HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

Videx

Indexing videos for the "official bulletin" of the Swiss Federal Parliament

Jonas Oesch
Prof. Andreas Sidler

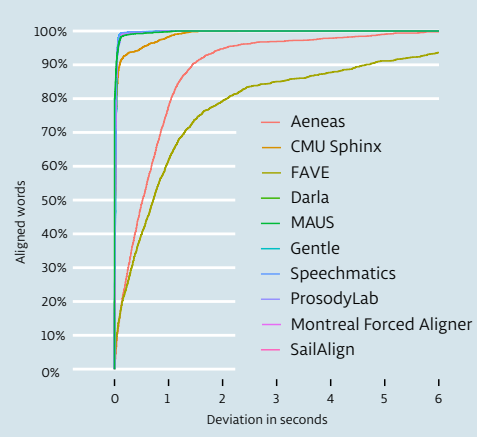
Media Engineering Institute
COMEM+
HEIG-VD

Situation

The Swiss Federal Parliament records all of its meetings on video and publishes them on its website. The videos are acompanied by a transcript created by the specialists of the parliamentary services. Searching the website reveals the correct transcript and the corresponding recording. Yet there is no way to find the position in the video where the search term is being said. A first approach using pure voice recognition produced unsatisfying results. [1] The parliamentary services wanted to know therefore if the videos could be indexed based on the existing transcripts.

Forced Alignment

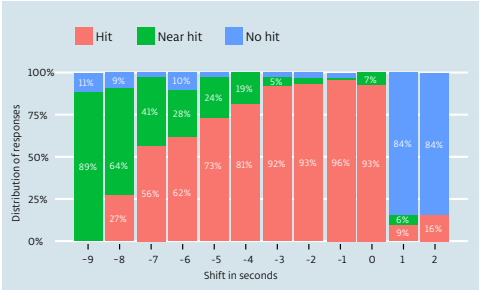
Relating words of a transcript with the corresponding sounds in a recording is known as "forced alignment" in literature. Tools for forced alignment (Aligners) exist primarily for the purpose of linguistic research. [2] A selection of the available tools [3] has been tested to find out how well forced alignment works in general. Because all tools work with English out of the box, a 13-minute speech of John F. Kennedy and an exact transcript were used as a test corpus. The results of most aligners are very impressive and show no deviation (<100 ms) from the hand-made alignment for >95% of the words.



Word-alignment quality of different tools for a 13-minute Kennedy speech.

Alignment precision

To understand the necessary precision of the alignment, we let 21 persons use a basic video-search interface. The entry point into the video was randomly shifted by −9 to +2 seconds compared to where the search term actually occured. The test showed that a shift of up to -4 seconds (early entry) is considered acceptable by most participants.

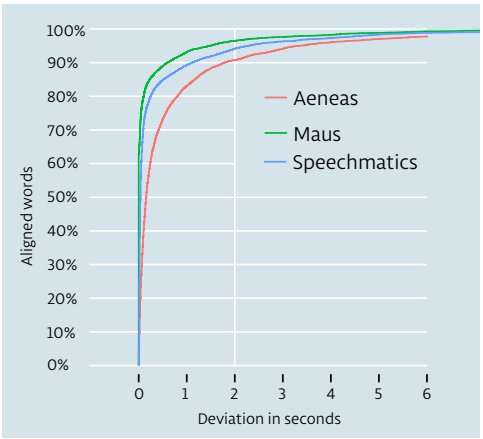


Impact of different shifts on the perception of "a hit"

Meeting Videos

Three aligners [4][8][10] met all necessary criteria for a future implementation. These were tested with a 78-minute corpus of videos from parliament meetings. The corpus covered different aspects of the meeting videos like language, differences between transcript and speech as well as the sound quality.

All aligners placed over 90% of the words inside of the available threshold of ±2 seconds.



Word-alignment quality for a 78-minute test corpus of parliament speeches

Results

From the three aligners compared, MAUS [8] produced the best results for the given test corpus. The biggest reason for false alignments were transcripts that differed strongly from the speech. Such "dissimilar" transcripts were defined by a difference of more than 10% between the words in the transcript and in the video.

MAUS	Share	±2 seconds
Highly similar	58%	98.7%
Dissimilar	42%	94.5%
Average		*96.5%*

Word-alignment quality of MAUS between the videos with a more and less precise transcript

Sources

[1] G. Szaszak, M. Cernak, P. N. Garner, P. Motlicek, A. Nanchen, and F. Tarsetti, "Automatic Speech Indexing System of Bilingual Video Parliament Interventions," 2013.
[2] N. Pörner, "Development of an automatic chunk segmentation tool for long transcribed speech recordings,"
[3] A. Pettarin, "GitHub - pettarin/forced-alignment-tools: A collection of links and notes on forced alignment tools," github.com/pettarin/forced-alignment-tools
[4] A. Pettarin, "Github - readbeyond/aeneas," github.com/readbeyond/aeneas
[5] "CMU Sphinx," cmusphinx.sourceforge.net
[6] Rosenfelder, Ingrid; Fruehwald, Joe; Evanini, Keelan and Jiahong Yuan. 2011. FAVE (Forced Alignment and Vowel Extraction) Program Suite. fave.ling.upenn.edu.
[7] S. Reddy and J. N. Stanford, "Toward completely automated vowel extraction: Introducing DARLA," Linguistics Vanguard, vol. 1, no. 1, 2015.
[8] F. Schiel, "Munich AUtomatic Segmentation," www.bas.uni-muenchen.de/Bas/BasMAUS.html
[9] "Gentle | Lowerquality," lowerquality.com/gentle
[10] "Speechmatics | API," www.speechmatics.com
[11] "Prosodylab-Aligner," prosodylab.org/tools/aligner
[12] McAuliffe, Michael, Michaela Socolof, Sarah Mihuc, and Michael Wagner (2016). Montreal Forced Aligner, montreal-corpustools.github.io/Montreal-Forced-Aligner
[13] A. Katsamanis, M. Black, and P. G. Georgiou, "SailAlign: Robust long speech-text alignment,"