisl, honzas, ircing}@kky.zcu.cz

Abstract

In this paper we present an online system for cross-lingual lexical (full-text) searching in the large archive of the Holocaust testimonies.

Video interviews recorded in two languages (English and Czech) were automatically transcribed and indexed in order to provide efficient access to the lexical content of the recordings. The engine takes advantage of the state-of-the-art speech recognition system and performs fast spoken term detection (STD), providing direct access to the segments of interviews containing queried words or short phrases.

Index Terms: speech recognition, information retrieval, human-computer interface

1. Introduction

The archive at hand contains about 52,000 interviews in 32 languages (116,000 hours of video in total) of personal memories of the Holocaust survivors and witnesses. Its collection was initiated by Steven Spielberg and his Shoah Visual History Foundation in mid-1990's and it is currently maintained by the USC Shoah Foundation Institute.[1]

Prior to the development of the described search engine, the archive was searchable through manually assigned keywords. However, it turned out that the coverage of the archive by this metadata is rather sparse. Thus the main idea behind the development of the presented online video search engine was to provide an easy-to-use tool where users could type an arbitrary textual query and receive a ranked list of pointers to the relevant passages of recorded testimonies that can be directly replayed[2]. Example of the search results for query "camp" is in Fig. 1.

Given the size of the archive, the new indexing method inevitably had to be automatic [1]. Two rather large portions of the archive (1,000 hours for English and the same amount for Czech)were therefore processed by proprietary automatic speech recognition systems and the resulting lattices were stored in two separate indexes. The indexes are then accessed via carefully designed GUI that also enables automatic translation of the queries and provides a cross-lingual search facility. The detailed description of all system components is given below.

[1] http://dornsife.usc.edu/vhi/
[2] Example is available at https://youtu.be/nQzZONSU2OY.

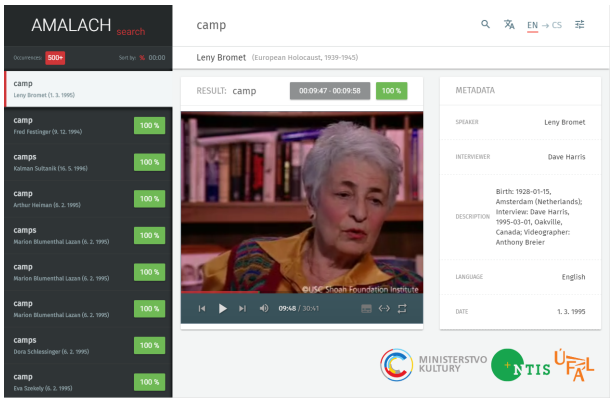

Figure 1: *The graphical user interface with ranked list of relevant passages for query "camp".*

2. System outline

2.1. Acoustic and language models

The acoustic models were of course build for each language separately and are based on the hidden Markov models (HMM) architecture – standard 3-state left-to-right models with a mixture of multiple Gaussians in each state are used. The data were parametrized as 15 dimensional PLP cepstral features including their delta and delta-delta derivatives. Given the highly specific nature of the audio data in this task (accented and often highly emotional speech of the Holocaust survivors), we have used only the manually transcribed portion of the testimonies for the acoustic model training (approx. 100 hours for each language). As for the language models, previous experiments showed that the Holocaust data are rather specific even from the language modeling perspective and thus the best strategy is to interpolate the n-gram model trained using the transcribed text with models estimated on the large body of "general" data (Google N-grams for English, selected newspaper articles for Czech).

3. Indexing and searching

The video recordings are automatically processed in the ASR pipeline. First, the video is converted into the common MP4 format, at the same time, the audio is extracted. Then, the audio signal is segmented on speaker changes into segments roughly equivalent to dialog turns. Each segment is then recognized in the ASR module. The recognizer produces the word lattice with word posterior scores. The words with posterior proba-

bility higher than a given threshold are then indexed into the inverted index.

The media files along with the ASR results and the inverted index are stored in the document database MongoDB. Such technical solution allows fast search in the index and real-time streaming of the media.

4. Application overview

The graphical user interface is designed with the IT non-professional in mind and is therefore as simple as possible (see the Fig. 1).

The text field on the top of the page is used to enter a searched word or phrase. The default search operation uses stemming to find all possible forms of the searched word (e.g. house and houses). The query can contain operators for more detailed control of the search engine behaviour – in the current version, the plus sign could be used to denote the optional word and enclosing the word or the phrase in quotes instructs engine to search for exact match.

The major portion of the GUI page is occupied with a list of retrieved results and detailed information of currently selected result. The sorted list of results is on the left side of the page and shows only basic information, mainly retrieved word(s) form, the value of confidence measure and the name of the interviewed person. The user can change the order and type of the sort algorithm (e.g. the results can be sorted by confidence or time). When sorting, the results from the same video are always grouped together to minimize the video reloading. The user can select a result by clicking on the item in the list. Immediately after selecting the result, the video is streamed from the server and starts playing just a couple of seconds ahead of the occurrence of the searched word in the audio signal.

The most important part of the GUI is the detail of the retrieved result. It contains the video player and the card with the video record metadata. The metadata are stored alongside the ASR results and the inverted index in the MongoDB. The corresponding metadata are sent together with the search results – they include information about the speaker, interviewer, short description, language in which the interview is recorded and the date of the interview.

It is essential to present relevant portions of the video within the graphical interface of the search engine simply and clearly. In most cases, the user wants to know not only when the search phrase uttered but also in what context. To fulfill this task, the result detail contains our own HTML5 video player. The received results contain relatively precise (narrow) boundaries of occurrence in the source video and therefore, it is difficult to quickly understand the meaning and the context, especially when the searched query is a single word. For this reason, the player expands the boundaries by a certain time epsilon (the default value is *5s* but it can be changed in settings); that way the user can easily hear the search term and its context. In order to facilitate the analysis of the additional results in the same video, the player marks all other outcomes on the player timeline, as shown in Fig 2. The user can click on the marked result and the player immediately jumps to this occurrence.

When the playback starts, the results are automatically replayed one after another, in the order defined by the list on the left side of the GUI. If such behaviour is not desired, the user can switch to endless repetition of the current result.

In some situations, the default time frame for playback is not sufficient for understanding the context. Thus the GUI contains a button for one-time expansion of the replayed time window.

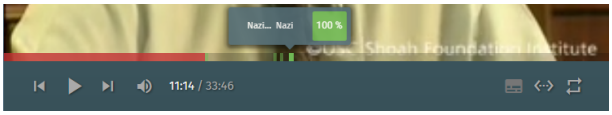

Figure 2: *The detail of the player timeline with the highlighted results in the video.*

As mentioned previously, the archive contains intreviews in many languages, therefore we have implemented the function for easy translation of the query. After selecting the source and the target language (see in Fig. 3), the translation is only a matter of clicking on the button. It should be noted that the query is always searched in one language at the time; the currently searched language is marked by a red line under the code language (see Fig. 1 and Fig. 3). The language of the search is automatically switched after translation.

Figure 3: *The detail of the search text field with the expanded menu with available languages for translation.*

5. Discussion

The search engine is currently being used by experts and also by general public in the Malach Centre for Visual History and Jewish Museum in Prague.

The architecture is highly modular, which will allow to add new functionality to the system. Currently, we plan to include the metadata search facility and also the handling of scanned and OCR processed text documents. The core of the system (with necessary modifications) could – and will – be therefore utilized for several other audiovisual archives.

6. Acknowledgements

This publication was supported by the project LO1506 of the Czech Ministry of Education, Youth and Sports.

7. References

[1] J. Psutka, J. Švec, J. V. Psutka, J. Vaněk, A. Pražák, L. Šmídl, and P. Ircing, "System for fast lexical and phonetic spoken term detection in a Czech cultural heritage archive," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2011, no. 1, p. 10, 2011.