Zürcher Hochschule
für Angewandte Wissenschaften

School of
Engineering
InIT Institut für angewandte
Informationstechnologie

Evaluation of Tools for Video Subtitling

Stephan Roth
ZHAW – InIT
07.12.2021

Table of Contents

1 Introduction 3
2 Test Videos 3
2.1 Evaluation Attributes for Video Cut 4
2.2 Overview of Test Videos 4
3 Tools 5
3.1 Reduce Number of Tools 5
4 Evaluation 6
4.1 Reference Data 6
4.2 Test Execution 6
5 Results 7
5.1 Preparation 7
5.2 WER – Word Error Rate 7
5.3 Detected Errors 8
5.4 Classification of Understandability 8
6 Discussion 9
6.1 Limitations 9
6.2 Further Improvement 9
7 Appendix 10
7.1 References 10
7.2 List of Files 10
7.3 Video Download 10

1 Introduction

The ETH aims to make lecture recordings and other video material accessible through subtitling. The ZHAW's ICT Accessibility Lab supports the ETH in evaluating currently available tools for automatic and semiautomatic subtitling of their videos.

This document compares potential service providers for subtitling of videos using sample videos from ETH and previously defined evaluation criteria.

(Refer too to offer «Offerte_Video_Captioning_v1.docx», 22.04.2021, ZHAW, A. Darvishy)

Numerical data: Every year 5'000 to 10'000 lecture recordings with a duration of 60 to 80 minutes each must be subtitled. For additional 1'000 to 5'000 recordings high quality subtitling is needed.

2 Test Videos

The tool evaluation is done with 8 representative videos[1] provided by the ETH.

Numerical data: These videos have a duration between 5 minutes and 2 hours in their raw version; the typical duration is 1 hour. Each video is shortened to a duration of approx. 5 minutes.

The videos are a conscious choice. It also had to be taken into account that certain records are copyright protected (i.e. D-MAVT). The selection includes recordings of different quality, duration and subject. The videos are to be treated confidentially, may only be used for the purpose of testing; under no circumstances they may be published at other locations.

Overview of the raw test videos:

No.	Description	Duration [hh:mm:ss]	Language	Recording Type	Link 1)
1	D-ARCH	01:24:54	DE	Zoom	Download
2	D-MATL	00:44:37	EN	older, lecture hall	Download
3	D-MATH	01:15:05	DE	lecture hall	Download
4	D-BSSE	01:53:56	EN	lecture hall (video conference)	Download
5	D-ITET	00:55:41	EN	lecture hall	Download
6	D-PHYS	01:33:00	DE	lecture hall	Download
7	event	01:58:15	EN	HG F 30 (Audimax)	Download
8	image video	00:05:09	EN	video production	Download

1) Download: check description in chap. 7.3

[1] The additional 9th video «opencast» is not used at the request of the ETH.

2.1 Evaluation Attributes for Video Cut

In order to shorten a video, its specific acoustic situations must be considered. Each video is classified according to acoustic attributes[2]; these are:

- Speaker
  - Pitch: female, male, mixed
  - Accent: strong, normal, none
  - Language: CH, DE, EN
- Audio
  - Echo: none, weak, strong
  - Background sound: none, some, loud
  - Noise: none, some, loud
  - Level: low (weak), normal, good
- Sampling rate

2.2 Overview of Test Videos

The file «overview.xlsx» (in German) gives an overview of the test videos, the cutting of the raw videos and the state of the captions by test video and tool. The overview contains this 4 tabs:

Tab name	Content
Übersicht	overview of the cut and evaluated test videos
Auswahl	evaluation criteria and their meanings
Schnitt	selection of the parts that are assembled from the raw videos
done	overview of which video is transcribed with which service

[2] Check tab «Übersicht» in file «overview.xlsx».

3 Tools

The commonly as «best practice» known video captioning tools as well as tools used by professional services like public television and a tool used in bachelor theses at the ZHAW are taken into account. This list of tools is coordinated with the ETH and contains 25 tools.

3.1 Reduce Number of Tools

Evaluation attributes are added to the list, each tool is evaluated with them. The main attributes are:

- Type («Typ»): service provider («Dienstleister»), service («Dienst»), software
- Test: test candidate («Testkandidat»)
  decision whether this tool is suitable for testing
- State («Status»)
  For future contact:
  - Source («Quelle»)
  - Link
  - Contact person («Ansprechperson»)
  - Contact («Kontakt»)
- Discount Price («Rabatt-Preis»)
  Only the standardized price in CHF per hour of video for automated captioning is used for the price comparison. The other details are given for information only.
- Languages («Sprachen»): DE/EN
- Location (Privacy) («Ort (Privacy)»)

The file «Anbieter.xlsx» (in German) contains the evaluation list and is split in this 3 tabs:

Tab	Content
Übersicht	tools overview with rated attributes *Hint: Filter column «Test» for «Testkandidat».*
Hinweise	references: exchange rates, numerical data (number of videos per year)
Auswahllilsten	evaluation attributes and their meanings

Based on the rating in the evaluation list the test candidates are filtered. This filtered tools are then tested in their application:

Tool	Comments
Azure	Microsoft
Happy Scribe	
Otter	
Sonix AI	
speech-to-text	Google
SWISS TXT	CH; own infrastructure; used by Swiss TV SRG
Transcribe	Amazon
Watson	IBM

4 Evaluation

All 8 tools are used to create the transcriptions for all 8 test videos. The quality of the transcriptions is compared using a numerical value (WER, chap. 5.2).

4.1 Reference Data

The transcriptions are compared against a reference. The reference from a test video is based on the automated transcription from the tool «Azure (MS)» and then manually corrected. The correction is done by viewing and listening the test video and reading the transcription in parallel. Each recognized error in the reference is corrected.

4.2 Test Execution

Usually a test account of a tool offers enough features to generate the transcriptions of all test videos. The transcriptions were carried out between 13.08.2021 and 18.08.2021.

5 Results

The quality of all transcriptions from all tools is compared with their specific references. The numerical WER value is used for this.

5.1 Preparation

Before the calculation each transcription is normalized.

- delete non-alphabet characters
- change to lower case
- replace newline with space

5.2 WER – Word Error Rate

$$WER = \frac{S + D + I}{N}$$

- WER word error rate
- S number of substitutions (words replaced with other words)
- D number of deletions (words that are not present but should)
- I number of insertions (words that are present but shouldn't)

The WER values of all transcriptions are calculated with the Python library datasets [2] [3].

Overview of all calculated WER values:

Video \ Service	Lang.-Select.	reference	wav2vec2	Azure (Microsoft)	Happy Scribe	Otter 1)	Sonix	speech-to-text (Google cloud)	Swiss TXT	Transcribe (Amazon)	Watson (IBM)
Price [CHF / h]				1.38	13.20	0.12	4.60	1.32	60.00	1.32	1.10
01	de-CH	X			40.9%					39.2%	
01	de-DE				27.3%					38.3%	
01	DE		48.4%	13.8%			26.8%	41.7%	26.8%		14.3%
03	de-DE	X								22.6%	
03	DE		29.1%	5.4%	15.9%		15.7%	20.8%	15.5%		6.9%
06	de-DE	X								22.7%	
06	DE		23.1%	9.4%	97.2%		16.2%	18.8%	16.1%		10.9%
Median			*29.1%*	*9.4%*	*27.3%*		*16.2%*	*20.8%*	*16.1%*	*22.7%*	*10.9%*
02	en-GB	X								8.5%	18.0%
02	EN		9.9%	2.0%	10.7%	5.2%	10.6%	17.9%	10.4%		
04	en-US	X								18.5%	15.5%
04	EN		24.3%	7.1%	14.8%	12.7%	14.8%	28.3%	14.9%		
05	en-US	X								19.9%	15.4%
05	EN		28.1%	12.7%	18.7%	17.6%	24.3%	44.2%	25.1%		
07	en-GB	X								16.9%	37.2%
07	EN		25.5%	13.3%	14.6%	12.6%	14.0%	37.8%	14.0%		
08	en-US	X								10.0%	11.5%
08	EN		21.7%	6.4%	8.5%	8.3%	8.3%	25.8%	8.0%		
08	EN								9.7%		
Median			*24.3%*	*7.1%*	*14.6%*	*12.6%*	*14.0%*	*28.3%*	*12.2%*	*16.9%*	*15.5%*

Legende	
	1) *Otter supports only English at the moment.*
	not included in Median
Ranking	
	best value
	2nd best value
	worst value

This table contains the additional column «wav2vec2», a tool used in bachelor theses at the ZHAW.

5.3 Detected Errors

In addition to the WER calculation the errors that occur are examined more closely. The following error classes were found:

- numbers: words or Arabic numerals
- proper names (DE: Eigennamen)
- compound words

5.4 Classification of Understandability

In addition, the understandability of the transcriptions is assessed. For this purpose, the WER limits of good and poor understandability is determined subjectively. This consideration takes place for transcriptions done with Azure and some transcriptions of the other tools.

Overview of the subjective transcription understandability:

Video	Language	Tool	WER	Understandability
1	DE	Azure	13.8%	medium
3	DE	Azure	5.4%	good
6	DE	Azure	9.4%	good
2	EN	Azure	2.0%	good
4	EN	Azure	7.1%	good
5	EN	Azure	12.7%	medium
7	EN	Azure	13.3%	medium
8	EN	Azure	6.4%	good
1	DE	wav2vec2	27.3%	bad
2	EN	Google	17.9%	bad
3	DE	HappyScribe	15.9%	bad
4	EN	Otter	12.7%	good
5	EN	Watson	15.4%	medium
6	DE	Watson	10.9%	good

6 Discussion

The tools are compared with one another on the basis of the transcription results.

- **Azure** (Microsoft) has an outstanding good WER.
- **Watson** (IBM) has also a very good WER (with 1 exception).
- At the other end of the scale wav2vec2 needs a lot of improvement and is not (yet) a practicable solution. (This may change in the future after additional bachelor theses.)

The following approximate rule applies for the **understandability classification** of the WER values:

- good below 12%
- medium between 12% to 15%
- poor greater than 15%

6.1 Limitations

The results stand for the **state of the tools in mid August 2021** (time of transcriptions). The tools are currently still improved. E.g. the transcription results from speech-to-text (Google cloud) could not be verified due the ongoing further development of the underlying model as well as the API.

The number of videos is **not sufficient** for a reliable **statistic evidence**. Therefore, the resulting numerical values must only be considered as an order of magnitude.

6.2 Further Improvement

We currently **check** the **reference transcription**. Each transcription tool has its own coding, i.e. write numbers as digits or words, write compound words together or with hyphen or separately. Since the reference transcription is based on Azure, the different encodings could lead to a systematic disadvantage for the other tools. The results of this clarification will be added to a next version of this document.

The reference transcription is improved by implementing **additional normalization roules** (chap. 5.1):

- no filler words (i.e. äh, ehm)
- keep word repetitions
- numbers as words

As part of a student work, first attempts with **adding a** very small **dictionary** (phrasebook) showed measurable and positive improvements.

7 Appendix

7.1 References

[1] Camtasia V21.0.3; TechSmith; https://www.techsmith.com/video-editor.html; video editing software

[2] Datasets V1.12.1; huggingface; https://huggingface.co/docs/datasets; Python library to calculate WER values

[3] Datasets github link to used version: https://github.com/huggingface/datasets/tree/549cd55e6d32ce03884963b1db47d2ff9bd64d5e/metrics/wer

7.2 List of Files

This is an overview of the files delivered via download link.

Filename	Content
Anbieter.xlsx	list for tool evaluation
overview.xlsx	• overview of test videos • overview of cutting of raw videos • state of transcriptions
VideoTranscriptionAnalysis.docx	this document
WER_results.xlsx	rated transcriptions: word error rates
Transcriptions/[01-08]/<tool>	[01-08].* • *.mp4: cutted test videos (downsized) • *.mp3: extracted audio • *.tscproj: cutting info for Camtasia <tool>: directory for each tool used containing transcription files directory _basis_: manually corrected transcription

7.3 Video Download

- click on «Download» link: webbrowser starts with lectures page from “ETH Zürich”
- unfold «Media», «Presentations»: select resolution and quality of video
- in webbrowser: open developer tools (F12)
- search for element <video> <source>
  playerContainer
    playerContainer_videoConntainer
      playerContainer_videoConntainer_videoContainer
        videoPlayerWrapper
          <video>
            <source>
- open source path in webbrowser: video starts
- right mouse click on video: “save video as”
- right mouse click on downloaded video: “settings > details”
  bitrate, number of channels, sampling rate